T. C.
SİVAS CUMHURİYET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY ZEKÂ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE İŞSİZLERİN ZAMANA GÖRE İŞSİZ KALMA RİSKLERİNİN TESPİTİ

YÜKSEK LİSANS TEZİ

YUSUF AKTAŞ

(20219258010)

Yapay Zekâ ve Veri Bilimi Ana Bilim Dalı

Tez Danışmanı: Prof. Dr. Hidayet TAKCI

SİVAS

EYLÜL 2024

Yusuf AKTAŞ'ın hazırladığı ve **"YAPAY ZEKÂ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE İŞSİZLERİN ZAMANA GÖRE İŞSİZ KALMA RİSKLERİNİN TESPİTİ"** adlı bu çalışma aşağıdaki jüri tarafından **YAPAY ZEKÂ VE VERİ BİLİMİ ANA BİLİM DALI**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı	**Prof. Dr. Hidayet TAKCI** Sivas Cumhuriyet Üniversitesi	……
Jüri Üyesi	**Doç. Dr. Kali GÜRKAHRAMAN** Sivas Cumhuriyet Üniversitesi	……..……
Jüri Üyesi	**Dr. Ögr. Üyesi Şengül BAYRAK** İstanbul Sabahattin Zaim Üniversitesi	……..……..

Bu tez, Sivas Cumhuriyet Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Nevcihan GÜRSOY
FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRÜ

Bu tez, Sivas Cumhuriyet Üniversitesi Senatosu'nun 20.08.2014 tarihli ve 7 sayılı kararı ile kabul edilen Fen Bilimleri Enstitüsü Lisanüstü Tez Yazım Klavuzu (Yönerge)'nda belirtilen kurallara uygun olarak hazırlanmıştır.

ETİK

Sivas Cumhuriyet Üniversitesi Fen Bilimleri Enstitüsü, Tez Yazım Kılavuzu (Yönerge)'nda belirtilen kurallara uygun olarak hazırladığım bu tez çalışmasında;

✓ Bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,

✓ Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,

✓ Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere, bilimsel normlara uygun olarak atıfta bulunduğumu ve atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,

✓ Bütün bilgilerin doğru ve tam olduğunu, kullanılan verilerde herhangi bir değişiklik yapmadığımı,

✓ Tezin herhangi bir bölümünü, Sivas Cumhuriyet Üniversitesi veya bir başka üniversitede, bir başka tez çalışması olarak sunmadığımı; beyan ederim.

16.08.2024

YUSUF AKTAŞ

ÖZET

YAPAY ZEKÂ MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE İŞSİZLERİN ZAMANA GÖRE İŞSİZ KALMA RİSKLERİNİN TESPİTİ

Yusuf AKTAŞ

Yüksek Lisans Tezi

Yapay Zekâ ve Veri Bilim Anabilim Dalı

Danışman: Prof. Dr. Hidayet TAKCI

2024, 156+xvi sayfa

İşsizlik hem oluşum sebepleri hem de sonuçları itibariyle çok boyutlu bir sorun olarak çözüm beklemektedir. İşsizlik sorunun çözümünde mikro düzeyde aktif istihdam politikası aracı olan iş arayan danışmanlığı önemli konumdadır. Danışmanlığın planı ise bireysel eylem planı ile sağlanmaktadır. İyi bir bireysel eylem planı geçerli bir iş profillemesi ile mümkündür. İş profilleme ile iş arayanlar iş arama özelliklerine göre ayırt edilebilir ve sınıflandırılabilir. Gerçekçi bir iş profilleme ise zamana göre işsiz kalma risk tespiti ile gerçekleşebilir. İşsiz kalma risk tespitine dayalı profilleme işlemi ise istatistiksel profillemedir. Bilimsel çalışmalar klasik istatistiksel modellemelere alternatif olarak yapay zekâ makine öğrenmesi modellerini sunmaktadır. Bu çalışmada Sivas İŞKUR'a başvuran işsizlerin bazı iş arama özellikleri üzerinden başvuru sonrası bir yıl içerisinde işe yerleşmeleri istatistiksel yöntemler ve makine öğrenmesi yöntemleri ile modellenmesi, sınıflandırılması, yöntemlerin karşılaştırılması yapılmıştır. Ayrıca işsizlik risk sınıfları ve işe yerleştirmede etkili olan değişkenler belirlenmiştir. Çalışma sonuçlarına göre makine öğrenmesi modelleri istatistiksel modellere göre çoğunlukla daha yüksek doğruluk performansı göstermiştir. Model doğruluk performansında ana belirleyici faktörün model algoritmalarının olduğu tespit edilmiştir. En yüksek doğruluk performansı %78,8 rasgele orman modellemesinde gerçekleşmiştir. En düşük performans ise lojistik regresyon ve naive bayes gibi istatistik tabanlı modellerde gerçekleşmiştir. Risk sınıflandırması için rasgele orman regresyon kullanılmış ve iyi bir açıklama oran(0,48) elde edilmiştir. Sınıflandırmaya göre Sivas da işsiz kalma riski taşıyanların oranı %76,6 olarak hesaplanmıştır. Modeldeki iş arama özelliklerinin işsizin işe yerleşmesi ile çok zayıf ilişkisi olduğu ve başkaca özellikler ile daha iyi bir modelleme yapılacağı tespit edilmiştir. Yapay zekâ

makine öğrenmesi algoritmaları ile yapılan iş profilleme işleminin il işsizlik sorunun çözümüne önemli katkılar sağlayacağı sonucuna varılmıştır.

Anahtar kelimeler: İşsizlik, Danışmanlık, Bireysel Eylem Planı, İş Profilleme, Risk Tespiti, İstatistiksel Modelleme, Makine Öğrenmesi

ABSTRACT

DETECTION OF THE RISK OF UNEMPLOYED PEOPLE BEING UNEMPLOYED BY TIME USING ARTIFICIAL INTELLIGENCE MACHINE LEARNING METHODS

Yusuf AKTAŞ

Yüksek Lisans Tezi
Yapay Zekâ ve Veri Bilim Anabilim Dalı
Danışman: Prof. Dr. Hidayet TAKCI
2024, 156+xvi sayfa

Unemployment awaits a solution as a multi-dimensional problem, both in terms of its causes and consequences. Job seeker consultancy, which is an active employment policy tool at the micro level, is important in solving the unemployment problem. The consultancy plan is provided with an individual action plan. A good individual action plan is possible with a valid job profiling. With job profiling, job seekers can be distinguished and classified according to their job search characteristics. A realistic job profiling can be achieved by determining the risk of becoming unemployed over time. The profiling process based on unemployment risk detection is statistical profiling. Scientific studies offer artificial intelligence machine learning models as an alternative to classical statistical modeling. In this study, the job placement of unemployed people who applied to Sivas İŞKUR within one year after the application, based on some job search characteristics, was modeled, classified and compared with statistical methods and machine learning methods. Additionally, unemployment risk classes and variables effective in job placement were determined. According to the study results, machine learning models generally showed higher accuracy performance than statistical models. It has been determined that model algorithms are the main determining factor in model accuracy performance. The highest accuracy performance was 78.8% in random forest modeling. The lowest performance was achieved in statistical-based models such as logistic regression and naive bayes. Random forest regression was used for risk classification and a good explanation rate (0.48) was obtained. According to the classification, the rate of people at risk of becoming unemployed in Sivas is calculated as 76.6%. It has been determined that the job search

features in the model have a very weak relationship with the job placement of the unemployed and that a better modeling can be done with other features. It has been concluded that job profiling with artificial intelligence machine learning algorithms will make significant contributions to the solution of the provincial unemployment problem.

Anahtar kelimeler: Unemployment, Consultancy, Indıvıdual Actıon Plan, Job Profiling, Rısk Detection, Statistical Modeling, Artificial İntelligence, Machine Learning

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

KISALTMALAR DİZİNİ

İŞKUR	**:**	Türkiye İş Kurumu Genel Müdürlüğü
OR	**:**	ODDS Ratio
TMS	**:**	Türk Meslekler Sözlüğü
YSA	**:**	Yapay Sinir Ağları
YZ	**:**	Yapay Zekâ

1. GİRİŞ

İşsizlik; ağırlıkla yapısal sebepler kaynaklı çeşitli tür ve boyutta olumsuz sonuçları olan bir yaşam gerçekliğidir. Yaşam gerçekliği olması ise işsizlik olgusunun önemini ortaya koymaktadır. İşsizlik olgusunun temelini "işsiz" oluşturmaktadır. İşsiz ise en sade ve sözlük manasıyla "geçinmek için iş bulamama durumu olarak" tanımlanmaktadır. İşsizlik ise işsizlerin bir arada oluşturduğu kurumsal bir yapıyı ifade etmektedir.

İşsizlik, geçmişten günümüze birçok ülkenin, farklı tür ve boyutta çözmesi gereken bir sorunudur. İşsizlik sorunun çözülmesi ile birçok alanda fayda sağlanacaktır. İşsizlik sorunun çözümünde ana belirleyici ve işsizliğin meydana getirdiği veya getireceği olumsuz sonuçların ortadan kaldırılması ve bir anlamda olumlu sonuçlara tebeddül ettirilmesi ancak işsiz bireylerin işe yerleştirilmesi ve istihdamıyla mümkündür. İstihdam en basit ve sözlük anlamıyla "bir insanı bir işte, bir görevde kullanma, çalıştırma anlamına" gelmektedir. İstihdamın sağlanması ise "mikro ve makroekonomik istihdam politikaları" ile gerçekleştirilmektedir. Makroekonomik istihdam politikalarında makro göstergeler üzerinden işsizlik sorunu incelenir. Mikro ekonomik istihdam politikalarında ise mikro düzeyde araçlarla soruna çözüm aranır. Mikro ekonomik istihdam politikaları aktif ve pasif istihdam politikalarıdır. Pasif istihdam politikaları işsizlik sigortası, kısa çalışma ödeneği, yarım çalışma ödeneği, ücret garanti fonu, iş kaybı tazminatı, kıdem ve ihbar tazminatı ve sosyal yardımlar gibi bir takım istihdam sübvansiyonlarıdır. Aktif istihdam politikaları ise danışmanlık hizmetleri, meslek edindirme hizmetleri, işbaşı eğitim programları ve istihdam teşvikleri gibi hizmetlerden oluşmaktadır Aktif istihdam politikalarının merkezinde danışmanlık hizmetleri yer almaktadır. Bir kamu istihdam kurumu olan İŞKUR danışmanlık hizmetlerini "İş ve Meslek Danışmanlığı" adı altında yürütmektedir. Danışmanlık hizmetleri iş arayan danışmanlığı, işveren danışmanlığı, meslek danışmanlığı gibi kısımlara ayrılmaktadır. İş arayan danışmanlığı, iş arayanın istihdam edilmesi sürecinin kurumsallaşmış biçimidir. Bu süreçte işe ve mesleğe yönlendirme, iş arama becerilerinin gelişimi ve işgücü piyasası hakkında bilgilendirme ana faaliyetleri yürütülmektir. Tüm bu faaliyetler bir iş arayan bireysel eylem planı ile yönetilmekte ve yürütülmektedir.

İyi bir bireysel eylem planı, iş arayanın tüm özelliklerine göre profillenmesi ile mümkündür. İş profilleme ile iş arayanlar kişisel, demografik ve iş arama özelliklerine göre ayırt edilebilir ve sınıflandırılabilir. Danışmanlık görüşmesinde danışman iş arayanı danışmanlık teknikleri ile tanır ve profilleme yapabilir. Ancak danışmanlık görüşmesinin muhteviyatı, danışman tecrübesi gibi bir takım nedenler çoğu zaman iyi bir iş profillemesini zorlaştırmaktadır. Bu engelin aşılması için bilimsel tekniklerin kullanılması gerekmektedir. Bu noktada iyi bir iş profilleme işleminin gerçekleştirilmesi iş arayanların iş bulma olasılıklarına göre işsiz kalma riskinin tespit edilmesinden geçmektedir. Başka bir deyişle istatistiksel profilleme işlemi gerçekleştirilmelidir. İstatistiksel profilleme de hedef değişken, işsizin işe yerleşme durumu olabileceği gibi nitelik ve beceri durumu da olabilir. İşsizlerin işe yerleşme durumuna göre yapılan profilleme işleminin temelini ise işsizin işe yerleşmeme riski oluşturmaktadır. Bu bağlamda işsizlerin belirli bir zaman dilim için işsiz kalma risk değerlendirmesinin yapılması gerekmektedir. Başka bir ifadeyle iş profillemenin en önemli parçalarından biri risk değerlendirmesidir. İşsiz kalma riski tespiti ve değerlendirilmesi işleminde iki ana yöntem bulunmaktadır. Bu yöntemler klasik çok değişkenli istatistiksel yöntemler ve yapay zekâ makine öğrenmesi yöntemleridir.

İstatistiksel yöntemler tek değişkenli ve çok değişkenli olmak üzere iki kısıma ayrılabilir. Verinin istatistiksel tekniklerle anlamlı bilgiye dönüştürülme sürecinde değişken sayısının türü istatistiksel tekniğin tek değişkenli mi yoksa çok değişkenli mi olduğunu ortaya koymaktadır. Hayatın olağan akışında bir olay ve sonuç birden fazla faktör ve değişkenin etkisinde meydana geldiğinden çok değişkenli istatistiksel analizler daha fazla önem kazanmaktadır. Hem tek değişkenli hem de çok değişkenli istatistiksel analizler betimsel ve çıkarımsal istatistiksel analizler olarak iki ana gruba ayrılmaktadır. Betimsel istatistiksel analizler veri madenciliğindeki veri keşfine, çıkarımsal istatistiksel analizler ise veri madenciliğindeki modelleme sürecine benzemektedir. Betimsel istatistiksel analizler ile veri görselleştirilir, merkezi eğilim ve yaylım ölçülerine göre tanımlanır. Çıkarımsal istatistiksel analizlerde ise bağımlı ve bağımsız değişkenler arasındaki ilişkiler ortaya konulur, veri setindeki gizli kalan bilgiler açığa çıkartılır ve geleceğe ilişkin tahminlerde bulunulur. Bununla birlikte gözlemler arası kümeleme ve ayırma işlemleri de yapılabilir. Çok değişkenli istatistiksel analizler, analizin amacı ve değişken yapısına göre Kümeleme Analizi,

Faktör Analizi, Path Analizi, Çok Değişkenli Regresyon Analizi, Çok Değişkenli Lojistik Regresyon Analizi, Temel Bileşenler Analizi farklı türlere ayrılmaktadır.

Yapay zekâ, 1950'lerden bu yana bilim insanlarının kafa yorduğu son dönemde ise hemen herkesin ilgi gösterdiği bir disiplindir [1]. Makineleşme süreci bilim dünyasın da Turing tarafından "Makineler düşünebilir mi? sorusunu da beraberinde getirmiştir. Aynı yıllarda, Ordinaryüs Profesör Cahit Arf, Erzurum Atatürk Üniversitesinde "Makineler Düşünebilir mi ve Nasıl Düşünebilir?" konulu çalışmasını sunmuştur [2]. Yapay zekânın amacı; insan gibi düşünen, onun gibi davranan, karar verme yeteneği olan yazılım ve donanım sistemleri geliştirmektir. Yapay zekâda sadece otonom hareket eden robotlarda değil aynı zamanda gelişmiş karar destek sistemlerinde karşımıza çıkmaktadır [1]. Yapay zekâda kat edilen yüksek gelişimler ve ilerlemeler yapay zekâya olan ilgiyi ve yapay zekâ üzerine yapılan araştırma ve uygulamaları da arttırmıştır. Bu durum yeni yapay zekâ tekniklerinin geliştirilmesine imkân tanıdığı gibi klasik bazı tekniklerin de yapay zekâ uygulamalarına entegrasyonunu da sağlamıştır. Yapay zekâ uygulamaları doğal dil işleme, görüntü işleme, makine öğrenmesi ve uzman sistemlerdir.

Bir yapay zekâ uygulaması olan makine öğrenmesi en basit tanımıyla mevcut veriden makinelerin bir takım öğrenme teknikleri ve algoritmaları ile eğitilerek tahmin edici ve açıklayıcı bilgiler üretmesi işlemidir. Bir başka tanımda makine öğrenimi, makinelerin mevcut verilerden öğrendiği ve kendi kendine öğrenip geliştirdiği bir konsept üzerinde çalışır. Büyük verilerin yansıra geçmiş deneyimlerden gelen algoritmayı ifade eder [3]. Makine öğrenmesinin temelini matematik, istatistik, bilgisayar ve yazılım gibi bilimler oluşturmaktadır. Makine öğrenmesi yöntemleri günümüzde ekonomi, sağlık, hukuk, bilgi iletişim, tarım olmak üzere tüm sektörlerde kullanılabilir durumdadır.

Makine öğrenmesi denetimli, denetimsiz ve yarı denetimli ve takviyeli öğrenme olarak ayırmak mümkündür. Makine öğrenmesi algoritmalarının kullanımı makine öğrenmesinin türüne ve model seçimine göre değişmektedir. Bazı algoritmalar hem denetimli öğrenme hem denetimsiz öğrenme için kullanıma elverişli iken bazıları ise sadece denetimli/denetimsiz öğrenime elverişlidir. Yine bazı algoritmalar ile hem sınıflama hem de regresyon gerçekleştirilirken bazı algoritmalar ise sadece sınıflandırma işlemine olanak sağlamaktadır. Denetimli makine öğrenmesi

sınıflandırma işlemi lojistik regresyon, rasgele orman, destek vektör makinaları, k en yakın komşu, naive bayes ve karar ağaçları gibi makine öğrenmesi algoritmaları ile regresyon işlemi ise basit doğrusal regresyon, çok değişkenli regresyon, rasgele orman regresyon, destek vektör regresyon, polinom regresyon ve karar ağaçları regresyon gibi çeşitli regresyon türleri ile gerçekleştirilir.

Bu çalışmada 2022 yılı Eylül, Ekim ve Kasım aylarında Sivas İŞKUR'a iş için başvuran işsizlerin kişisel, demografik ve işgücüsel bazı bilgileri üzerinden başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmemeleri klasik istatistiksel yöntemler ve yapay zekâ makine öğrenmesi yöntemleri ile modellenmesi, sınıflandırılması, yöntemlerin karşılaştırılması, işsiz kalma risk sınıflarının ve işe yerleşmede etkin değişkenlerin belirlenmesi gerçekleştirilmiştir. Çalışmanın iki ana sorusu/problemi bulunmaktadır, Bunlardan birincisi; "İşsizlerin zamana göre işsiz kalma risklerinin tespitinde yapay zekâ makine öğrenmesi yöntemleri, istatistiksel yöntemlere göre daha mı etkili?" olduğu ikincisi ise "İşsizlerin işe yerleşmesinde etkin faktörlerin/değişkenlerin neler olduğu?" dur. Ayrıca çalışmanın ana soruları altında makine öğrenmesi algoritmaların, model eğitim yöntemleri ve veri setine göre model performansları da irdelenmiştir

Tez çalışması üç bölümden oluşmaktadır. Birinci bölümde tez çalışmasının genel teşekkülü hakkında bir giriş yapılmış olup tez çalışmasının problemi, konusu, amacı, önemi, kapsamı, kısıtları, varsayımları, veri kaynağı, veri kitlesi ve önceki çalışmalara ilişkin literatür taraması, işsizlik ve işsiz kalma riski değerlendirmesine yer verilmiştir. Çalışmanın materyal ve metot başlıklı ikinci bölümünde ise yapay zeki teknolojileri, yapay zekâ uygulamaları, veri madenciliği ve makine öğrenmesine ilişkin tanımlayıcı ve açıklayıcı bilgilere yer verilmiştir. Bulgular başlıklı çalışmanın üçüncü bölümünde ise tez çalışmasına ilişkin bulgular yer almaktadır. Çalışmanın son bölümü olan dördüncü bölümde ise çalışma bulgularına ilişkin tartışma ve sonuçlar yer almaktadır. Tez çalışması muhteviyatı ile çalışma ekonomisi, istatistik, yapay zekâ, veri bilimi ve yazılım bilimleri ortak olarak kullanılarak çok disiplinli bir yapı arz etmektedir. Bu bağlamda çalışma hem yapay zekâ hem de veri bilimi alanlarına doğrudan ilişkili bulunmaktadır.

1.1 Çalışmaya Ait Genel Bilgiler

Çalışma, Sivas ilinde yaşayan ve 2022 yılı Eylül, Ekim ve Kasım aylarında İŞKUR'a iş için başvuru yapan işsizlerin birbiriyle ilişkili olduğu varsayılan kişisel, demografik ve işgücüsel bazı bilgileri üzerinden başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmemeleri klasik istatistiksel yöntemler ve yapay zekâ makine öğrenmesi yöntemleri ile modellenmesi, sınıflandırılması, yöntemlerin birbirine olan üstünlüklerinin ortaya konması, işsiz kalma olasılıklarının ve risk gruplarının belirlenmesi şeklinde gerçekleştirilmiştir. Çalışmanın genel çerçevesi aşağıda yer alan akış diyagramına göre şekillendirilmiştir.

Şekil 1.1 İşsizlik Sorunun Çözümüne İlişkin Mikro Düzeyde Akış Diyagramı

İşsizlik sorunun çözümünde mikro düzeyde aktif istihdam politikası aracı olan iş arayan danışmanlığı önemli konumdadır. Danışmanlığın planı ise bireysel eylem planı ile sağlanmaktadır. İyi bir bireysel eylem planı geçerli bir iş profillemesi ile mümkündür. İş profilleme ile iş arayanlar iş arama özelliklerine göre ayırt edilebilir ve sınıflandırılabilir. Gerçekçi bir iş profilleme ise zamana göre işsiz kalma risk tespiti ile gerçekleşebilir. İşsiz kalma risk tespitine dayalı profilleme işlemi ise istatistiksel profillemedir. Bilimsel çalışmalar klasik istatistiksel modellemelere alternatif olarak yapay zekâ makine öğrenmesi modellerini sunmaktadır. Bu bağlamda tez çalışması da işsizlik sorunun çözümünde önemli bir katkı sağlayacağı ön görülmektedir. Özellikle çalışma çıktılarının bir karar destek sistemi olarak iş arayan danışmanlığına entegre edilmesi veya daha da geliştirilerek yapay zekalı bir danışmana başka bir ifadeyle e-danışmanlık hizmetini dönüştürülmesi tez çalışmasının önemini daha da arttırmaktadır.

Çalışmanın ana amacı işsizlik sorununa bilim ve teknolojinin yardımıyla yeni çözümler geliştirmektir. Çalışmanın bu ana amacına ulaşmak için birtakım ara amaçlarda belirlenmiştir. İşsizlerin İŞKUR'a başvuru gerçekleştirdikten sonraki bir yıllık süre içerisinde işsiz kalma risklerinin klasik istatistiksel yöntemlerle ve yapay zekâ makine öğrenmesi yöntemleriyle tespit edilmesi ve birbirlerine olan üstünlüklerinin karşılaştırılması, işsizlik risk tespitinde en uygun yöntemin ve tekniğin belirlenmesi ve bu tespitlere göre ideal iş arayan danışmanlık uygulamasına kaynak teşkil edecek analize dayalı bilginin elde edilmesi ara amaç olarak belirlenmiştir.

Çalışmanın amacına uygun olarak iki ana sorusu/problemi bulunmaktadır, Bunlardan birincisi; "İşsizlerin zamana göre işsiz kalma risklerinin tespitinde yapay zekâ makine öğrenmesi yöntemleri klasik istatistiksel yöntemlere göre daha mı etkili?" olduğu ikincisi ise "İşsizlerin işe yerleşmesinde etkin faktörlerin/değişkenlerin neler olduğu?" dur. Ayrıca tez çalışmasının ana soruları altında makine öğrenmesi algoritmalarının ve model geçerleme yöntemlerinin, yığın ve örneklem verilerine göre oluşturulan modellerin birbirine üstünlüğü alt problemler olarak irdelenmiştir. Bu sorulara uygun olarak tez çalışmasının konusu "Yapay Zekâ Makine Öğrenmesi Yöntemleriyle İşsizlerin Zamana Göre İşsiz Kalma Risklerinin Tespiti" olarak belirlenmiştir.

Tez çalışmasının sorusuna/problemine bağlı olarak çalışmanın varsayımları şekillenerek iki ana varsayımdan oluşmuştur. Bu varsayımlardan birincisi; İŞKUR'a başvuran işsizlerin, başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmeme durumu ile işsizin kişisel, demografik ve işgücüsel bilgileri arasında istatistiksel anlamlı bir ilişkinin olup olmadığı,

H_{10}=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki yoktur.
H_{11}=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki vardır.
İkincisi; İŞKUR'a başvuran işsizlerin, başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmeme doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir farkın olup olmadığı;

H_{20}=İşsizin işe yerleşmesinin doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir fark yoktur.

H_{21}=İşsizin işe yerleşmesinin doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir fark vardır.

Çalışmanın coğrafi, zamansal ve öz niteliklere göre kapsamı ve kısıtlamaları bulunmaktadır. Ülke genelindeki tüm işsizlerin zamana göre işsiz kalma risklerinin tespitinin oldukça güç ve maliyetli olması aynı zamanda doğru araştırma verisinin elde edilmesinin muhal olması nedeniyle tez çalışmasında sadece İŞKUR aracılığıyla Sivas ilinde iş arayan işsizler kapsama alınmıştır. Bununla beraber işsizin İŞKUR'a başvurusundan sonra geçen bir yıllık zaman içerisinde iş bulma durumunun tespit edilmesi için çalışmaya analiz döneminden bir yıl önceki başvuran işsizler yani 2022 yılı Eylül, Ekim ve Kasım döneminde başvuran işsizler kapsama alınmıştır. Başka bir ifadeyle yıl boyu başvuran işsizlerin tamamı dikkate alınmamıştır. Ayrıca İŞKUR'a başvuran işsizlerin sosyal/kişisel durumu, cinsiyet, yaş, eğitim, meslek, medeni durum, başvuru türü, sosyal yardım alma durumu, işsizlik maaşı alma durumu ve istihdam durumuna göre bilgiler yer almasına rağmen işsizin iş tecrübesi, mesleki bilgi, beceri ve tecrübesi, iş arama bilgi, beceri ve tecrübesi, iş arama sıklığı, işe olan ihtiyacı, isteği gibi birtakım iş arama donanımı ve iş arama davranışına dair bilgileri İŞKUR iş arayan kayıt bilgilerinde olmadığı için çalışmaya dâhil edilememiştir. Bu noktada çalışmanın yıl içerisinde başvuru gerçekleştiren ülke genelindeki tüm işsizleri ve tüm işsizlerin demografik ve işgücüsel özelliklerini kapsamamış olması ve işsizlerin bir yıl öncesinde ve bir yıl sonrasındaki işsiz kalma risklerinin ve yeniden istihdam edilmelerini tespit edememiş olması tez çalışmasının kısıtları olarak değerlendirilmektedir.

1.2 Önceki Çalışmalar ve Literatür Taraması

Bu tez çalışmasında işsizlerin İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşme/ yerleşmeme durumuna göre işsiz kalma risklerinin tespiti amaçlanmıştır. Ancak tez çalışmasının bir kamu istihdam kurumu olan İŞKUR'a başvuru ile sınırlandırılması literatür taramasını kısıtlayacağından, işsizlerin iş araması ile başlayan süreç içeresinde işsiz kalma riski tespiti ile ilişkili yapılar üzerinden literatür taraması yapılmıştır. Dolayısıyla iş profilleme gibi yapılara da değinilmiştir. Literatür taraması Google akademik ve YÖK tez veri tabanları üzerinden yapılmıştır.

Google akademik üzerinden yabancı kaynaklardan "work profiller, statistical profil unemployment, machine learning unemployment, risk prediction unemployment" kelimeleri üzerinden araştırma yapılmış ve 48 adet konuyla doğrudan ve dolaylı ilgili çalışma tespit edilmiştir. Yapılan çalışmalar ağırlıkla ülkelerin kamu istihdam kurumları aracılığıyla gerçekleştirdiği istatistiksel ve yapay zekâlı profilleme araçlarının etkileri, uygulanabilirliği, olumlu ve olumsuz sonuçları, eksiklikleri, avantaj ve dezavantajları noktasında ağırlık göstermektedir.

YÖK Tez ve Google akademik üzerinden Türkçe kaynaklardan "iş profilleme, istatistiksel profilleme, makine öğrenmesi iş profilleme, makine öğrenmesi işsizlik, makine öğrenmesi, istihdam, yapay zekâ işsizlik ve istihdam, risk ve işsizlik" kelimeleri üzerinden yapılan araştırmalarda istatiksel/yapay zekâ metotları ile işsizlerin profillemesine ilişkin sadece bir çalışmaya rastlanmıştır. Bu çalışmada Emeç ve arkadaşları "Türkiye İstatistik Kurumunun 2014 - 2017 yılları arasında yaptığı Gelir ve Yaşam Koşulları Araştırması Anketi Panel Verileri kullanılmıştır. Anket verilerine göre 15 - 29 yaş aralığında örneklemdeki 38.109 gencin 23.375'i işsizdir. Bireyin işsizliğini etkileyen faktörler fert ve hane bazında, tesadüfi etkiler dengesiz panel logit model kullanılarak analiz edilmiştir. Analiz bulgularına göre kadınların erkeklere göre, evli olanların bekârlara göre işsiz olma olasılığının daha fazla olduğu ortaya çıkmıştır [4], sonucuna varmışlardır.

Bu literatür taramasına ek olarak tez çalışması konusuyla doğrudan alakalı kurum olan İŞKUR uzmanlık tezleri incelenmiş ve tez çalışması konuyla doğrudan alakalı bir uzmanlık tezi tespit edilmiştir. "Profil temelli danışmanlık kapsamında bireysel eylem planlarının oluşturulması: dünya uygulamaları ve Türkiye simülasyonu" isimli çalışmada İŞKUR'a belirli bir dönem de kayıt yaptıran iş arayanların kayıt sonrası "6 Aydan Uzun Süreli İşsiz Olma Durumu" birçok değişkenli istatistiksel yöntem olan lojistik regresyon analizi ile modellenmiştir. Çalışma Kapsamında iş arayanların istatistiksel modelle profillenmesi aşamasında ikili lojistik regresyon modeli kullanılmıştır. Bağımlı değişken iş arayanların kuruma kayıt olduktan sonraki ilk işe girene kadar geçen sürenin 6 ay dan (180 gün) fazla olup olmamasına göre alınmıştır [5]. Çalışmada 12 adet bağımsız değişken kullanılmış olup bu değişkenlerden dokuz tanesi doğrudan ve üç tanesi dolaylı olarak iş arayan ilgilidir. Çalışma kapsamında %31,8 açıklama oranı ve %72,2 doğru sınıflama oranı elde edilmiştir. Oluşturulan model %95 güven düzeyinde anlamlıdır. Çalışma kapsamında dörtlü risk

sınıflandırması oluşturulmuş ve iş arayanların %5,2'si çok yüksek işsiz kalma riski taşıdığı sonucuna varılmıştır. Çalışma sonucunda bir takım önerilerde bulunulmuş ve bu önerilerden en fazla göze çarpanı "İstatistiksel modellerden İŞKUR'un azami düzeyde faydalanmasını sağlayacak modeller geliştirilmesi ve kullanılmasını teşvik etmesi [5]" olarak tespit edilmiştir. Aynı zamanda çalışmanın bir diğer önemli sonucu ise "Çalışma kapsamında uygulanan modelin bağımsız değişkenler tarafından ne kadarının açıklandığını ifade eden R^2 değeri % 31,8 olarak bulunmuştur. Bu oran 6 aydan uzun süreli işsiz olma durumunu kestirmek için daha fazla açıklayıcı değişkene ihtiyaç duyulduğunun bir göstergesidir [5]". Bu çalışma alanın da ilk olması, doğrudan ilgili Kurum ve uzman personeli eli ile yürütülmesi ve iyi seviyede açıklama ve doğruluk performansı göstermesi açısından büyük önem arz etmektedir. Ancak çalışma sonuçlarında da bahsedildiği üzere çalışmanın hem modele girecek değişkenler yönünden hem de modelleme tekniği yönünden iyileştirilmesi ve geliştirilmesi gerekmektedir. Bu tez çalışmasında model tekniği yönünden yapay zekâ makine öğrenmesi algoritmaları bu gelişimin sağlanmasına imkân tanımaktadır.

Tablo 1.1 OECD Genelinde İstatistiksel Profil Oluşturma Modellerinin Özellikleri

Ülkeler	Model
Avustralya	Lojistik Regresyon
Avusturya	Lojistik Regresyon
Belçika	Random Forest Model
Danimarka	Büyük Veri Model
İrlanda	Probit Regresyon
İtalya	Lojistik Regresyon
Letonya	Faktör Analizi
Hollanda	Lojistik Regresyon
Yeni Zelenda	Random Forest (LET), Gradient boosting (SEM)
İsveç	Lojistik Regresyon
Amerika	Lojistik Regresyon

OECD tarafından yayımlanan "Herkese açık istatistiksel profil oluşturma istihdam hizmetleri: Bir uluslararası karşılaştırma *(Statistical Profiling in Public Employment Services: An international comparison)*"isimli çalışmaya göre OECD ülkeleri genelinde ağırlıkla bir istatistiksel metot olan lojistik regresyon modellemesi kullanılmaktadır. Yine aynı rapora göre daha gelişmiş metot olan yapay zekâ makine öğrenmesi algoritmaları kullanan Belçika ve Yeni Zelanda gibi ülkeler bulunmaktadır [6].

1.2 İşsizlik ve İşsiz Kalma Riski Değerlendirmesi

İşsizlik; ağırlıkla yapısal sebepler kaynaklı oluşan, bireysel, ailesel ve toplumsal çeşitli tür ve boyutta olumsuz sonuçları olan bir yaşam gerçekliğidir. Yaşam gerçekliği olması ise işsizlik olgusunun önemini ortaya koymaktadır. İşsizlik olgusunun temelini "işsiz" oluşturmaktadır. İşsiz ise en sade ve sözlük manasıyla "geçinmek için iş bulamama durumu olarak" tanımlanmaktadır. İşsizlik, işsizlerin bir arada oluşturduğu kurumsal bir yapı veya bir olgudur. İşsiz somut olarak bir işi olmayıp iş arayarak iş bulamayanları ifade ederken işsizlik ise tüm işsizlerin oluşturduğu soyut bir yapıyı ifade etmektedir. Bu soyut yapı işsizlik nedenleri, işsizlik etkileri, işsizlik sonuçları ve işsizliğin boyutları gibi kavramlar ile somutlaştırılmaya çalışılmaktadır. İşsizliğin daha somut bir şekilde gösterimi ise işsizlik oranı, işsiz sayısı gibi temel işgücü göstergeleri üzerinden doğrudan yapılmaktadır. Ayrıca işsizlikle ilintili dolaylı göstergeler ile de bu somutlaştırma işlemine destek sağlanmaktadır.

İşsizliğin oluşmasında birden çok faktör bulunmaktadır. Genel olarak işsizlik işgücü arzı ve işgücü talebi kaynaklı mikro ve makro boyutta nedenlere dayanmaktadır. Nüfus, göç, eğitim ve meslek gibi nedenler işgücü arzı kaynaklı işsizlik nedenleri iken ekonomik ve teknolojik gelişmeler, istihdam maliyetleri, küresel gelişmeler ise işgücü talebi kaynaklı işsizlik nedenleridir. İşsizliğin oluşum nedeni, yoğunluğu ve etki süresindeki farklılaşma işsizlik türlerini iradi-gayri iradi, açık-gizli işsizlik gibi ana türelere ayırmaktadır. Bu ana ayırıma ek olarak açık işsizlik türü de arızı, yapısal, mevsimsel, konjoktürel ve teknolojik işsizlik olmak üzere alt türelere ayrılmaktadır. İradi işsiz piyasadaki cari ücret seviyesinde çalışmayı kabul ederken gayri iradi işsiz ise kabul etmemektedir. İşgücü piyasasında istihdam da bulunmalarına karşın mal ve hizmet üretimine katkı sağlamayan bireyler gizli işsizlerdir. Açık işsizlikte eğitim, meslek gibi yapısal nedenler, ekonomik kriz veya daralma gibi konjoktürel nedenlerle, üretimde robotlaşma gibi teknolojik nedenler iş arayanların işsiz kalmalarına neden olmaktadır. İşsizlik nedenleri ve türlerinde olduğu gibi sonuçları itibariyle de çok çeşitlik göstermektedir. İşsizlik en küçük birim olan işsizden en büyük yapılanma olan topluma kadar birçok olumsuz sonuçlar içermektedir. Bu olumsuz sonuçlar mikro boyutta, işsiz bireyin çalışma istek ve arzularının ve çalışma yeteneklerinin azalmasına ve zamanla kaybolmasına, gelir kaybı ile ekonomik sıkıntılar yaşamasına ve suç işlemeye yönelme şeklinde kendini gösterirken makro boyutta ise ülkeyi ekonomik krizlere ve politik çıkmazlara, hukuki uyuşmazlıklarda artışa, yetersiz eğitim

faaliyetlerine sürükler. İşsizliğin bu denli olumsuz etkileri toplumsal barış, huzur ve sükûnu bozarak hem bireysel hem toplumsal hem de ülkesel gelişime ket vurur. İşsizlik sorunun kısa ve orta vadede çözülmemesi ise sorunu kronik hale getirerek büyütür ve daha içinden çıkılmaz bir yapıya dönüştürür. Bu haliyle işsizlik sorunu ekonomik, sosyal, siyasal, hukuksal ve eğitime ilişkin birçok ana yapı üzerinden meydana gelecek değişim ve gelişimlerle çözüm beklemektedir.

İşsizlik sorunun olağan dönemlerde çözümü ise iki şekilde gerçekleşir. Birinci olması istenmeyen bir gerçeklik olarak işsizlerin işgücü piyasasından çekilmesidir. İşsizlerin iş bulma ümitlerini kaybederek işgücü piyasasından çekilmesi durumunda nicel anlamda işsizliği düşürmesine karşın, sorunun ötelenerek daha büyük olumsuzluklara neden olma riskini taşımaktadır. İkinci ise işsizlerin istihdam edilmesidir. İstihdam ise gerçek anlamda işsizliği çözen ve olması istenen bir çözümdür. İstihdam en basit ve sözlük anlamıyla "bir insanı bir işte, bir görevde kullanma, çalıştırma anlamına" gelmektedir. İstihdam geniş anlamda üretim faktörlerinin üretime dâhil edilmesini, dar anlamda ise emeğin arz edilmesi anlamına gelmektedir. İstihdam işleminin gerçekleştirilmesinde birtakım ana ve alt faktör belirleyici ve etkileyici konumdadır. İstihdamın gerçekleşmesinde üç ana faktör bulunmaktadır. Bu ana faktörler iş arayan, açık iş ve işgücü piyasasıdır. Başka bir ifadeyle işgücü arzı, işgücü talebi ve eşleşme ortamıdır.

İşsizlerin istihdam işleminin gerçekleştirilmesi "mikro ve makroekonomik istihdam politikaları" marifetiyle gerçekleştirilmektedir. Makroekonomik istihdam politikaları, para, maliye ve gelir politikasıdır. Makroekonomik istihdam politikalarında esasında istihdam, enflasyon, büyüme gibi makroekonomik göstergeler üzerinden işsizlik sorunu incelenir ve yine makro düzeydeki araçlar ile işsizlik sorununu çözümüne ilişkin politikalar geliştirilir. Mikro ekonomik istihdam politikaları ile mikro ölçekte soruna çözüm aranır. Mikro ekonomik istihdam politikaları ise aktif ve pasif istihdam politikaları olarak ikiye ayrılmaktadır. Pasif istihdam politikaları işsizlik sigortası, kısa çalışma ödeneği, yarım çalışma ödeneği, ücret garanti fonu, iş kaybı tazminatı, kıdem ve ihbar tazminatı ve sosyal yardımlar gibi bir takım istihdam sübvansiyonlarıdır. Aktif istihdam politikaları ise danışmanlık hizmetleri, meslek edindirme hizmetleri, işbaşı eğitim programları ve istihdam teşvikleri gibi hizmetlerden oluşmaktadır. Pasif istihdam politikalarının amacı ağırlıkla istihdamı koruyucu, işsizliği önleyici ve işsizlik sonrası gelir kaybını azaltıcı faaliyetlerden oluşturmaktadır.

Mikro düzeyde ve aktif bir şekilde istihdamın korunması ve geliştirilmesi işlemi aktif istihdam politikası araçları marifetiyle gerçekleşmektedir. Aktif istihdam politikası araçlarının hedef kitlesi işgücü arzının ve talebinin ana birimleri olan iş arayanlar ve işverenlerdir. Aktif istihdam politikası ile işgücü piyasasındaki mesleksizlik, tecrübesizlik, iş ve eleman arama davranış yetersizliği kaynaklı işsizlik sorununa çözüm üretilmesi amaçlanmaktadır. Bu kapsamda iş arayan, işveren ve öğrencilere danışmanlık hizmeti, işyerlerinde işbaşı eğitim programı, meslek edindirme programları düzenlenmektedir. Ancak önemle belirtmek gerekir ki aktif istihdam politikalarının merkezinde tüm hizmetlerle ilişkili yapıda olan danışmanlık hizmetleri yer almaktadır.

Danışmanlık hizmetleri aktif istihdam politikalarının kalbi hükmünde olup hizmet yürütümünün ana omurgasının oluşturmaktadır. Danışmanlık hizmetleri; Bir danışan tarafından ihtiyaç duyulan bilginin, bir danışman tarafından belirli bir süre içerisinde nitelikli ve kurumsal olarak verilmesi işlemidir. Burada danışan bilgi hakkında hiç veya gereği kadar bilgi sahibi olmaması, danışmanın ise tam veya gereği kadar bilgi sahibi olması durumu vardır [7]. İş arayan, eleman arayan ve mesleki gelişim çabasında olan danışan farklılığı danışmanlık hizmetinin çeşitlenmesine sebep olmuştur. Bu çeşitlilik iş arayan danışmanlığı, işveren danışmanlığı, meslek danışmanlığı, iş kulübü lideri, engelli koçu gibi danışmanlıkta ihtisaslaşmayı birlikte getirmiştir. Dünyada olduğu gibi ülkemizde bir kamu istihdam kurumu olan İŞKUR danışmanlık hizmetlerini "İş ve Meslek Danışmanlığı" hizmetleri adı altında gerçekleştirmektedir.

İş arayan danışmanlığı faaliyetinde danışanlar iş arayanlar, danışmanlar ise iş ve meslek danışmanlarıdır. İş arayan ile iş ve meslek danışmanı arasında istişare edilen ana mevzular ise iş arayanın işe yerleşmesinde belirleyici olan ve etkileyen faktörlerdir. Bu faktörler iş arayanın işe olan ihtiyacı, iş arama ve çalışma isteği, iş arama donanımı ve iş arama davranışıdır. İş arayan danışmanlığı hizmeti kapsamında iş arama statüsüne göre işe ve mesleğe yönlendirme, iş arama becerilerinin gelişimi ve işgücü piyasası hakkında bilgilendirme ana faaliyetleri yürütülmektir. İş arayan danışmanlarının en önemli görevi ise iş arayanın iş arama davranışının geliştirilmesi, güncel tutulması ve hazırlanan eylem planına göre iş arama süreci takibinin yapılmasıdır. Burada iş arama eylem planı/bireysel eylem planı iş arayan danışmanlığının ana omurgasını oluşturmaktadır.

Bireyin ihtiyaçlarını, becerilerini, fiziksel ve ruhsal olarak çalışma hayatına hazır olup olmadığını, sosyal statüsünü vb. özelliklerini birlikte değerlendirerek uygun faaliyetin gerçekleştirilmesi doğrultusunda bir yol haritası düzenleme ve takip ederek sonuca ulaşma olgusunun bütünü "Bireysel Eylem Planı" olarak tanımlanmaktadır [5]. Bireysel eylem planı ile iş arayanın iş arama süreç yönetimi planlanmaktadır. Bu plan kapsamında iş arayanın iş arama sürecindeki yapması gereken faaliyetler ile alması gereken iş arayan danışmanlık hizmetleri planlanmaktadır. Planın hem oluşturulmasında hem de icrasında iş arayan ve danışman birlikte hareket etmelidir. Planın etkin ve verimli uygulanması iş arayanın şartlarına uygun bir işe kısa sürede yerleşmesine olanak sağlayacaktır. Plana uyulmaması durumunda ise iş arayan beklenti ve tercihleri dışında bir işe girme veya uzun süre işe girmeme riski oluşacaktır. Bu bağlamda bireysel eylem planın temelini iş arayanının mevcut iş arama özelliklerine göre işgücü piyasasında istihdam edilip edilmeyeceği konusu/riski oluşturmaktadır. Bu noktada işgücü piyasasında iş arayanların iş bulma olasılığının bilinmesi iş arayanın işsiz kalma riskinin belirlenerek profillemesine ve bu profile göre iyi bir eylem planın hazırlanmasına olanak sağlayacaktır.

Profil kavramı sözlük anlamına göre, "bir kişi veya eşya için ayırt edici özelliklerin bütünü [8]" anlamına gelmektedir. Profil aynı zamanda görüntü anlamına da gelmektedir. Bu tanımlara göre profilleme ayırt edici özelliklere göre görüntüleme anlamına gelmektedir. Profilleme de amaç profillenen şeyin ayırt edici özelliklerinin ortaya çıkarılmasıdır. İş profilleme ise bir iş için gerekli nitelik ve beceriye göre ayırt edilmesi veya bu iş için gerekli nitelik ve beceriye uygun iş arayanların ayırt edilmesi olarak tanımlanabilir. KİK'ler(Kamu İstihdam Kurumları) nezdinde profilleme ise; hizmet sunulan tüm bireylerin çalışma hayatı bakımından ayırt edici özelliklerinin tanımlanması olarak değerlendirilmektedir [5]. İş arayan profilleme ile iş arayanlar kişisel, demografik ve iş arama özelliklerine göre ayırt edilebilir ve sınıflandırılabilir. Bununla birlikte iş arayanlar işsiz kalma risklerine göre sınıflandırılabilir. Bu sınıflandırma uygun olarak bireysel eylem planı hazırlanarak iş arayan danışmanlık hizmetleri iş arayana sunulur. Böylece iş arayanın daha hızlı bir sürede işe yerleşmesi sağlanabilir. İş arayan profilleme de istatistiksel profilleme, kural tabanlı profilleme, soft profilleme, danışman yaklaşımıyla profilleme gibi yöntemler bulunmaktadır [5] İstatistiksel profilleme de hedef değişken, işsizin işe yerleşme durumu olabileceği gibi nitelik ve beceri durumu da olabilir. İşsizlerin işe yerleşme durumuna göre yapılan

profilleme işleminin temelini ise işsizin işe yerleşmeme riski oluşturmaktadır. Bu bağlamda işsizlerin belirli bir zaman dilim için işsiz kalma risk değerlendirmesinin yapılması gerekmektedir.

Risk sözlük anlamı olarak, zarara uğrama tehlikesi [8] iken terimsel anlam olarak kullanım alanına göre de şekillenmektedir. Örneğin çalışma hayatında risk, iş arayanın gerekli iş arama davranışı sergilememe tehlikesi ile işsiz kalma ihtimali olarak tanımlanabilir. Riske neden olan sebepler, istenmeyen ve zarara neden olan tehlikelerdir. Risk aynı zamanda bir yönetim süreci olup, tanımlama, ölçme ve analiz etme, değerlendirme, çözüm önerileri, uygulama ve iyileştirme gibi aşamalardan oluşmaktadır. Risk değerlendirmesi risk yönetim sürecinin bir parçası olup tehlikenin bir risk teşkil edip etmeyeceği, risk teşkil ederse riskin boyutunun ne olduğu etraflıca irdelenir ve değerlendirilir. Birçok risk değerlendirme yöntemi bulunmakla beraber bunlardan en sık kullanılanı karar matrisidir. Karar matrisi ile olasılık ve şiddet üzerinden yapılan bir puanlama ile risk değerlendirmesi işlemi yapılır. İşsizin işe yerleşme olasılığı ile danışmanın kanaati üzerinden gerçekleştirilecek puanlama karar matrisine örnek olarak verilebilir. İşsiz kalma risk değerlendirilmesinde iş arayanın işsiz kalma riskine neden olan tehlikelerin önemi ve boyutları ortaya konularak etkin tehlikeler belirlenir ve kıymetlendirilir. Bu kıymetlendirme işlemi ağırlıkla modelleme üzerinden gerçekleştirilir.

İşsiz kalma risk değerlendirilmesi işleminde gerekli olan modelleme işleminin yapılmasında iki ana yöntem bulunmaktadır. Bu yöntemler klasik olasılık hesaplarına dayanan çok değişkenli istatistiksel yöntemlerle beraber farklı türlü matematiksel ve istatistiksel hesaplara sahip yapay zekâ makine öğrenmesi algoritmalarıdır. Klasik çok değişkenli istatistiksel yöntemlerin uygulanmasında birçok varsayımın olması ve mevcut sorunun çözümünden daha çok analizin uyum iyiliğine odaklanılması gibi nedenler yapay zekâ makine öğrenmesi metotlarını daha da ön plana çıkarmaktadır. Yapay zekâ makine öğrenmesi metotlarının herhangi bir varsayıma tabi tutulmaması ve doğrudan kullanılmaya ve sonuç almaya elverişli olması bu metotların çok değişkenli istatistiksel yöntemlerin yerine sıklıkça tercih edilmesi durumunu ortaya çıkarmıştır. Ayrıca makine öğrenmesi metotlarının daha doğru sınıflandırma işlemi gerçekleştirmesi bu metotların kullanılmasını daha da öncelemektedir.

2. MATERYAL VE METOD

Tez çalışmasında “Çok Değişkenli İstatistikler Yöntemler” ve “Yapay Zekâ Makine Öğrenmesi Yöntemleri” olmak üzere iki ana modelleme tekniği kullanılmıştır. Bu iki ana tekniğin etkin kullanılması için ise “Veri Madenciliği Süreci” işletilmiştir Dolayısıyla üç ana teknik materyal ve metot olarak çalışmanın ana omurgasını oluşturmaktadır. Bu bölümde yapay zekâ ve makine öğrenmesi, veri madenciliği ve çok değişkenli istatistiksel yöntemler hakkında bilgi verilmiş olup bu üç yöntemin birbiriyle ilişkileri ve etkileşimleri irdelenmeye çalışılmıştır.

2.1 Yapay Zekâ, Veri Madenciliği Ve Makine Öğrenmesi

Kâinatın merkezine hayat, hayatın merkezine ise insan konumlandırılmıştır. İnsan özel ve yüksek bu konumu itibariyle konumuna uygun maddi ve manevi cihazlarla donatılmıştır. Bu cihazların varlığı insanda maddi ve manevi bir gelişmişliği gerekli kılmaktadır. İnsanın bu ikiz gelişmişlik süreci birbirini tamamlayıcı niteliktedir. İlerlemenin ve gelişimin temelini bilgi oluşturmaktadır. Bilgiye ulaşmanın ise akıl ve nakil başta olmak üzere farklı türde kaynakları bulunmaktadır. İnsan nakil yolu ile edinilmiş tecrübe ve bilgilere ulaşırken akıl vasıtasıyla doğruyu yanlıştan ve gerçeği yalandan ayırarak yine bilgiye ve hakikate ulaşır. Doğru ve gerçekçi bilgiye ulaşımda akıl zekâ ile etkileşim halindedir. Zekâ ile insan olayları tanımlar, anlamlandırır, ölçer ve somut çözümler üretir. Bu noktada bilgiye ulaşımda akıl ve zekâ birbirinden ayrılmaz birliktelik arz ederek insanın fikir ve düşünce dünyasının temelini oluşturur.

Bilgi kimi zaman hazır ve paket halinde insanın fikir ve düşünce alanına gelirken kimi zaman ise henüz işlenmemiş ham veri şeklinde işlenmeyi beklemektedir. Esasında bilginin kaynağı veridir. Veri işlenmiş ve işlenmemiş veri olarak iki kısıma ayrılmaktadır. İşlenmiş veri bilgiye dönüşmeye hazır iken işlenmemiş verinin bilgiye hazır hale gelmesi için bir takım veri madenciliği işlemlerine tabi tutulması gerekmektedir. Bilginin elde edilmesinde veri madenciliği büyük önem arz etmektedir.

Madencilik ile yer altındaki bir takım cevherler araştırılır, tespit edilir, cevher olmayan unsurlardan ayıklanır, çıkarılır ve nihayetinde kullanıma elverişli hale getirilerek kullanıcıların bilgisine sunulur. Veri madenciliğinde de klasik gerçek madencilikte olduğu gibi ham veri gereksiz ve değersiz verilerden ayrıştırılarak cevher niteliğindeki değerli bilgiye dönüştürülür. Günlük hayatta insan görsel, işitsel, sözel ve sayısal birçok bilgiye maruz kalmaktadır. İnsan zekâsının tüm bu bilgileri tanımlayıp

anlamlandırması, bilgiye dönüştürmesi ve hafızaya kaydetmesi fizyolojik bir veri madenciliği süreci olup bu süreç insan beyninde yüksek bir hızla gerçekleşmektedir. Ancak verinin nicelik olarak artması, nitelikli hale gelmesi ve uğraşılan problemlerin daha karmaşık olması insan beyninde gerçekleşen verinin bilgiye dönüşüm sürecini büyük ölçekte kısıtlamakta ve çoğunlukla da imkânsız hale getirmektedir.

Günümüz dünyası büyük verinin hâkim olduğu bir yapıya bürünmüş olup bu yapı gün geçtikçe daha da büyümekte ve devası bir hal almaktadır. Bu oluşum büyük verinin bilgiye dönüştürülmesinde ve depolanmasında insan beyni ve hafızasını yetersiz kılmaktadır. Bilim ve teknolojinin ilerlemesi büyük verinin oluşumuna kaynaklık yaptığı gibi büyük verinin bilgiye dönüştürülmesinde gerekli bilimsel ve teknolojik yeni yöntem ve tekniklerin geliştirilmesine büyük katkı sağlamıştır. Bu tekniklerden en önemlisi yapay zekâdır. Yapay zekâ ile yetersiz kalan insan beyni taklit edilmeye çalışılmış olup suni ve sanal bir mekanizma geliştirilmiştir. Geliştirilen bu yapay zekâ düzeneği ile ön işlemeye tabi tutulan veriler anlamlı bilgiye dönüştürülür. Anlamlı bilgiye dönüşüm süreci yapay zekânın doğal dil işleme, görüntü işlem, makine öğrenmesi ve uzman sistemler olmak üzere bir takım uygulama çeşitliliğine imkân tanımıştır.

Makine öğrenmesi, farklı tür ve teknikle elde edilmiş algoritmalar ve öğrenme tekinleri marifetiyle makinenin öğrenerek, veriyi modellemesi ve bilgiye dönüştürmesi sürecini ifade etmektedir. Makine öğrenmesi, veri madenciliği ve yapay zekâ ile doğrudan ilişki halinde olup sistemsel bir bütünlük sağlamaktadır. Başka bir ifadeyle makine öğrenmesi bilgi üretim sisteminin bir parçasını oluşturmaktadır. Makine öğrenmesi, üretilecek bilginin yapısına göre denetimli, denetimsiz, yarı denetimli ve takviyeli öğrenme şeklinde türlere ayrılmaktadır. Makine öğrenmesi ile elde edilen bilginin doğru, güvenilir ve geçerli olmasında verinin yapısı, işlenişi, model tekniği, model seçimi, model eğitimi ve model algoritmaları belirleyici konumdadır. Bu bağlamda performansı yüksek modellerle bilgiye ulaşma süreci ancak iyi yönetilmiş ve yürütülmüş veri madenciliği ile ancak mümkündür.

2.1.1 Zekâ ve Yapay Zekâ(YZ)

TDK ’ya göre zekâ, “İnsanın düşünme, akıl yürütme, öğrenme, kavramları ve nesneleri zihinde canlandırabilme, objektif gerçekleri algılama, yargılama, sonuç çıkarma, bedeni kontrol edebilme, duyguları doğru algılayabilme, değerlendirebilme, icat

edebilme vb. yeteneklerinin ve becerilerinin tamamı [8]" olarak tanımlanmaktadır. Bu tanıma göre zekâ kavramsal olarak insan ve yetenek/beceri olmak üzere iki ana temele oturtulmuştur. Yine TDK 'ya göre yapay kelimesi "Doğadaki örneklerine benzetilerek insan eliyle yapılmış veya üretilmiş; yapma, bileşimli, suni, tasni, sentetik, doğal karşıtı [8]"olarak ifade edilmiştir. Bu iki tanım bir araya getirildiğinde yapay zekâ, insan zekâsına benzetilerek, insan eliyle yapılmış bir düzenek olarak tanımlanabilir. TDK ise yapay zekâyı, "Bir bilgisayarın, bilgisayar kontrolündeki bir robotun veya programlanabilir bir aygıtın insana benzer biçimde algılama, öğrenme, fikir yürütme, karar verme, sorun çözme, iletişim kurma vb. işlevleri sergileyebilme yeteneği [8]" şeklinde sözlük anlamından daha fazla terimsel anlam olarak tanımlamıştır. Terimsel anlamdan da anlaşılacağı üzere yapay zekâ insan dışı bir bilgisayarın veya robotun insan zekâsı gibi hareket etmesidir.

Yapay zekâ ile ilgi birçok tanım bulunmaktadır. Bu tanımlardan bazıları şöyledir. Yapay zekâ, karmaşık problemlerin çözümü için makinelerin, insanın düşünme yapısını temel alarak çözüm üretmesini sağlayan, uygulamalı bilgisayar biliminin bir alt dalıdır. Bir başka tanıma göre yapay zekâ, bir problemin çözümü için düşünme, anlama, kavrama, yorumlama, öğrenme yapılarının programlama ile taklit edilmesidir. Kısacası, zekâ ve düşünme gerektiren işlemlerin, bilgisayarlar tarafından yapılmasını sağlayarak araştırmaların ve yeni yöntemlerin geliştirilmesi konusuyla ilgili bir bilim dalıdır [9].

2.1.2 Yapay Zekânın Amacı ve Tarihçesi

İnsan sürekli gelişime açık olarak kendi ve çevresi ile iletişim ve etkileşim halinde var edilmiştir. İnsandaki bu sürekli gelişim ihtiyacı bilim ve teknolojide yeni tekniklerin elde edilmesini gerekli kılmıştır. Bilim ve teknikte ilerleme sanayi alanında da ilerlemeyi ve üretimi beraberinde getirerek sanayi devrimlerine yol açmıştır.

İlk sanayi devrimi (1.0) su ve buhar gücünü kullanarak mekanik üretim sistemleri ile ortaya çıktı. İkinci sanayi devrimi (2.0) ile elektrik gücünün yardımıyla seri üretim tanıtılmıştı. Üçüncü sanayi devriminde (3.0) ise dijital devrim, elektroniklerin kullanımı ve BT (Bilgi Teknolojileri)'nin gelişmesiyle üretim daha da otomatikleştirildi. Dördüncü sanayi devrimi (4.0) birçok çağdaş otomasyon sistemini, veri alışverişlerini ve üretim teknolojilerini içeren kollektif bir sistemdir. Endüstri 4.0, gömülü sistem teknolojisiyle akıllı ürün üretim süreçlerini birleştirecek, yeni bir

teknolojiyi ortaya çıkaracak, bu teknolojiyi iş modellerine, üretim zincirlerine ve sanayiye aktaracaktır. Endüstri 4.0 nesnelerin internetinin üretim içerisinde hayat bulması anlamına gelmektedir. Endüstri 4.0 ile nesnelerin interneti, internetin hizmetleri ve siber-fiziksel sistemler oluşmuştur, aynı zamanda bu yapı akıllı fabrika sisteminin oluşmasında büyük rol oynamaktadır. Bu durum, üretim ortamında her bir verinin toplanmasına ve iyi bir şekilde izlenip analiz edilmesine olanak sağlamaktadır [10].

Bilim ve sanayi alanındaki bu devrimler üretim-tüketim döngüsü daha ileri tekniklerde makineleşmeyi ve üretimde insan emeğini azaltan otomasyon sistemlerine geçişe olanak sağlamıştır. Tüm bu gereksinimler daha az maliyetlerle daha kısa sürede bir üretim amacıyla ortaya çıkmıştır. Makineleşme süreci, elektrik ve seri üretim kapasitelerinin icat edilmesinden sonra toplumların gelişme göstergesi haline gelen sanayi toplumunu tetiklemeye devam etmiştir. Arayışlar devam etmiş ve otomatik çalışan makine hayalleri bilgisayarların icat edilmesine ve bilgi teknolojisindeki ilerlemelere yol açmıştır. Bu sayede işletmelerin otomatik makineler ve yazılım operasyonları yaygınlaşmaya başlamıştır. Özellikle 1950'li yıllarda bilim insanlarından bazıları artık otomatik makineler yapabiliyorsa yapay beyinde yapabileceğini düşünerek çalışmaları o yöne doğru kaydırmıştır [11]. Böylece makineleşme süreci bilim dünyasın da Turing tarafından "Makineler düşünebilir mi? sorusunu da beraberinde getirmiştir. Aynı yıllarda, Ordinaryüs Profesör Cahit Arf, Erzurum Atatürk Üniversitesinde "Makineler Düşünebilir mi ve Nasıl Düşünebilir?" konulu çalışmasını sunmuştur [2]. Sonrasında başta Allen Newell ve John McCarthy olmak üzere olmak üzere bazı araştırmacılar bugünkü anlayışa göre çok sınırlı olsa da bazı programlar yazmayı başararak robotlara zekâ kazandırmanın yolunu açmayı başarmışlardır. 1956 yılında Dartmouth'ta yapılan bir konferans ile bu gelişmelere "yapay zekâ" adını vererek yeni bir bilimin doğmasına yol açmışlardır [11].

Yapay zekânın kronolojik değişim ve gelişim tarihi incelendiğinde 1950 yılının milat olarak kabul edileceği söylenebilir. Günümüz yapay zekâ teknolojileri halen Turing'in "Makineler düşünebilir mi?" sorusunun üzerine konumlandırıldığı dikkat çekici bir

gerçeklik olarak durmaktadır. 1950-2024 döneminde yapay zekâ kimi zaman bilim[1] kimi zaman teknoloji[2], kimi zaman da endüstri olarak gelişimine devam etmiştir.

Yapay zekânın tarihi perspektifi esasında yapay zekânın amacını da ortaya koymaktadır. Bu bağlamda yapay zekânın amacı insan gibi davranış sergileyen ancak insandan çok daha fazla işlevselliğe ve hafızaya sahip mekanik ve elektronik bir düzeneğin geliştirilmesidir. Burada önemli olan husus yapay zekâlı sistemlerin insan gibi hareket etmesine karşın bir insan zekâsı ve hafızasının çok çok üstünde biri veri elde etme, depolama, işleme, bilgi üretme ve üretilen bilgi ile değer katma özelliğine sahip olmasıdır. Dolayısıyla yapay zekânın iki ana amaca hizmet ettiği söylenebilir. Bunlardan birincisi "insan gibi davranış sergileyen" ikincisi ise " yüksek işlevselliğe ve üretkenliğe" sahip bir yapının oluşturulmasıdır. Bu noktada klasik tanımlardaki "insan gibi düşünen, davranan, karar veren yazılım ve donanım sistemleri geliştirmek" şeklindeki tarif edilen yapay zekâ amacı yetersiz olup yapay zekânın amacının sadece bir bölümünü ifade etmektedir.

2.1.3 Yapay Zekâ Uygulamaları

Yapay zekâda kat edilen yüksek gelişimler ve ilerlemeler yapay zekâya olan ilgiyi ve yapay zekâ üzerine yapılan araştırma ve çalışmaları da arttırmıştır. Bu durum yeni yapay zekâ tekniklerinin geliştirilmesine imkân tanıdığı gibi klasik bazı tekniklerin de yapay zekâ uygulamalarına entegrasyonunu sağlamıştır. Genel olarak yapay zekâ uygulamaları doğal dil işleme, görüntü işleme, makine öğrenmesi ve uzman sistemlerdir. Bununla birlikte yapay sinir ağı yaklaşımı, optimizasyon teknikleri, bulanık mantık ve karar destek sistemlerinde yapay zeka uygulamaları arasında sayılmaktadır. Yapay zekâ uygulamalarına geniş bir perspektifte bakıldığında bir insanda olan konuşma, yazma, görme, analiz etme, anlamlandırma, tahmin etme ve çıkarımda bulunma özelliklerinin daha gelişmiş şekilleriyle farklı türdeki yapay zekâ uygulamaları ile makinelere ve robotlara kazandırılma çabaları görülmektedir. Dolayısıyla farklı türdeki her bir yapay zekâ uygulaması yapay zekânın geliştirilmesi aşamasına hem özel hem de genel bir katkı sağlamaktadır.

[1] Bilim; Evrenin veya olayların bir bölümünü konu olarak seçen, deneye dayanan yöntemler ve gerçeklikten yararlanarak sonuç çıkarmaya çalışan düzenli bilgi; ilim [8]
[2] Teknoloji; Bir sanayi dalı ile ilgili yapım yöntemlerini, kullanılan araç, gereç ve aletleri, bunların kullanım biçimlerini kapsayan uygulama bilgisi; uygulayım bilimi [8]

Doğal dil işleme(DDİ); İnsanın konuşma, anlama yeteneğini taklit eden yapay zekâ alt alanıdır. Doğal dil işleme çalışmaları kapsamında hem konuşma dili hem de yazılı dil üzerinde çalışmalar yapılmaktadır. Temel amacı dil içeriği üretimi ve bunların anlaşılmasıdır [1]. Bilişim firmalarını geliştirildiği çeviriler, yönetici asistanları doğal dil işleme kapsamında değerlendirilebilir. Doğal dil işleme ile insan gibi konuşan, yazan, tercümanlık ve asistanlık/sekreterlik yapan düzenekler hedeflenmiştir.

Görüntü İşleme(Gİ); İnsanın görme kabiliyetinin taklit eden yapay zekâ uygulaması görüntü işlemedir. Bilgisayar bilimlerindeki önemli çalışma konularından biri de görüntü işlemedir. Bir görüntüden faydalı bir bilgi çıkarılarak yorumlanması gerektiğinde görüntü işleme tekniklerinden faydalanılmaktadır. İşlenecek görüntü, kameralar, optik tarayıcılar ve fotoğraf makineleri yardımıyla elde edilebilir. Bu dijital görüntülerin sayısallaştırılmasıyla üzerinde farklı işlemler uygulanarak anlamlı yorumlanabilir sonuçlar elde edilebilir. Tıp, Askeri, Endüstriyel ve Coğrafi Sistemler gibi birçok alanda kullanılan görüntü işleme teknikleri, güvenlik sistemleri alanında da yaygın olarak kullanılmaktadır. Parmak izi, iris ve yüz tanıma gibi uygulamalar güvenlik alanında görüntü işleme teknikleri kullanılarak yapılabilmektedir [12].

Uzman Sistemler(US); İnsanın kendisi dışındaki diğer varlıklardan ayıran en büyük özelliği zekâsıdır. İnsan gibi davranış sergileyen bir mekanizma içinde olmazsa olmaz fonksiyon zekâ olması gerekmektedir. Bazı problemler vardır ki insan uzmanlığına olanak sağlayan zekâyı gerektirmektedir. Uzman sistemler işte tam da bu ihtiyaca hizmet etmektedir. Uzman sistemler, özel bir alanda ele alınan problemi konu ile ilgili uzmanların çözdüğü şekilde çözebilen bilgisayar programlarıdır. Uzman sistemlerin kökeni, insan zekâsının bilgiyi işleme sürecinin makine tarafından otomatik olarak gerçekleştirilebilmesi amacıyla sürdürülen çalışmalardır. Bunu yapabilmek için uzmanların sahip olduğu bilgi ve tecrübelerin bilgisayara aktarılabilmesi ve yine bilgisayar tarafından saklanması gerekmektedir. Böylelikle uzman sistemler, bilgi tabanında saklanan bilgileri kullanarak insan karar verme sürecine benzer bir yapıyla ele alınan probleme çözüm üretir [9]. Uzman sistemler endüstri mühendisliği, iş süreçleri, arıza tespit sistemleri, tıpta teşhis ve karar verme, finans, sigortacılık, konfigürasyon hazırlama, kütüphanecilik ve sistem kontrolü gibi alanlarda uygulanmaktadır [13].

Makine Öğrenmesi; Geleneksel programlama yöntemleri kullanılmadan, çeşitli algoritmalar ile öğrenme yeteneği kazandırılmış sistemler oluşturan bir yapay zekâ alanıdır [14]. Makine öğrenmesi, öğrenme yeteneği nedeniyle yapay zekânın en çok gelişen koludur. Bugün yapay zekâ alanında ortaya çıkan çalışmaların birçoğunda makine öğrenmesi teknikleri kullanılmaktadır. Makine öğrenmesinin ana fonksiyonu geçmiş deneyimler ve örenme algoritmaları yardımıyla sistemlerin öğrenmesidir [1]. Makine öğrenmesinin temelini matematik, istatistik, bilgisayar ve yazılım gibi bilimler oluşturmaktadır. Makine öğrenmesi yöntemleri günümüzde ekonomi, sağlık, hukuk, bilgi iletişim, tarım olmak üzere tüm sektörlerde kullanılabilir durumdadır. Makine öğrenmesi denetimli, denetimsiz ve yarı denetimli ve takviyeli öğrenme olarak ayırmak mümkündür.

2.1.4 Yapay Zekâ Teknolojileri

Yapay zekâda bir bilim olarak ilerlemeye kaydedip önem arz ettiği gibi bu bilime dayanılarak elde edilen teknolojide büyük önem arz etmektedir. Teknoloji esasında teorik bilimin uygulamalı karşılığıdır. Bu bağlamda yapay zekâ teknolojileri denilince yapay zekâya dayalı günlük hayatın her bir alanında kullanılan teknolojik araç, gereç ve aletler kastedilmektedir. Genel olarak yapay zekâ teknolojileri eğitim, sağlık, mühendislik, kamu hizmeti, sosyal, hukuk, kültür, tarım, sanayi, savunma sanayi ve uzay olmak üzere tüm alanlarda artık kendine bir edinmiştir. Otonom şoförsüz taksiler, hastalık tespit sistemleri, karanlık fabrikalar, e-danışmanlık, yapay zekâ teknolojilerine örnek gösterilebilir. Günlük hayatın kolaylaştırılması ve maliyetlerin düşürülmesi yapay zekâ teknolojilerinin gelişmesine bağlıdır.

2.1.5 Veri ve Veri Madenciliği

TDK ‘ya göre veri, “Gözlem ve deneye dayalı araştırmanın sonuçları [8]” olarak tanımlanmaktadır. Yine TDK ‘ya göre madencilik ise “Yer altındaki bir takım cevherler araştırılır, tespit edilir, cevher olmayan unsurlardan ayıklanır, çıkarılır ve nihayetinde kullanıma elverişli hale getirilerek kullanıcıların bilgisine sunulur [8]” şeklinde tanımlanmaktadır. TDK tarafından henüz yapılmış bir veri madenciliği tanımı olmamakla beraber veri ve madencilik kelimeleri bir araya getirildiğinde veri madenciliği kavramsal olarak “veri tabanları gibi bir takım veri kaynaklarında yer alan ham veri içinde gizlenmiş ve cevher niteliğinde değer taşıyan anlamlı bilginin

araştırılma, tespit etme ve işleme gibi bir dizi işlemlere tabi tutularak açığa çıkartılması işlemidir.” olarak tanımlanabilir.

Veri madenciliğinin terimsel olarak birçok tanımı bulunmaktadır. Genel olarak, veri madenciliği, verilerin farklı bir bakış açısından analiz edilmesi ve kullanışlı bilgi halinde özetlenme sürecidir. Teknik olarak veri madenciliği, büyük ve birbiriyle ilişkili veri tabanları içinde düzinelerce alan arasında korelasyonlar ve düzenler bulma sürecidir. Veri madenciliği üzerinde fikir birliğine varılmış ortak bir tanım yoktur [15]. Ancak veri madenciliği tanımlanır incelendiğinde veri madenciliğinin bazı belirleyici unsurlar üzerine bina edildiği gözlemlenmektedir.

Veri madenciliğinde belirleyici unsurlar;

1. Veri tabanları ve diğer veri kaynakları
2. Büyük ve karmaşık veri,
3. Bilimsel Teknikler/Farklı Disiplinler(Matematik, istatistik, yapay zekâ, vb.)
4. Veri Analizi/Modelleme
5. Anlamlı/Gizli/Değerli Bilgi olmak üzere beş unsurdan oluşmaktadır.

Günlük hayatta insan görsel, işitsel, sözel ve sayısal birçok elde eder ve zekâsıyla tüm bu bilgileri tanımlayıp anlamlandırması, bilgiye dönüştürmesi ve hafızaya kaydetmesi fizyolojik bir veri madenciliği süreci olup bu süreç insan beyninde yüksek bir hızla gerçekleşmektedir. Ancak verinin nicelik olarak artması, nitelikli hale gelmesi ve uğraşılan problemlerin daha karmaşık olması insan beyninde gerçekleşen verinin bilgiye dönüşüm sürecini büyük ölçekte kısıtlamakta ve çoğunlukla da imkânsız hale getirmektedir. Dolayısıyla büyük ve karmaşık verilerin madenciliğinde bilgisayar teknolojisi gerekmektedir. Bu bağlamda literatürde yapılan veri madenciliği tanımlamalarındaki belirleyici unsurlara veri madenciliği işleminin gerçekleştirileceği alanlara bilgisayar teknolojileri bir belirleyici unsur olarak eklenmesi gerekmektedir. Tüm bu belirleyici unsurular üzerinden bir tanım yapıldığında veri madenciliği; Veri tabanları ve diğer veri kaynaklarında yer alan karmaşık ve büyük veride gizli kalmış anlamlı ve değerli bilginin bir takım bilimsel teknikler uygulanarak bilgisayar teknolojileri ile elde edilmesi sürecidir. Bu bağlamda veri madenciliği bir süreç olarak da değerlendirilmektedir.

Veri madenciliği sürecinde ham veri giriş(input), değerli bilgi ise çıkış(output) olarak yer almaktadır. Ham veriden değerli bilgi elde edilmesinde gerçekleştirilen veri keşfi,

veri ön işlemesi, istatistiksel analizler ve modelle teknikleri ise veri madenciliği sürecine ait faaliyetlerdir. Dolayısıyla veri madenciliği kendini oluşturan alt alanların hepsinden daha fazlasını ifade eder [1].

Bilginin parasal olmayan güç olarak konumlandırılması günümüz dünyasında bilgiyi ve bilgiye kaynaklı eden veriyi daha da değerli kılmıştır. Bilim ve teknolojide alanında yaşanan gelişimler bilgiye ulaşımı ve verinin depolanmasını daha da kolay kılmıştır. Ancak bu kolaylık nicel anlamda büyük, nitel anlamda ise karmaşık bilgi/veri kitlelerine de olanak sağlamıştır. Veri tabanlarında depolanan ve büyük veri olarak ifade edilen veri kitlesinden değerli ve anlamlı bilginin üretilmesi hem devlet hem de özel işletmeler için finansal ve finansal olmayan birçok faydalar sağlamaktadır. Dolayısıyla veri madenciliği hem kamu hem de özel kurum ve kuruluşların vazgeçilmez bir aracı haline dönüşmüştür. Genel olarak güvenlik, sağlık, eğitim, kültür, sosyal ve ekonomi alanları başta olmak üzere birçok alanda veri madenciliği işlemleri gerçekleştirilmektedir. Özel olarak ise iş zekâsı, iş profilleme, hastalık tespiti, insan kaynakları yönetimi, ürün önerimi gibi birçok farklı sektör ve alanda veri madenciliği işlemleri uygulanmaktadır.

Veri madenciliğinin geniş bir yelpazede kullanımı veri madenciliğinin önemini ortaya koymaktadır. Yapay zekâ çalışmalarındaki yüksek ivme veri madenciliği sürecinin daha gelişmesine imkân tanıdığı gibi veri madenciliğinin daha da önemli hale geleceğini göstermektedir. Bu noktada veri madenciliği, yapay zekâ teknolojilerinin geliştirilmesinde büyük önem arz etmektedir.

2.1.6 Veri Madenciliği İşlem Süreci

Süreç; olguların ya da olayların belli bir taslağa uygun ve belli bir sonuca varacak biçimde düzenlenmesi ve art arda sıralanması. Bir şeyin yapılış, üretiliş biçimini oluşturan sürekli işlemler, eylemler dizisi olarak da tanımlanır [16]. Veri madenciliği veri girişi ile başlayıp bir takım sıralı faaliyetlerin geçekleşmesi sonrası bilgi çıkışı ile son bulan bir süreçtir. Literatür incelendiğinde veri madenciliği sürecini açıklamak ve bir standarta uygun hale getirmek üzere farklı şematik gösterimler bulunmaktadır. Bu gösterimler CRISP-DM (*Cross Industry Standard Process for Data Mining*), SEMMA

(*Sample, Explore, Modify, Model, Assess*)[3] ve KDD (*Knowledge Discovery in Databases*) olmak üzere üç farklı süreç modelinden oluşmaktadır.

CRISPM-DM veri madenciliği süreci;

1. Problemin tanımlanması
2. Verilerin anlaşılması
3. Verilerin hazırlanması
4. Modelleme
5. Değerlendirme
6. Modellerin kullanımı olmak üzere altı aşamadan oluşmaktadır.

KDD bilgi keşfi süreci ise;

1. Veri seçimi
2. Veri ön işleme
3. Veri dönüştürme
4. Veri madenciliği
5. Değerlendirme/Yorumlama olmak üzere beş aşamadan oluşmaktadır.

SEMMA süreci ise;

1. Veri toplama
2. Veri keşfi
3. Veri dönüşümü
4. Modelleme
5. Değerlendirme olmak üzere beş aşamadan oluşmaktadır.

Her üç veri madenciliği süreci birlikte değerlendirildiğinde esasında aynı amaca hizmet ettiği net olarak meydana çıkmaktadır. Ancak her üç süreç yine de standart bir veri madenciliği süreci oluşturamamıştır. Ayrıca veri madenciliği sürecinde olması zorunlu olmayan bazı aşamalar kesin olması gerekli gibi gösterilmiştir. Örneğin CRISPM-DM veri madenciliği süreci problemin tanımlanması ile başlamıştır. Hâlbuki bazen veri madenciliği bir problem olmaksızın da başlatılır. Bununla birlikte KDD bilgi keşfinde veri madenciliği süreci keşfin bir parçası şeklinde gösterilerek veri

[3] Sample, Explore, Modify, Model, ve Assess kelimelerinin baş harflerinden oluşan bir metodolojidir. İstatistik ve İş Zekâsı yazılımı geliştiren SAS Enstitüsü tarafından geliştirilen ardışık adımlar listesidir. [17]

madenciliği daraltılmıştır. SEMMA süreci veri madenciliği sürecini daha özet ve net ortaya koyarken anlamlı bilgiye değinilmemiştir. Ayrıca tüm süreçler verinin kaynağı ve madenciliğin bizzat yapıldığı veri tabanlarına ve geri bildirime hiç değinmemiştir. Dolayısıyla veri madenciliği sürecinin tanımına ve yapısına uygun daha iyi veri madenciliği sürecinin gösterilmesi gerekliliği ortaya çıkmıştır.

Tablo 2.1 Veri Madenciliği Sürecinin Ana ve Alt Aşamaları

Ana Aşamalar	Alt Aşamalar
Amacın Belirlenmesi (İşin Tanımı)	Problemin/İşin Tanımlanması
	Amacın Tanımlanması
	Planlamanın Yapılması
Veri Kaynağının Belirlenmesi	Veri Tabanları, İdari Kayıtlar, Deneyler, Veri Ambarı Anketler vb. Belirlenmesi
Verinin Seçilmesi	Hedef Verinin Seçilmesi(Yığın, Örneklem vb.)
	Hedef Verinin İncelenmesi
	Hedef Verinin Düzenlenmesi
Veri Keşfi	Verinin Betimlenmesi (Eğilim ve Dağılım Ölçüleri)
	Verinin Görselleştirilmesi(Histogram,Diyagram, Dağılım G.)
	Verinin Anlamlandırılması(İstatistiksel Testler)
Verinin Hazırlanması	Veri Ön İşleme(Veri Temizleme, Ayıklama, Doldurma vb.)
	Veri Dönüştürme(Normalizasyon, Standartlaştırma vb.)
	Veri Boyutlandırma(Feature Selection, Feature Extraction)
	Veri Dengeleme(Oversampling, Undursampling, SMOTE)
Modelleme	Model Seçimi(Tahmin Edici, Tanımlayıcı Modeller)
	Model Tekniğinin Belirlenmesi (İstatistiksel Yöntemler, Yapay Zekâ Yöntemleri vb.)
	Model Yönteminin Seçimi (Sınıflandırma, Regresyon, Kümeleme, Diskriminant, Birliktelik, Korelasyon vb.)
	Model Algoritmasının Seçimi (Lojistik Regresyon, Rasgele Orman vb.)
	Model Geçerleme/Doğrulama Yönteminin Seçimi (Hold Out, K-Çapraz Doğrulama vb.)
	Özellik/Öznitelik Seçimi
	Model Eğitimi ve İnşası
	Model Performans Değerlendirme
	Model Performans İyileştirme
	Amaçlanan Modelin Seçimi
	Yeni Veri İle Modelin Test Edilmesi
	Modelin Sürekli İzlenmesi
	Model Geliştirme
Anlamlı Bilgiye Ulaşma	Bilginin Elde Edilmesi
	Bilginin Değerlendirilmesi ve Yorumlanması
	Bilginin Eyleme Dönüştürülmesi(Karar Destek,İş Profilleme)
Geri bildirim/ Yenileme/ Güncelleme	Geri Bildirim(Bilginin Yenilenmesi ve Güncellenmesi)
	Bilginin Veri Madenciliği Sürecine Dâhil Edilmesi
	Bilginin Depolanması

Tarafımızca hazırlanan veri madenciliği süreci diğer süreçlerden biraz farklı olarak sekiz ana aşamadan oluşmaktadır. Bu ana aşamalarda kendi içerisinde alt aşamalardan oluşmaktadır.

1. Amacın belirlenmesi, Problemin/İşin tanımı
2. Veri kaynağının belirlenmesi
3. Verinin seçilmesi
4. Veri Keşfi
5. Verinin hazırlanması
6. Modelleme
7. Anlamlı bilgiye ulaşma
8. Geri bildirim/Yenileme/Güncelleme

Veri madenciliği ana ve alt aşamaları incelendiğinde amaçtan bilgiye ulaşmanın çokta kolay olmadığı ortaya çıkmaktadır. Veri madenciliği süreci öncelikle veri madenciliğine gereklilik hissedilen amacın belirlenmesi ile başlar. Bu amaç kimi zaman bir araştırma sorusudur, kimi zaman bir meraktır, kimi zaman ise işletmesel bir problemden kaynaklanabilir. Amacın belirlenmesine müteakip gerçekleştirilecek veri madenciliği sürecinin planlaması yapılarak en önemli veri madenciliği aşaması tamamlanır.

Amaca uygun bilginin elde edilmesinde doğru veri setinin temin edilmesi gerekmektedir. Doğru veri setinin temini ise veri kaynağının doğru belirlenmesinden geçmektedir. Veriler deneysel, gözlemsel yollarla toplanır ve kimi zaman idari kayıt şekliden kimi zamanda başka şekillerde veri tabanlarında depolanır. Bazen ise anket tekniği ile elde edilen verile ise depolanmadan analize dâhil edilebilir. Veri madenciliğinde talep edilen veri için doğru veri kaynağına gidilerek ikinci aşama geçilir.

Üçüncü aşamada veri setinin türü belirlenerek veri seti veri keşfi için incelenir ve düzenlenir. Veri seti incelenmesinde verini boyutu, öznitelik ve gözlem sayısı, veri metinin tutarlılığı, verinin kendi içerisinde tutarlılığı ve uyumu gibi özellikleri ile veri seti incelenir ve veri keşfine uygun olması için gerekli düzenlemeler yapılır.

Veri keşfinde veriler mod, medyan, aritmetik ortalama gibi merkezi eğilim ölçüleri ve varyans, standart sapma ve değişim katsayısı gibi dağılım ölçülerine göre betimlenir, histogram, diyagram, kutu grafiği, daire ve çubuk grafiği gibi araçlarla

görselleştirilerek ve istatistiksel testlerle anlamlı hale getirilerek veri madenciliğinin dördüncü aşaması geçilir. Veri keşfi ile veri seti hakkında detaylı bilgi sahibi olunur ve verinin bilgiye dönüştürülme sürecinde önemli bir yol alınmış olur.

Veri madenciliğinin beşinci aşamasında veri seti veri ön işleme işlemleri ile ayrık, uç, hatalı ve gereksiz verilerden temizlenir, eksik hücreler doldurulur ve gerektiğinde gözlem ve öznitelikler arasında satır ve sütun birleştirme işlemler yapılır, veri ön işleme sonrasında veri numerik ve kategorik olmasına göre, birim farklılıklarına göre gerekli normalizasyon ve standardizasyon işlemlerinden geçirilir, gerektiğinde veri seti öznitelik çıkarma ve seçme işlemleri ile boyutlandırılır ve yine gerektiğinde veride dengeleme işlemi gerçekleştirilerek veri modellenmeye hazır hale getirilir.

Veri madenciliğinin altıncı aşaması ile modelleme işlemine başlanır. Öncelikle modellemenin tekniği belirlenir. Modelleme tekniğinde istatistiksel yöntemler kullanılacağı gibi makine öğrenmesi, uzman sistemler gibi yapay zekâ teknikleri veya fuzzy[4], optimizasyon gibi başkaca modelleme tekniklerinde kullanılabilir. Modelleme tekniğinden sonra model seçimi yapılır. Model seçimi veri madenciliğinin amacına uygun olarak tahmin edici ve tanımlayıcı modeller şeklinde yapılır. Model seçiminin yapılmasından sonra model uygun yöntem ve algoritma belirlenir. Seçilen model tekniği bir yapay zekâ makine öğrenmesi tekniği ise model eğitimi için uygun model geçerleme/doğrulama yöntemi seçilir. Modele girecek özniteliklerin belirlenmesinden sonra model eğitimi ve inşasından sonra model performansları doğruluk, kesinlik gibi performans göstergeleri üzerinden hesaplanır ve model performansları değerlendirilir Gerektiğinde hesaplanan model performanslarının arttırılması için performans iyileştirme işlemi yapılarak amaçlanan en uygun modele ulaşılır. Amaçlanan model yeni veriler ile test edilir ve testi geçen model uygulamaya konulur ve sürekli izlenerek geliştirilmesi sağlanır. Artık model anlamlı ve değerli bilgi üretmeye hazır hale gelmiştir.

Veri madenciliğinin yedinci aşaması ile anlamlı bilgi üretimi başlamıştır. Üretilen bilginin değerlendirilmesi ve yorumlanması işlemi ile bilginin eyleme dönüşmesi

[4] Doğada nasıl ki karmaşık davranışlar, şekiller bulunuyor, her durumda bir kesinlik olmayabiliyor, kalıplar çok net ve doğrusal değilse bunların makinelere de aktarılışındaki bu bulanıklık ve belirsizlik "Bulanık Mantık" (Fuzzy Logic) kavramının ortaya çıkışında etkili olmuştur. Klasik mantıktaki "0" ve "1" gruplarına ayırma mantığının aksine iki değer arasındaki diğer ihtimallere de fırsat tanınır, bunlar mutlak değerlerden ziyade gri bölge için sınır değerleri olarak görülür. Bulanık Mantık; elektronik aletler, endüstri, robotik, otomasyon, görüntü işleme ve örüntü tanıma gibi birçok alanda kullanılmaktadır [18].

kararı verilir. Bu karar eyleme dönüşme şeklinde olduğunda veriden elde edilen bilgiye dayalı karar destek sistemleri, iş zekâları gibi teknolojik araçlar geliştirilebilir. Verinin dinamik yapısı veri güncelliğinin beraberinde getirmektedir. Güncel olmayan verilerden elde edilen bilgiler ise gerçeklik ve tutarlılıktan uzaktır.

Veri madenciliğinin bir süreç olması geri bildirimi de zorunlu kılmaktadır. Bu noktada veri madenciliği süreci sonunda elde edilen bilginin güncellenmesi ve yenilenmesi etkin bir geri dönüşüm ile mümkündür. Bilginin elde edilmesi kadar bir diğer önemli husus bilginin depolanması ve yeni bilgi üretimi için veri madenciliği sürecine dâhil edilmesidir. Elde edilen bilginin veri madenciliği sürecine dâhil edilmesi ile veri madenciliği sürecinin son aşaması biter ancak süreç yeni bilgi ile tekrar işletilir.

Veri madenciliği işlem süreçlerine genel olarak bakıldığında içeriğinde matematiksel ve istatistiksek teknikler, yapay zekâ teknikleri ve bilişim teknolojileri gibi birçok teknik ve yöntem birlikte kullanılmaktadır. Dolayısıyla veri madenciliği sürecinin söz konusu tekniklere göre daha kapsayıcı ve daha büyük bir yapı arz ettiği söylenebilir. Bu bağlamda bir yapay zekâ uygulaması olan makine öğrenmesi tekniği de veri madenciliği süreçlerinin bir parçası olarak değerlendirilmektedir. Belki de günümüzde veri madenciliğinde en fazla tercih edilen teknik olup büyük önem arz etmektedir.

Ver madenciliğe, özellikle yapısal olan ve olmayan büyük veri yığınlarından yararlanarak yeni durum ve yöntem bilgilerinin açığa çıkarılması amacıyla yürütülen çalışmaları içermektedir. Bu çalışmalarda yararlanılan yöntemler iki kategoride ele alınabilir: İstatistiksel yöntemler ve yapay zekâ yaklaşımları. Ver madenciliğinde istatistiksel yaklaşımların kullanımında klasik istatistiksel yaklaşımlara göre önemli bir fark bulunmaktadır. Klasik yaklaşımlarda, gerçekleşenle model arasındaki farkların yan hata karelerinin azaltılması esas iken büyük veride artan ver hacmi, hataların göz ardı edilmesine olanak vermektedir. Öte yandan kesinlik içermeyen yapay zekâ algoritmaları, verilerdeki örüntülerin yakalanmasında veri büyüklüğünden istatistiksel yöntemler kadar etkilenmemektedir [19]. Yapay zekâ algoritmalarının başta büyük veriye duyarlılığı olmak üzere istatistiksel yöntemlere göre başkaca üstünlükleri bu yöntemlerin uygulanma sıklığını arttırmıştır.

2.1.7 Makine Öğrenmesi

Makine öğrenmesi bir yapay zekâ ve veri madenciliği uygulamasıdır. Makine öğrenmesinin tarihsel gelişimi yapay zekâ ile paralellik göstermekte olup Turing'in 1952'de "makineler düşünebilir mi?" sorusuna dayanmaktadır. Veri madenciliğinde olduğu gibi makine öğrenmesine ilişkinde pek çok tanım bulunmaktadır. En basit tanımıyla makine öğrenmesi mevcut veriden makinelerin bir takım öğrenme teknikleri ve algoritmaları ile eğitilerek tahmin edici ve açıklayıcı bilgiler üretmesi işlemidir. Bir başka tanımda makine öğrenimi, makinelerin mevcut verilerden öğrendiği ve kendi kendine öğrenip geliştirdiği bir konsept üzerinde çalışır. Büyük verilerin yansıra geçmiş deneyimlerden gelen algoritmayı ifade eder [3]. Bununla birlikte makine öğrenimi programlanmamış sonuçları bile ortaya çıkarabilen bir tür yapay zekâ ve insan müdahalesi olmadan sonuçları tahmin etmede daha doğru olmasını sağlayan bir yapay zekâ türü olarak da tanımlanabilir [3].

Makine öğrenmesi de tıpkı veri madenciliğinde olduğu gibi bir süreç olup veri madenciliği süreci ile benzerlikler içermektedir. Esasında verinin bilgiye dönüştürülmesi sürecinde makine öğrenmesinin model tekniği olarak belirlenmesi durumunda veri madenciliği süreci doğrudan makine öğrenmesi süreci olmaktadır. Bu noktada veri madenciliği süreci(CRISP-DM) ve makine öğrenmesi süreci(CRISP-ML) olarak büyük benzerlikler içermektedir.

Makine öğrenmesinin ana fonksiyonu geçmiş deneyimler ve öğrenme algoritmaları yardımıyla sistemlerin öğrenmesidir. Deneyimlerde öğrenen makine öğrenme algoritmaları bugün ve gelecek için faydalı bilgiler üretmektedir [1]. Bu noktada üretilen bilginin türü ve üretim tekniklerini daha net ortaya koyan makine öğrenmesi sistem tasarımı makine öğrenmesini daha iyi anlamamıza ve tanımlamamıza yardımcı olmaktadır.

2.1.8 Makine Öğrenmesi Sistem Tasarımı ve Aşamaları

Makine öğrenmesi süreci verinin hazırlanması ve modellenerek bilgiye dönüştürülmesi olarak iki ana aşamadan oluşmaktadır. Bu iki ana aşama makine öğrenmesi süreci "veri tasarımı" ve "makine öğrenmesi sistem" tasarımı olarak da adlandırılabilir.

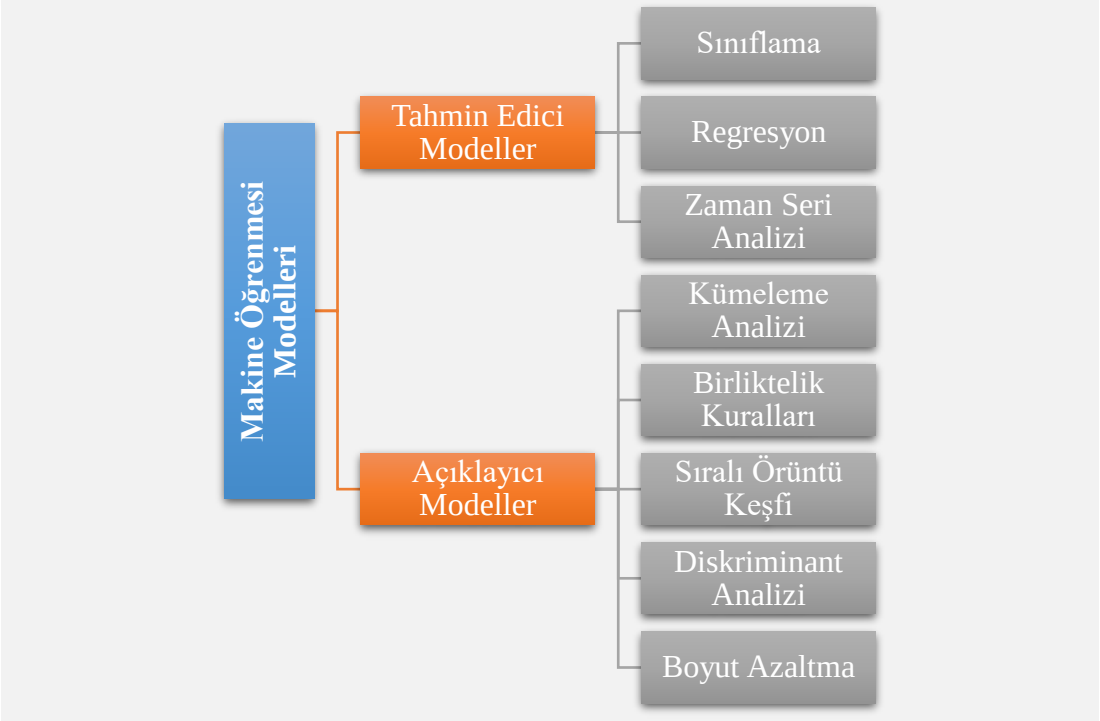
Tablo 2.2 Makine Öğrenmesi Sistem Tasarımı ve Aşamaları

	Aşamalar	**Açıklama**
1	**Problem/Amaç**	Anlamlı bilgiye ulaşma amacı
2	**Model Tekniği**	Makine Öğrenmesi Teknikleri
3	**Model Seçimi**	Tahmin Edici, Tanımlayıcı Modeller
4	**Model Öğrenme Türü**	Denetimli Öğrenme, Denetimsiz Öğrenme, Yarı Denetimli Öğrenme, Takviyeli Öğrenme
5	**Model Algoritmaları**	Rasgele Orman, K-En Yakın Komşu, Naive Bayes, Karar Ağaçları ve Destek Vektör Makineleri, Yapay Sinir Ağları, Lojistik Regresyon
6	**Model Geçerleme/Doğrulama Yöntemi**	Hold Out, K-Çapraz Doğrulama
7	**Veri Seti Seçimi**	Yığın ve Örneklem
8	**Özellik Seçimi**	Veri setindeki özniteliklerin seçilmesi işlemi
9	**Model Eğitimi ve İnşası**	(Eğitim=%70, Test=%30), (Eğitim=%80, Test=%20) K-5 Ve K-10 Çapraz Doğrulama
10	**Model Performans Değerlendirme**	Doğruluk(Accuracy),Kesinlik(Precision), Duyarlılık(Recall), F-Ölçütü, ROC Eğrisi
11	**Model Performans İyileştirme**	Veri Seti Değişimi, Algoritma Değişimi, Geçerleme Değişimi, Özellik Değişimi, Eğitim Değişimi, Algoritma Parametre Değişimi
12	**Nihai Modelin Seçimi**	Model Performans Değerleri En İyi Model
13	**Model Veri Tahmini**	Yeni Veri İle Nihai Model Üzerinden Tahmin /Tanımın yapılması

Makine öğrenmesi sistem tasarımı anlamlı bilgiye ulaşmak için gerekli olan problemin tanımlanması ile başlayıp ve model tekniği olarak makine öğrenmesi tekniklerinin uygulanmasına karar verdikten sonra model seçimi, model geçerleme yöntemlerinin belirlenmesi, özellik seçimi, model eğitimi ve inşası, model performans değerlendirme ve yorumlama, model performans iyileştirme, nihai modelin seçimi ile devam edip modelin yeni veri ile test edilmesi ve yeni veriye ilişkin tahmin yapılması ile son bulmaktadır. Esasında yeni veri ile nihai model üzerinden gerçekleştirilen tanımlama veya tahmin ile anlamlı bilgiye ulaşılmış olmaktadır.

2.1.9 Makine Öğrenmesi Modelleri ve Yöntemleri

Hem veri madenciliği sürecinde hem de makine öğrenmesi sürecinde verinin bilgiye dönüştürülmesi bir amaca istinaden gerçekleştirilir. Bu amaç aynı zamanda model seçimini ve tekniğini de belirleyici konumdadır. Eğer makine öğrenmesi sürecinin amacı mevcut veri setindeki de deneyimlerden geleceğe yönelik tahmin edici bilgiler üretmek olduğu gibi veri setindeki öznitelikler ve gözlemler arasındaki gizli ancak değerli ilişkinin açıklanması şeklide de olabilir. Dolayısıyla hem veri madenciliği hem de makine öğrenmesi süreçleri iki ana bilgi üretmek üzere kurgulanmıştır. Bunlar veriyi açıklamak/tanımlamak ve veriden tahminlerde bulunmaktır. Bu noktada makine öğrenmesi modelleri tahmin edici ve açıklayıcı modeller olmak üzere iki ana kısma ayrılır.



Şekil 2.1 Makine Öğrenmesi Modelleri ve Yöntemleri

Tahmin Edici Modeller (Predictive)

Sonuçları bilinen verilerden hareket ederek bir model oluşturup, sonuçları bilinmeyen veri kümeleri için sonuç değerlerinin tahmin edilmesidir [17]. Gerçekleştirilen tahmin türü veri setinin yapısı ile doğrudan alakalıdır. Eğer veri seti kategorik yapıda ise tahmin sınıflandırma şeklinde, veri seti sürekli yapıda ise tahmin regresyon şeklinde ve veri seti bir zamana bağlı olarak yapılandırılmış ise tahmin zaman serisi şeklinde

gerçekleştirilmektedir. Dolayısıyla tahmin edici modeller sınıflandırma, regresyon ve zaman serisi analizi olmak üzere üç kısıma ayrılmaktadır. Örneğin işsizlerin iş aramaya başladıktan sonra bir yıl içerisinde işe yerleşip yerleşmeme durumlarına ilişkin tahmin edici model bir sınıflandırma modeli iken, aynı işsizlerin işe yerleşme sürelerinin tahmin edici model regresyon modeli ve işsizlerin yıl içerisinde iş aramaya başlama dönemine göre işe yerleşme dönemini tahmin edici model ise zaman serisi modelidir.

Tanımlayıcı Modeller (Descriptive)

Karar vermeye rehberlik etmede kullanılabilecek verilerdeki örüntülerin tanımlanmasını sağlamaktadır [17]. Veriler arasındaki gizli ilişkilerin açığa çıkarılması ile veriler anlamlı bir şekilde açıklanır ve karar vericiye önemli bir bilgi sunar. Veri seti, özniteliklere kümeleme ve ayırma(Diskriminant) analizi ile gözlemlere göre birliktelik analizi ve sıralı örüntü keşfi yöntemleri ile modellenir ve açıklanır. Makine öğrenmesi işlem sürecinde elde edilecek bilginin türü, model seçimini, model yöntemini, model öğrenme türünü ve model algoritmalarını belirleyici konumdadır.

2.1.10 Makine Öğrenmesi Öğrenme Türleri

Öğrenme, bireyin yaşantılar sonucu davranışlarda meydana gelen oldukça uzun süreli değişmelerdir. Bir bilgi ve becerinin, öğrenme sayılması için davranışta değişiklik yapması ve davranıştaki değişikliğin uzun süreli olması gerekmektedir. Yeni öğrenmeler ile kişinin kapasitesi gelişir, önceden yapamadığı bir şeyi yapabilir hale gelir. Daha geniş anlamda, öğrenme sonucu, birey içinde bulunduğu evrene yeni bir anlam yükler ve evrendeki konumunu yeniden tanımlar [20].

Öğrenme kavramının tanımına bakıldığında bireylerden bahsedilmiştir asıl sorulması gereken soru ise makinelerinin öğrenip öğrenmeyeceğidir. Bu soru 1940'lı yıllardan itibaren sorulmuş ve makinelerin öğrenip öğrenmeyeceği konusunda birtakım deneyler gerçekleştirilmiştir. Günümüz teknolojisinde artık makinelerin öğrenmesi, yapay bir zekâya sahip olması işlemi gerçekleşmiş olup otonom makineler git gide yaygınlaşmış ve makine öğrenmesi yöntemleri ve birçok model geliştirilmiştir.

Makine öğrenmesi öğrenme türleri oluşturulması planlanan modele göre şekillenmekte olup genel olarak dört kısıma ayrılmaktadır. Bunlar denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve takviyeli öğrenmedir.

Denetimli öğrenme

Bir öğretmen eşliğinde hangi girdilerin hangi çıktılarla eşleşeceğini ifade eden bir makine öğrenmesi görevidir [1]. Denetimli öğrenme, gözetimli, danışmanlı ve güdümlü öğrenme olarak da isimlendirilmekte olup günümüzde sıklıkla kullanılmakta olan bir yöntemdir. Denetimli öğrenmenin temel çalışma prensibi bilinenden bilinmeyeni tahmine çalışmaktadır. Bu noktada bir bağımlı hedef/karar/etiket değişkene karşılık birden fazla bağımsız değişken/öznitelik bulunması gerekmektedir. Öz niteliklerin hedef değişken üzerindeki mevcut etkileri hesaplanarak farklı öz nitelik değerlerine karşılık yeni hedef değişkenin hesaplanması veya sınıflandırılması hedeflenmektedir. Bu bağlamda girdiler ve çıktılar arasında bir modelin oluşturulması gerekmektedir. Denetimli öğrenme regresyon ve sınıflandırma işlemi olarak ikiye ayrılmaktadır. Regresyon işleminde bağımsız değişkenlere karşılık sürekli, nicel bağımlı değişkene ilişkin tahmin yapılmakta olup bu durumun istisnası ise lojistik regresyon modellemesidir. Sınıflandırma işlemi ise nitel kategorik yapıda olan hedef değişkeninin tahmin edilmesidir. Denetimli öğrenme sürecinde veri eğitim ve test verisi olarak ikiye ayrılır, eğitim verisi ile modelin öğrenmesi ve akabinde test verisi ile değerlendirmesi işlemi gerçekleştirilir. Modelin başarı oranı test verisi ile ölçülür. Sınıflandırma işlemi ikili, üçlü veya çoklu şekilde gerçekleştirilebilir. Denetimli makine öğrenmesi sınıflandırma işlemi lojistik regresyon, rasgele orman, destek vektör makinaları, k en yakın komşu, naive bayes ve karar ağaçları gibi makine öğrenmesi algoritmaları ile regresyon işlemi ise basit doğrusal regresyon, çok değişkenli regresyon, rasgele orman regresyon, destek vektör regresyon, polinom regresyon ve karar ağaçları regresyon gibi çeşitli regresyon türleri ile gerçekleştirilir.

Denetimsiz öğrenme

Etiketsiz veri üzerinde önceden bilinmeyen kalıpları bulmaya yardımcı olan ve kendi kendine organize olan bir öğrenme türüdür [1]. Denetimsiz öğrenmede danışmansız ve gözetimsiz öğrenme olarak adlandırılmakta olup özellikle verinin anlamlandırılmasında sıklıkla kullanılmaktadır. Denetimsiz öğrenme problemlerinde, veri setinde sınıflara ayrılamamış veriler ile çalışılır. Denetimsiz öğrenme ile veri setindeki etiketlenmemiş veriler arasındaki gizli ilişkileri ortaya çıkarmak amaçlanır [14]. Denetimsiz öğrenmede özniteliklere karşı bir hedef bağımlı değişken bulunmamakta olup makinenin kendi kendine öğrenme işlemi sağlanmaktadır.

Denetimli öğrenmede bilinen x ve y değişkenleri ile bir bağ aranırken denetimsiz öğrenmede sadece bilinen x değerleri ile aynı özniteliklere sahip olanların gruplanması veya kümelenmesi hedeflenmektedir. Bu kapsamda sıklıkla kullanılan yöntem kümeleme yöntemidir. Kümeleme yöntemi dışında anormallik tespiti, birliktelik analizi, ayırma analizi, sıralı örüntü keşfi de denetimsiz öğrenme işlemleri arasında sayılmaktadır. Bu analizlerin gerçekleştirilmesi için bir takım makine öğrenmesi algoritmalar ve çok değişkenli istatistiksel analizler kullanılmaktadır.

Yarı denetimli öğrenme

Denetimsiz öğrenme ve denetimli öğrenme arasında bir köprü kuran ve hem etiketlenmiş hem de etiketlenmemiş verilerden yararlanarak modellerin eğitildiği bir makine öğrenimi yaklaşımıdır. Bu yöntem, etiketli verilerin sınırlı olduğu durumlarda, büyük miktarda etiketsiz veriden de faydalanarak modellerin performansını önemli ölçüde arttırır. Yarı denetimli öğrenme, genellikle iki adımda gerçekleşir: Öncelikle, bir miktar etiketlenmiş veriyle bir başlangıç model eğitilir ve ardından bu model, etiketlenmemiş veriyi kullanarak kendisini geliştirir. Bu süreçte, sözde etiketleme, kendi kendini eğitme ve graf tabanlı yöntemler gibi çeşitli teknikler kullanılabilir [21].

Diyelim ki bir e-ticaret şirketi, müşterilerin ürünler hakkındaki duygu ve tercihlerini anlamak için yorumları analiz etmek istiyor. Mevcut yorumların çoğu etiketlenmemiş. Etiketlenmiş olanlar ise sadece küçük bir kısmı oluşturuyor. Bu durumda yarı denetimli öğrenme kullanılabilir. Öncelikle, küçük bir etiketlenmiş veri setiyle bir duygu analizi modeli eğitilir. Ardından, bu model, etiketlenmemiş yorumlar üzerinde tahminler yapmak için kullanılır. Tahminler daha sonra elle doğrulanır ve doğrulanmış olanlar etiketlenmiş veri setine eklenir. Bu doğrulanmış verilerle model yeniden eğitilir ve performansı artırılır. Bu süreç, modelin daha fazla etiketlenmiş veriyle geliştirilmesi ve daha geniş bir yelpazede yorumları analiz etmesi için tekrarlanır. Sonuç olarak, şirket müşteri geri bildirimlerini daha etkili bir şekilde değerlendirebilir ve ürünlerini geliştirmek için daha iyi kararlar alabilir [21].

Takviyeli öğrenme, pekiştirici ve destekleyici öğrenme olarak ta isimlendirilmektedir. Takviyeli öğrenme hedef odaklı öğrenme olup maksat en iyi sonuçlar elde edecek eylemlerin yapılmasının teşvik edilmesidir. Bu türden öğrenim bir nevi deneme-yanılma yöntemine benzemektedir [1]. Takviyeli Öğrenme, denetimli öğrenme yönteminde olduğu gibi öğrenme danışman kontrolünde gerçekleştirilir

ancak denetimsiz öğrenme yönteminde olduğu gibi çıktı verileri ağa verilmemektedir. Bu yöntemde gerçekleşen her öğrenme sonucunda danışman doğru-yanlış, yeterli-yetersiz gibi bir skor oluşturularak öğrenme sürecinin tamamlanması veya devam etmesine karar verir. Bazı popüler takviyeli öğrenme teknikleri Q-learning, State-Action-Reward-State Action(SARSA), Deep Q Network(DQN)'dır [14].

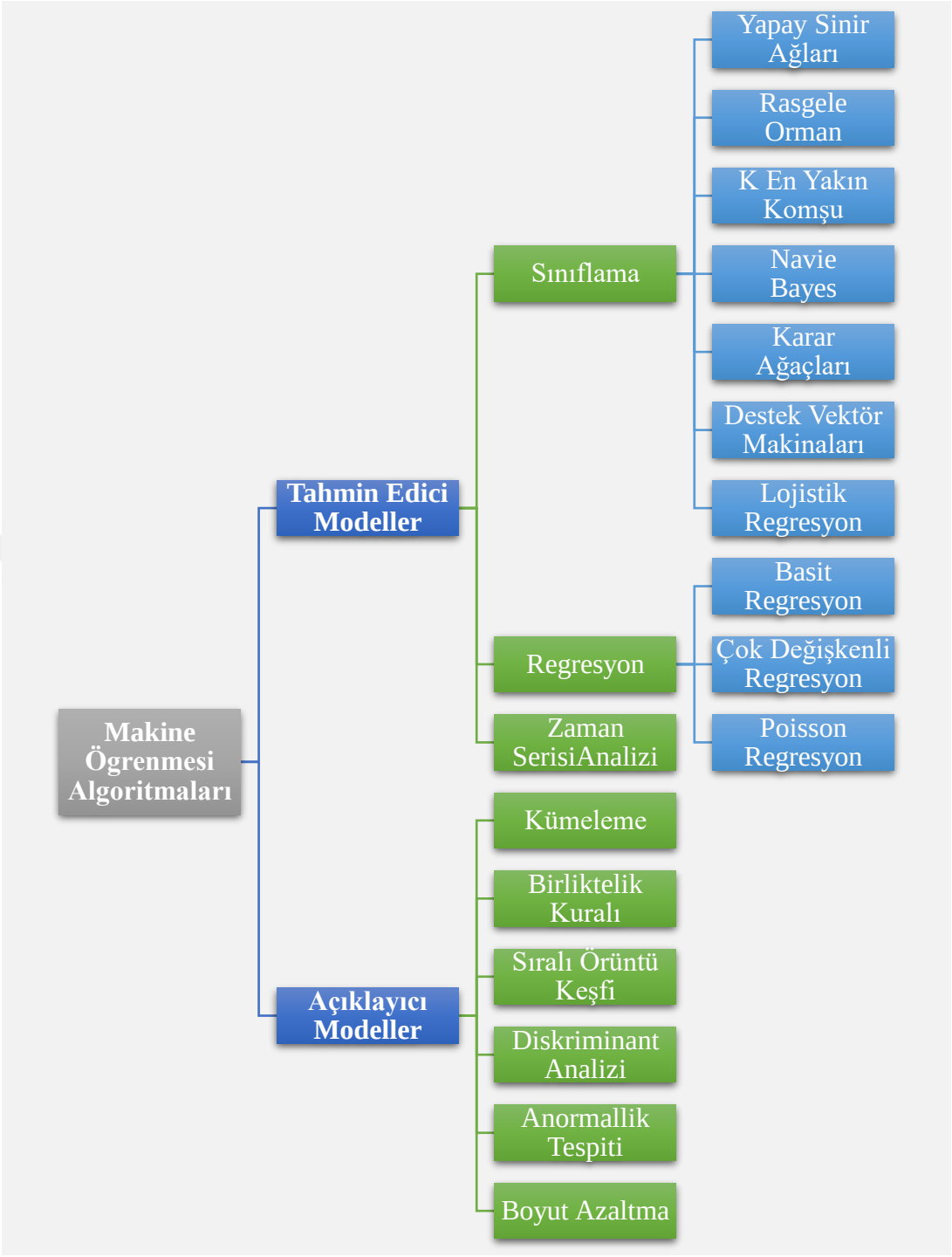
2.1.11 Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritmaları da aslında bir klasik algoritma çeşidi olmakla beraber klasik algoritmalardan bazı yönlerden ayrılmaktadır. Klasik algoritmaların en önemli özelliklerinden birisi açık, net, sonlu olması iken makine öğrenmesi algoritmalarında her zaman problemlerin çözümü için net ve sonlu çözümler olmayabilir. O nedenle klasik algoritma yetenekleri ile çözülemeyen problemler, bunlar kimi zaman yapay zekâ problemleri olarak da karşımıza çıkar, makine öğrenmesi algoritmaları ile çözülebilmektedir [1]. Başka bir deyişle makine öğrenmesi algoritmaları klasik algoritmaları tamamlayıcı ve geliştirici nitelik taşımaktadır.

Makine öğrenmesi algoritmaları oluşturulacak modelin türüne göre farklılık arz etmektedir. Bazı algoritmalar sadece bir modele özgülenmişken bazı algoritmalar ise birden fazla model oluşumuna fırsat tanımaktadır. Örneğin naive bayes algoritması ile sadece sınıflama modeli oluşturulurken, rasgele orman ve karar ağaçları gibi algoritmalar ile hem sınıflama hem de regresyon modele oluşturulabilir. Ayrıca lojistik regresyon algoritması ile sınıflama işlemi yapıldığı gibi regresyon işlemi de yapılabilir. Bununla birlikte lojistik regresyon algoritması aynı zaman da boyut indirgeme işlemi ile açıklayıcı bir model oluşturulmasına da fırsat tanımaktadır.

Makine öğrenmesi algoritmaları meydana getirildiği matematiksel ve istatistiksel disiplinler yönünden de farklılık arz etmektedir. Örneğin naive bayes, lojistik regresyon, çok değişkenli regresyon algoritmalarında istatistiksel yöntemler kullanılarak oluşturulmuştur. Karar ağaçları, k en yakın komşu ve rasgele orman algoritmaların da ise matematiksel yöntemler tercih edilmiştir.

Günlük hayatta karşılaşılan bir sorunun çözümünde ya da elde edilen etkinin, karın ve faydanın arttırılmasında veya iş ve işlemleri kolaylaştıracak bir teknolojinin geliştirilmesinde makine öğrenmesi algoritmaları sıklıkla kullanılmaktadır.



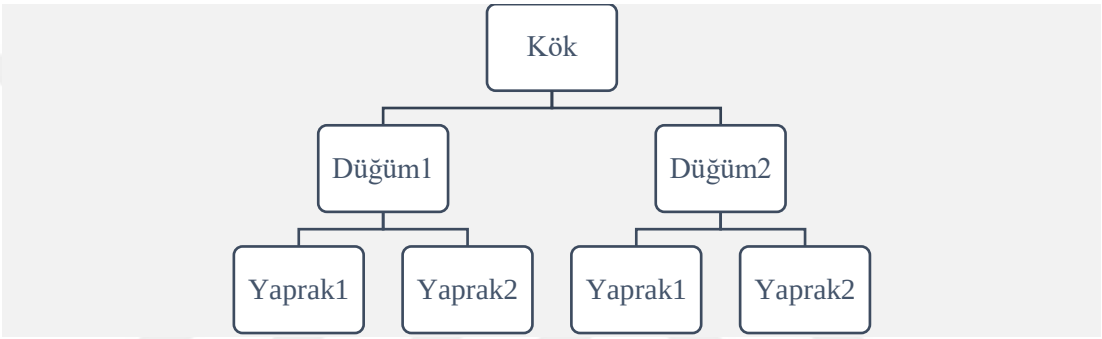
Şekil 2.2 Makine Öğrenmesi Algoritmaları

Makine öğrenmesi algoritması işin amacına uygun olarak tek kullanıldığı gibi birden fazla makine öğrenmesi algoritması birlikte kullanılarak model performansının arttırılması hedeflenmektedir. Makine öğrenmesi algoritmalarına ilişkin akademik yayınlar incelendiğinde çok fazla ana ve ana algoritmaya bağlı alt algoritma türleri olduğu gözlemlenmektedir. Örneğin karar ağaçları algoritmasının birden fazla türü bulunmaktadır. Dolayısıyla bu tez çalışmasında bu algoritmaların tamamı

incelenmemiş olup sadece tez çalışmasında kullanılan denetimli öğrenme algoritmalarına yer verilmiştir.

Karar Ağaçları(Decision Tree) Algoritması

Karar ağacı bir denetimli makine öğrenmesi algoritması olup hem sınıflandırma hem de regresyon için kullanılır. Bir karar ağacı, kararları ve karar almayı görsel ve açık bir şekilde temsil etmek için kullanılır. Karar ağacı algoritmasında tahminci değişkenler düğümleri, hedef değişkenler yaprakları oluşturur ve model ağaç veri yapısı ile sunulur, Kökten yapraklara kadar bilgi kazancı metriğine dayalı olarak bölümleme yapılarak ağaç meydana getirilir.



Şekil 2.3 Karar Ağacı Oluşum Şeması

Bir karar ağacı algoritmasında kök düğümün yaprak düğüme ayrılması şu şekilde açıklanabilir; Kök Düğüm; Kök düğüm tüm veri kümesini temsil eder ve ağacı başlatmak için kullanılır. Ağacın başlangıç noktasıdır ve verileri maksimum bilgi kazancı veya minimum Gini Impurity sağlayan özelliğe göre böler. İç Düğüm; Her bir iç düğüm, verileri iki veya daha fazla alt kümeye ayıran bir özelliği temsil eder. Bölme işlemi özelliğin değerine göre gerçekleştirilir ve her bir gözlemin izleyeceği yolu belirler. İç düğüm daha sonra birden fazla alt düğüme bölünür. Yaprak Düğüm; Yaprak düğüm, verilerin daha fazla bölünemeyen bir alt kümesini temsil eder. Kendisine ulaşan gözlemler için nihai tahmini içerir. Tahmin, alt kümedeki çoğunluk sınıfına veya hedef değişkenin ortalama değerine dayanır [22]. Karar ağaçları ile sınıflandırma yapılırken, bir veri kümesi giderek daha küçük alt kümelere ayrılarak karar ağacı kademeli olarak geliştirilir. Sonuç olarak elde edilen ağaç, karar düğümleri ve yaprak düğümlerinden oluşan bir karar ağacıdır. Oluşturulan karar ağacında bir karar düğümünün iki veya daha fazla sayıda elemanı vardır. Yaprak düğümler ise bir sınıfı veya kararı temsil eder. Karar ağacı ile sınıflandırma yapılırken, veri setinden

elde edilen karar ağacının olabildiğince az sayıda düğümden oluşturulması hedeflenmektedir [23].

Karar ağaçlarında kökten yapraklara kadar bölünmenin geri planında ağaç tümevarımı bulunmaktadır. Ağaç tümevarımının dayanak noktası bilgi kazancı metriğidir. Bilgi kazancı bir düğümün bölünme kararıyla ilgilidir [1]. Bilgi kazancı eğer pozitif bir değerse ağaç yeni düğümlere bölünür, aksi takdirde ağaç yeni düğümlere bölünmez. Homojenlik ölçümünde de gini index, entropy ve missclassification error gibi değerlere göre karar verilir [1]. Gini indexi 0-0,5 aralığında entropy ise 0-1 aralığında değer alır. C4.5, ID3, CART, SLIQ, SPRINT ve Hunt's Algorithm gibi karar ağacı tümevarımları bulunmaktadır.

Karar Ağaçlarının Avantajları [24],

- Anlaması ve yorumlaması kolaydır. Kullanılan ağaç yapılar görselleştirilebilir.
- Bir karar ağacı, verilerin normalleştirilmesini gerektirmez.
- Bir karar ağacı, verilerin ölçeklendirilmesini de gerektirmez.
- Az oranda bir veri hazırlığına ihtiyaç duyar. Fakat unutulmamalıdır ki bu model kayıp değerleri desteklememektedir.
- Kullanılan ağacın maliyeti, ağacı eğitmek için kullanılan veri noktalarının sayısıyla logaritmiktir.
- Hem sayısal hem de kategorik verileri işleyebilir.
- Çok çıktılı problemleri ele alabilmektedirler.
- İstatistiksel testler kullanılarak bir modelin doğrulanması mümkündür.
- Karar ağaçları, parametrik olmayan bir yöntem olarak düşünülebilir. Yani uzay dağılımı ve sınıflandırma yapısı hakkında bir yaklaşıma sahip değillerdir.

Karar Ağaçlarının Dezavantajları [24],

- Verilerdeki küçük bir değişiklik, karar ağacının yapısında büyük bir değişikliğe neden olarak kararsızlığa neden olabilir.
- Veriyi iyi bir şekilde açıklamayan aşırı karmaşık ağaçlar üretilebilir. Bu durumda ağaç dallanması takip edilemeyebilir.
- Bir Karar ağacı için bazen hesaplama, diğer algoritmalara kıyasla çok daha karmaşık olabilir. Karar ağacı genellikle modeli eğitmek için daha fazla zaman

gerektirir. Karar ağacı eğitimi, karmaşıklığı ve aldığı zaman daha fazla olduğu için nispeten pahalıdır.

- Ezbere öğrenme yaşanabilir ("overfitting"). Bu problemin çözümü için model parametrelere kısıtlamalar ve budama gibi yöntemler kullanılabilir. Budama işlemi, az sayıda nesneyi barındıran yaprak düğümlerin karar ağacı grafiğinden atılmasını ifade etmektedir.
- Karar Ağacı algoritması, regresyon uygulamak ve sürekli değerleri tahmin etmek için yetersizdir.

Karar Ağacı algoritması hangi verilerde daha başarılı sonuçlar vermektedir? [24],

- Sonlu sayıda öznitelik olduğunda, feature'lar, az sayıda kategorik değişkenler (örneğin: sarı, mavi, mor) içerdiğinde.
- Hedef değişkeni binary (1 ve 0) olduğunda. Ancak yine de ikiden fazla hedef değişkeni olduğunda da başarılı tahminlerde bulunabilir.
- Karar ağaçları doğal olarak ayrık ifadeleri temsil eder.
- Eğitim verileri hatalar içerebilir. Örneklerin sınıflandırılmasındaki veya bu örnekleri tanımlayan öznitelik değerlerindeki hatalar, karar ağaçları tarafından iyi bir şekilde işlenir ve onları sağlam bir öğrenme yöntemi haline getirir.
- Eğitim verileri Nan değerleri içerebilir. Karar ağacı yöntemleri, bazı eğitim örneklerinin bilinmeyen değerleri olduğunda bile kullanılabilir (örneğin, örneklerin yalnızca bir kısmı için nem bilinmektedir).

Rasgele Orman(Random Forest) Algoritması

Rastgele orman algoritması, ilk olarak Leo Breiman ve Adele Cutler tarafından literatüre kazandırılmıştır. Rastgele orman algoritması hem regresyon analizine hem de sınıflandırma problemlerine dayanmaktadır. Rastgele orman algoritması her özniteliği ağaç gibi düşünerek, tüm veri setini orman gibi düşünen tahmin edicilerden oluşan bir topluluktur [25].

Rastgele ormanlar, her ağacın bir değere bağlı olduğu ağaç tahminlerinin bir kombinasyonudur [26]. Random Forest (Rastgele Orman) algoritması; birden çok karar ağacı üzerinden her bir karar ağacını farklı bir gözlem örneği üzerinde eğiterek çeşitli modeller üretip, sınıflandırma oluşturmanızı sağlamaktadır. Kullanım kolaylığı ve esnekliği; hem sınıflandırma hem de regresyon problemlerini ele aldığı için benimsenmesini ve kullanımının yaygınlaşmasını hızlandırdı. Algoritmaya yönelik en

beğenilen nokta ise; veri kümeniz üzerinde çeşitli modellerin oluşturulması ile kümenizi yeniden ve daha derin keşfetme imkânı sunmasıdır [27].

Rastgele orman algoritmasının sınıflandırması şu şekilde çalışmaktadır [25]:

- Rastgele ağaç sınıflandırıcısı, öznitelikleri girdi vektörü olarak alır, veri setindeki her bir veriye göre sınıflandırır ve değer olarak en çok girdi değerini alan değişkeni etiketler. Ana gövdeden yapraklara doğru bir ilerleme söz konusudur. Herhangi bir gerileme durumunda, elde edilen bilgi, ormandaki tüm ağaçların elde ettiği yanıtların ortalamasıdır.
- Bu durumda, herhangi bir doğruluk tahmin prosedürüne gerek yoktur. Sonrasında, çapraz doğrulama, ön yükleme ve ayrı bir eğitim hatası ele alınmaz. Hata, ağaç sınıflandırması sırasında test edilmektedir.

Rastgele Orman Regresyonunun Uygulamaları [28],

- Sürekli sayısal değerleri tahmin etmek: Ev fiyatlarını, hisse senedi fiyatlarını veya müşteri yaşam boyu değerini tahmin etmek.
- Risk faktörlerinin belirlenmesi: Hastalıklar, finansal krizler veya diğer olumsuz olaylara ilişkin risk faktörlerinin tespit edilmesi.
- Yüksek boyutlu verileri işleme: Çok sayıda giriş özelliğine sahip veri kümelerini analiz etme.
- Karmaşık ilişkileri yakalama: Giriş özellikleri ile hedef değişken arasındaki karmaşık ilişkileri modelleme.

Rastgele Orman Regresyonunun Avantajları [28],

- Kullanımı kolaydır ve karar ağacına göre eğitim verilerine daha az duyarlıdır.
- Karar ağacı algoritmasından daha doğrudur.
- Birçok özelliğe sahip büyük veri kümelerinin işlenmesinde etkilidir.
- Eksik verileri, aykırı değerleri ve gürültülü özellikleri işleyebilir.

Rastgele Orman Regresyonunun Dezavantajları [28],

- Modelin yorumlanması da zor olabilir.
- Bu algoritma, karar ağaçları sayısı, her ağacın maksimum derinliği ve her bölmede dikkate alınacak özelliklerin sayısı gibi uygun parametrelerin seçilmesi için bazı alan uzmanlığı gerektirebilir.

- Özellikle büyük veri kümeleri için hesaplama açısından pahalıdır.
- Modelin çok karmaşık olması veya karar ağaçlarının sayısının çok fazla olması durumunda aşırı uyum sorunu yaşanabilir.

Naive Bayes Algoritması

Naive bayes algoritması bir denetimli öğrenme algoritması olup sınıflandırma işlemi gerçekleştirmektedir. Naive Bayes(Simple Bayes), veri setindeki özniteliklerin istatistiksel olarak birbirinden bağımsız olduğu varsayımı üzerine oluşturulmuş ve Bayes Teoremine dayanan bir sınıflandırma algoritmasıdır. Birçok sınıflandırma tekniğinde, öznitelikler arasında korelasyon olabileceği varsayılmaktayken, Naive Bayes sınıflandırıcılar öznitelikler arasında ilişki olmadığını varsayar [14]. Naive Bayes algoritması genellikle büyük veri setlerinde kullanılmaktadır. Genellikle verilerin sınıflandırılmasında ve literatürde spam maillerin filtrelenmesi gibi örnek olaylarda kullanılmıştır [25],

Algoritmanın temel fikri sınıflandırılacak bir kaydın bütün sınırlar için olasılığının hesap edilmesi ve en yüksek olasılığı veren sınıfa kaydın atanmasıdır [1]. Naive Bayes algoritması, bayes teoremine dayanır. A ve B olayları üzerinde Bayes Teoremi için aşağıdaki eşitlikler yazabiliriz.

$$P(A\backslash B)=\frac{P(A\cap B)}{P(B)} \text{ ve } P(B\backslash A)=\frac{P(A\cap B)}{P(A)} \quad \textbf{(1.1)}$$

Böylece, A ve B olaylarının birlikte gerçekleşme olasılığı ile B olayının olasılığını biliyorsak B gerçekleştiğinde A olayın gerçekleşme olasılığını bulabiliriz. Bu yaklaşım yardımıyla sınıflandırmada, eğitim verisinden elde edilen olasılıklara dayalı olarak test verisinin sınıfı bulunabilecektir [1]. Naive Bayes algoritmasının genel çalışma mantığı, aynı kriterlerin sonuca olan etkilerinin olasılıksal olarak hesaplanması temeline dayanmaktadır. Birçok yazılım uygulamasında kullanılmasının ana nedeni de budur [29].

Naive bayes sınıflandırıcılar hızlı çalışan ve kolay yorumlanabilen sınıflandırıcılardır. Ayrıca veri setindeki öznitelikler arasında bir korelasyon yok ise çok daha iyi performans göstermesi mümkündür. Bir diğer nokta ise veri setindeki öznitelik sayısındaki artışın birçok sınıflandırma algoritmasında bir yük oluşturduğu gerçeğidir. Bu dezavantajlı durumu, Naive bayes sınıflandırıcılar için bir avantaja dönüşebilmektedir ve iyi sonuçlar elde edilmektedir [14]. Naive Bayes algoritmasında

sıralı verilerin kullanılması tercih edilmez. Bu nedenle sıralı değerler içeren, bağımlı ya da bağımsız değişkenler kategorik verilere dönüştürülmelidir [29].

Naive Bayes Sınıflandırıcısının Avantajları,

- Her özellik birbirinden bağımsız kabul edildiği için lojistik regresyon gibi modellerden daha iyi performans gösterebilir [30].
- Az veriyle iyi işler başarabilir [30].
- Sürekli ve kesikli veriler ile kullanılabilir [30].
- Yüksek boyutlu verilerde iyi çalışabilir [30].
- Hızlı çalışan ve kolay yorumlanabilen sınıflandırıcılardır [14].
- Öznitelik artışı bir yük oluşturmaz [14].
- Dengesiz veri kümesinde başarısız sonuçlar verebilir [1].
- Değişkenler arasında korelasyon zayıfladıkça daha iyi sonuçlar verir [31].
- Hesaplama hızlı [31].
- Uygulanması kolay [31].
- Küçük veri kümeleri ile iyi çalışır [31].
- Naive varsayım mükemmel bir şekilde karşılanmasa bile iyi performans gösterir. Çoğu durumda, yaklaşım iyi bir sınıflandırıcı oluşturmak için yeterlidir [31].

Naive Bayes Sınıflandırıcısının Dezavantajları [31],

- Özellikler birbirinden bağımsız varsayılarak işlem yapıldığı için değişkenler arası ilişkiler modellenemez.
- Sıfır olasılık problemi ile karşı karşıya kalabilirsiniz. Sıfır olasılık istediğimiz örneğin veri setinde hiç bulunmaması durumudur. Yani herhangi bir işleme alındığında sonucu sıfır yapacaktır. Bunun için en basit yöntem tüm verilere minimum değer ekleyerek (genellikle 1) bu olasılık ortadan kaldırılabilir. Bu duruma Laplace kullanılarak tahminleme de denmektedir.

Destek Vektör Makinaları(Support Vector Machines)

Destek vektör makineleri algoritması ile sınıflandırma, regresyon ve aykırı değer tespiti yapılabilmektedir. Destek vektör makineleri gözetimli öğrenme temeline dayanan bir metottur [25]. Destek vektör makinaları, kalkulüs, vektör geometrisi ve kısıtlı en iyileme gibi matematik konularına dayanan; doğrusal ve doğrusal olmayan sınıflandırma problemlerini gerçekleştirmeye olanak tanıyan bir makine öğrenmesi

tekniğidir. İlk olarak AT&T laboratuvarlarında Vapnik ve arkadaşları tarafından geliştirilmiştir. Daha öncesinde ise 1963 yılında yine Vapnik tarafından geliştirilen bir algoritmaya dayanmaktadır [14].

Destek vektör makineleri denetimli öğrenme tekniğini kullanan bir algoritmadır. Algoritma tahminsel modellere uygun şekilde çalışır. Destek vektör makinaları nümerik giriş değerleri ve kategorik çıkış değerleri üretebilen ve sıklıkla sınıflandırma işlemlerinde kullanılan etkili bir algoritmadır [1]. Algoritmanın destek vektör makinesi adını almasının sebebi şudur. Karar yüzeyi adını verdiğimiz iki sınıfı birbirinden ayıran düzelmenin iki tarafında farklı sınıflara ait veriler bulunur. Bu veriler üzerinden geçen vektörlere destek vektörleri, destek vektörlerini ayıran çizgi veya düzleme de destek vektör makineleri adı verilir [1].

Destek vektör makineleri ile sınıflandırmanın temel amacı, iki boyutlu bir uzayda sınıflandırma yapmanın yanı sıra yüksek boyutlu öznitelikler uzayında da hiper düzlemlerin iyi bir şekilde ayrılarak en uygun sınıflandırmanın sağlanmasıdır [14]. Algoritmanın çalışması esnasında verilerin türüne bağlı olarak çekirdek fonksiyonlarda kullanılabilmektedir. Bu sayede hem doğrusal hem de doğrusal olmayan sınıflandırma işlemlerini gerçekleştirilebilmektedir. Eğer sınıflandırma işleminde, tam ayrıştırılabilir veriler kullanılırsa genellikle tüm veriler bir hiper düzlem ile sınıflandırılabilmektedir. Fakat eğer tam ayrıştırılamayan veriler kullanılırsa, çoğunlukla aynı boyutta tek bir düzlem ile sınıflandırılamamaktadır. Bu nedenle de farklı çekirdek fonksiyonları kullanılmaktadır [29]. Destek Vektör Makinesi Türleri karar sınırının doğasına bağlı olarak Destek Vektör Makineleri (SVM) iki ana bölüme ayrılabilir:

Doğrusal DVM: Doğrusal DVM'ler, farklı sınıfların veri noktalarını ayırmak için doğrusal bir karar sınırı kullanır. Veriler tam olarak doğrusal olarak ayrılabildiğinde doğrusal DVM'ler çok uygundur. Bu, tek bir düz çizginin (2B'de) veya bir hiperdüzlemin (daha yüksek boyutlarda) veri noktalarını tamamen ilgili sınıflara bölebileceği anlamına gelir. Sınıflar arasındaki marjı maksimuma çıkaran hiperdüzlem karar sınırıdır [32].

Doğrusal Olmayan DVM: Doğrusal Olmayan DVM, verileri düz bir çizgiyle iki sınıfa ayıramadığında (2B durumunda) sınıflandırmak için kullanılabilir. Doğrusal olmayan DVM'ler, çekirdek işlevlerini kullanarak doğrusal olmayan şekilde ayrılamayan

verileri işleyebilir. Orijinal girdi verileri, bu çekirdek işlevleri tarafından, veri noktalarının doğrusal olarak ayrılabileceği daha yüksek boyutlu bir özellik uzayına dönüştürülür. Bu değiştirilmiş alanda doğrusal olmayan bir karar sınırının yerini belirlemek için doğrusal bir DVM kullanılır [32].

Her algoritmanın kullanım alanı, avantajı, dezavantajı olduğu gibi destek vektör makinelerinin de avantajları ve dezavantajları bulunmaktadır.

Destek vektör makineleri algoritmasının avantajları şu şekilde sıralanabilir:

- Büyük hacimli veriler için uygundur. Özellikle, boyut sayısının özellik sayısından daha büyük olduğu durumlar için uygundur [25].
- Karar noktasında farklı çekirdek fonksiyonları kullanılabilmektedir ve hafıza açısından algoritma bunu tutabilmektedir [25].
- Dağılımı hakkında herhangi bir ön bilgiye sahip olunmayan veri setlerinde de yüksek başarımlar elde edilir [1].

Destek vektör makineleri algoritmasının dezavantajları ise şu şekilde sıralanabilir [25],

- Örnek sayısı, öznitelik sayısından az olduğu durumlarda çekirdek fonksiyonları seçiminden ve düzenlileştirme aktivitelerinden kaçınılması gerekmektedir
- Destek vektör makineleri diğer algoritmaların aksine olasılık tahminleri ortaya koymaz.

K En Yakın Komşu Algoritması(K Nearest Neighbor-KNN)

K-En Yakın Komşular (KNN) algoritması, sınıflandırma ve regresyon problemlerini çözmek için kullanılan denetimli bir makine öğrenme yöntemidir. Evelyn Fix ve Joseph Hodges bu algoritmayı 1951'de geliştirdiler ve daha sonra Thomas Cover tarafından genişletildi [33].

(K-NN) algoritması, öncelikle basitliği ve uygulama kolaylığı nedeniyle kullanılan, çok yönlü ve yaygın olarak kullanılan bir makine öğrenme algoritmasıdır. Temel veri dağılımı hakkında herhangi bir varsayım gerektirmez. Ayrıca hem sayısal hem de kategorik verileri işleyebilir, bu da onu sınıflandırma ve regresyon görevlerinde çeşitli veri kümesi türleri için esnek bir seçim haline getirir. Belirli bir veri kümesindeki veri noktalarının benzerliğine dayalı olarak tahminler yapan parametrik olmayan bir

yöntemdir. K-NN, diğer algoritmalarla karşılaştırıldığında aykırı değerlere karşı daha az duyarlıdır [33].

K-NN bilinen en eski ve basit sınıflandırma algoritmalarından birisidir. Algoritmanın genel mantığı, yeni gelen örnek bir veri ile K adet komşu arasındaki mesafenin ölçülmesine dayanmaktadır. K-NN algoritması denetimli bir algoritma olması nedeniyle kullanım esnasında veriler ilk olarak eğitim ve test olarak ikiye ayrılmaktadır. Sonrasında;

- K parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısını ifade etmektedir. K değeri, komşu sınıflar arasındaki uzaklık ve sınıflandırma performansını doğrudan etkilemektedir.
- Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı tek tek hesaplanır.
- Uzaklık hesaplanması için farklı fonksiyonlar kullanılabilir. Literatürde ise yaygın olarak Minkowski, Öklid, Manhattan ve Chebyschev fonksiyonları kullanılmaktadır.
- Hesaplanan uzaklıklardan en yakın K. komşu ele alınır. Öznitelik değerlerine göre K. komşunun sınıfına atanır.
- Atanılan sınıf, tahmin edilmesi beklenen sınıf değeri olarak kabul edilir. Yani yeni verinin sınıfı bulunmuş olur [29].

En yakın k komşu algoritması gerçek manada bir öğrenim algoritması değildir. Daha çok bir arama metodudur. Bununla birlikte en ham tekniklerinden birisidir. Çünkü referans olarak verinin kendisi kullanılır. N tane eleman içeren bir veri setinde her bir eleman için bir tahmin yapılmak istendiğinde her bir kaydı bir diğeriyle karşılaştırmak gerekmektedir. Bu ise büyük veri setleri için çok uygun değildir. Çünkü karmaşıklık çok fazladır [1].

KNN algoritmasında, algoritmadaki komşuların sayısını tanımlamak için k değeri çok önemlidir. K-en yakın komşular (k-NN) algoritmasında k değeri, giriş verilerine göre seçilmelidir. Giriş verilerinde daha fazla aykırı değer veya gürültü varsa, daha yüksek bir k değeri daha iyi olur. Sınıflandırmada bağlardan kaçınmak için k için tek bir değer seçilmesi önerilir. Çapraz doğrulama yöntemleri, verilen veri kümesi için en iyi k değerinin seçilmesine yardımcı olabilir [33]. K parametresinin seçiminde dikkat

edilmesi gereken önemli bir nokta ise K parametresinin sadece tek sayı olarak belirlenmesi zorunluluğudur.

KNN algoritmasının avantajları;

- Eğitim işlemlerinin diğer algoritmalara nazaran daha kolay olması [34],
- Süreçlerin ve analizlerin analitik/sayısal olarak takip edilebilir olması [34],
- Kompleks ya da karmaşık (Gürültülü) eğitim verilerine karşı etkili olması [34],
- Uyarlanabilirliğinin (uygulanabilirliği) kolay olması dile getirilebilir [34].
- Az sayıda Hiperparametre ile çalışabilir olması [33],
- KNN algoritmasının işleyişi gereği tüm verileri bellekte saklar ve dolayısıyla yeni bir örnek veya veri noktası eklendiğinde algoritma kendisini bu yeni örneğe göre ayarlar ve gelecek tahminlerine de katkıda bulunur [33].

KNN algoritmasının dezavantajları;

- İşlem hacmi ve işlem adımı fazla olduğundan dolayı yüksek donanıma ihtiyaç duymaktadır, (maliyet) [34],
- Yüksek hacimli verilere karşı dirençli olsa da adım ve işlem sayısı fazla olduğundan dolayı zaman almaktadır [34],
- Performans adına uygun algoritmanın bulunması kimi zaman dilimlerinde uzun sürmektedir, (Uzaklık denklemi, parametreler vb.) dile getirilebilir [34],
- Ölçeklenme ve aşırı uyuma eğilim sorunu. Bu sorunun üstesinden gelmek için genellikle özellik seçimi ve boyut azaltma teknikleri uygulanır [33].

Yapay Sinir Ağları(Artificial Neural Networks)

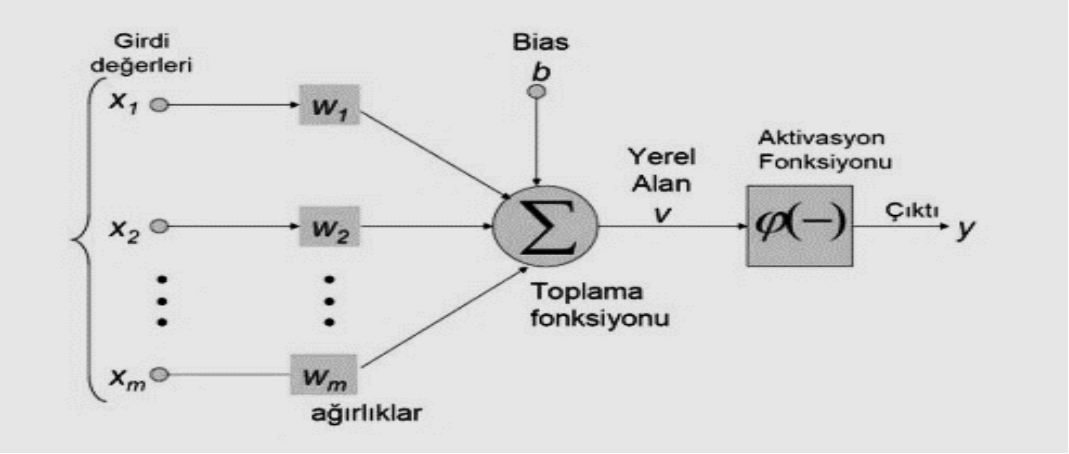
İlk yapay sinir ağı modeli 1943 yılında, bir sinir hekimi olan Warren McCulloch ile bir matematikçi olan Walter Pitts tarafından gerçekleştirilmiştir [35]. Yapay Sinir Ağları(Artificial Neural Networks), makine öğrenmesinde yaygın kullanıma sahip bir modeldir. Temeli 1940'lı yıllara dayanan yapay sinir ağları hakkında geçmişte birçok çalışma yapılmış ve bu çalışmalarda önemli başarılar elde edilmiştir. Yapay sinir ağları, son yıllarda popüler olan deri öğrenme çalışmalarının da temelini oluşturmaktadır [14].

Geçmişten günümüze yapay sinir ağları birçok alanda ve uygulamalarda kullanılmaktadır. Bu uygulama alanlarına; üretim planlama, kan analizlerinin sınıflandırılması, beyin modellemesi çalışmaları, kalite kontrolü, parmak izi tanıma,

otomatik araç denetimi, kredi kartı hilelerini saptama, zeki araçlar ve robotlar için optimum rota belirleme, mekanik parçaların ömürlerinin tahmin edilmesi, ses tanıma, denetim, meteorolojik yorumlama, elektrik işareti tanıma, el yazısı tanıma, hastalıkların tanımlanması ve tedavisi, radar ve sonar sinyalleri sınıflandırma, spam maillerin filtrelenmesi olarak örnekler verilebilir [36].

Bilim ve teknolojinin gelişmesinde mevcudattaki çalışma sistemler defaten örnek alınmıştır. Yapay sinir ağları da biyolojik sinir ağaları örnek alınarak yapılmış ve geliştirilmiştir. Tipik bir sinir hücresi ya da nöron, dentrit, hücre gövdesi, aksonlar ve sinapslardan oluşmaktadır. Dendiritler nöral iletişimin önemli alıcılarıdır. Bir nörondan diğerine geçen mesajlar, mesajı yollayan hücrenin terminal butonlarıyla mesajı alan hücrenin dendirit membranı ya da soma (hücre gövdesi) bölümü arasındaki birleşme yerleri olan sinapslar aracılığıyla iletilir/transfer edilir [37]. Yapay sinir ağlarında dentritlere karşılık toplama fonksiyonu, hücre gövdesine karşılık transfer fonksiyonu, aksonlara karşılık yapay nöron çıkışı ve sinapslara karşılık ağırlıklar bulunmaktadır. Bu sistemler ise biyolojik sinir sistemi taklit edilerek yapay sinir sistemi tasarlanmıştır.

Yapay sinir ağları ile yapay zekâ alanındaki birçok probleme çözüm üretmek mümkündür. Onunla sınıflandırma, kümeleme ve optimizasyon başta olmak üzere birçok alandan problemler çözebilirsiniz [1]. Yapay sinir ağları, normalizasyon, temizleme ve kümeleme analizleri sonrasında veriler arasındaki ilişkiyi tahmin etmek için kullanılmaktadır. Denetimli ağ yapısı olan perceptron ağ yapısı kullanılmaktadır. Bu ağ yapısında, yeni hesaplanan çıktı değeri, her aşamadaki gerçek değerlerle ilerlemektedir [25].



Şekil 2.4 Yapay Sinir Ağı Çalışma Sistemi

En temel ve basit bir yapay sinir ağı girdi katmanı, gizli/ara katman ve çıktı katmanından oluşmaktadır. Bu yapı daha detaylı olarak girdiler, ağırlıklar, toplama/transfer fonksiyonu, aktivasyon fonksiyonu ve çıktılardan oluşmaktadır. Bir yapay sinir ağ çalışma sistemi ise genel olarak şu şekilde gerçekleşmektedir. Girdi katmanı ile alınan girdi değleri(x_m) ara/gizli katmana aktarılır burada ağırlıklar(w_m) ile çarpılarak işlenir daha sonra eşik değer(bias) ile toplanarak aktivasyon fonksiyonuna gönderilir ve y değeri olarak çıktı katmanın da işlenmiş veriler çıkar. Bu yapay sinir ağı eğer geri beslemeli bir özelliğe sahip ise bu işleme geriye doğruda işleyerek tekrarlanır.

Yapay sinir ağları katman sayısına göre tek katmanlı ve çok katmanlı algılayıcılar olarak ikiye ayrılmaktadır. Tek katmanlı algılayıcılar yalnızca girdi ve çıktı katmanlarından oluşmaktadır. Tek katmanlı ağ yapısına sahip ağ modellerinde bir veya birden fazla girdi girebilir fakat tek bir çıktı oluşturulur. Doğrusal ilişkinin olduğu durumlarda etkilidir ve Perceptron, Çoklu Adaptif Doğrusal Eleman (Madalıne) ve Adaptif Doğrusal Eleman (Adalıne) en önemli tek katmanlı algılayıcılardır. Çok Katmanlı Algılayıcılar, girdi katmanı, ara (gizli) katmanlar, çıktı katmanından oluşmaktadır. Çok katmanlı algılayıcıların gizli katmanlarında kullanılan doğrusal olmayan aktivasyon fonksiyonları sayesinde girdiler ve çıktıların arasında doğrusal ilişkinin olmadığı durumlarda kullanılmaktadır. Kullanılan veri kümesinin karmaşıklık düzeyine göre gizli katman sayısı değişiklik gösterebilir.

Bağlantı şekline göre yapay sinir ağları ileri beslemeli yapay sinir ağı ve geri beslemeli yapay sinir ağı olmak üzere ikiye ayrılır. İleri beslemeli yapay sinir ağlarında nöronlar girişten çıkışa doğru düzenli katmanlar şeklindedir. Bir katmandan sadece kendinden sonraki katmanlara bağ bulunmaktadır. Yapay sinir ağının girişine gelen bilgiler bir değişime uğratılmadan orta noktaya diğer bir deyişle gizli katmandaki hücrelere iletilir. Daha sonra sırasıyla çıkış katmanından işlenerek geçer ve dış ortama aktarılır [38]. Geri Beslemeli Yapay Sinir Ağları: En az bir hücrenin çıkışı, diğer herhangi bir hücreye giriş olarak verilir bundan dolayı genellikle geri beslemeli yapay sinir ağlarında bir geciktirme eleman üzerinden yapılır. Besleme işlemi bir katmandaki hücreler arasında olmayabilir, bu sebeple doğrusal ilişkinin varlığından söz edilemez. Bu sebeple yapay sinir ağlarının geri beslemenin yapısı veri setine göre değişkenlik gösterebilir [39].

Yapay Sinir Ağları avantajları olarak [40];

- Bir kez eğitildiklerinde yeni bir veri kümesini doğrudan analiz edebilirler, (Örüntü Deneyimleri)
- Yüksek hassaslıkta örüntülü ilişkilendirme ve genel sınıflandırma için ideal bir yöntem olarak nitelendirilmektedirler,
- Ağırlık ve Ağ Yapısı gibi analiz modelindeki dinamiklerin değişiminde kendilerini yeni modele adapte edebilirler,
- Doğrusal olmayan analizler için yüksek doğruluk değerleri sağlamaktadırlar, YSA hücreleri doğrusal değildir,
- Algılamaya ve eğitilmeye yönelik güçlü bir algoritma olarak görülmektedirler.

Yapay Sinir Ağları dezavantajları olarak [40];

- Yüksek nitelikte donanım bağımlı algoritmalardır, (Yetersizlikte güvenilirlik sağlamaz.)
- Analiz edilecek probleme yönelik ağ yapısını deneme yanılma yöntemi ile elde edebilmektesiniz, (Zaman kaybı oluşturabilir.)
- Ele alınacak girdi parametreleri-değerleri için herhangi bir kural bulunmamaktadır,
- Algoritmanın çıktı oluşturma süresi önceden net olarak öngörülememektedir,
- Her durumda yüksek doğruluk değeri elde edileceği garantisi verilememektedir.

2.1.12 Model Geçerleme/Doğrulama Yöntemleri

Makine öğrenmesi algoritmaları üzerinden makinenin öğrenmesi için makinenin eğitilmesi gerekmektedir. Makinenin eğitilmesi işlemi ise model geçerleme/doğrulama yöntemleri olarak adlandırılmaktadır. Makinenin eğitilmesi için parametrik ve parametrik olmayan farklı yöntem bulunmaktadır.

Parametrik Yöntemler: Bu yöntemlerde veri setindeki verilerin dağılımının belli bir formülle ifade edilebildiği varsayılır. Örneğin; Holdout ve Simple Cross Validation gibi yöntemler verişe tinin belirli bir yüzdesini test verisi, geri kalanını ise eğitim verisi olarak kullanır. Non-Parametrik (Parametrik Olmayan) Yöntemler: Bu yöntemler, veri setindeki verilerin dağılımının belli bir formülle ifade edilemediği varsayılır. Örneğin; K-Fold Cross Validation, Leave One Out Cross Validation, Time Series Cross Validation, Bootstrap gibi yöntemler veriş etinin tümünü eğitim ve test için kullanır

[41]. Parametrik ve parametrik olmayan birçok yöntem bulunmakla beraber literatürde sıklıkla kullanılan Holdout Ve K-Fold Cross Validation yöntemleri incelenmiştir.

Holdout geçerleme yöntemi

Holdout yönteminde veri seti eğitim(trade) ve test(test) verisi olarak ikiye ayrılır. Bu ayırma işlemi genellikle %80 eğitim-%20 test veya %70 eğitim-%30 test şeklinde olur. Veri setinin eğitim-test şeklinde ayrılmasında mutlak sınırlamalar bulunmamakta olup araştırmacının farklı ayrımlar ile yüksek performans denemesi yapmasına olanak sağlamaktadır. Kimi zamanda aşırı uyumu(overfitting) önlemek için veri seti eğitim, test ve doğrulama(validation) olmak üzere üç kısıma ayrılır.[5] Eğitim setindeki veriler ile makine öğrenmesi modeli eğitilir. Eğitim sonunda yeterli performans elde edilirse test setindeki veriler(sistemin daha önce görmediği veriler) ile model test edilir [14]. Hold-out yöntemi; veri setinin yeterli büyüklükte olduğu durumlarda kullanışlıdır. Çünkü test verileri yeterli sayıda veri noktası içermelidir, aksi durumda test verileri yeterli sayıda veri noktası içermeyebilir ve modelin doğruluğu yanıltıcı olabilir [41].

K-Katlı Çapraz Doğrulama(K-Fold Cross Validation)

Veri setinin eğitim seti ve test seti olarak ayrılması esnasında veri setinin dağılımda oluşabilecek düzensizlikler, makine öğrenmesi modelinin performansını olumsuz etkileyebilmektedir. Bu problem k-katlı çapraz doğrulama(k-fold cross validation) yöntemi ile çözüme kavuşturulmaktadır. Çapraz doğrulamada veri seti k olarak ifade edilen sayı kadar parçaya ayrılır. Daha sonra her adımda k-1 adet parça makine öğrenmesi algoritmasında eğitilir ve kalan parça üzerinde test edilir. Burada önemli olan nokta, her adımda daha önce denenmemiş parçanın test verisi olarak kullanılmasıdır. Son olarak da her adım neticesinde elde edilen hata değerlerinin ortalaması alınarak modelin toplam hatası elde edilmiş olur [14].

Model eğitiminin k-katlı çapraz doğrulama yöntemi ile gerçekleştirildiğinde dikkat edilecek husus ise veri setinin dağılımıdır. Veri seti bir kurala bağlı olarak oluşturulduğunda k-katlı çapraz doğrulama ile inşa edilen modelin orta ve düşük bir doğruluk performansı göstermesidir.

[5] Overfitting durumunda model eğitim verisini ezberlemiş olup modele test verisi uygulandığında model düşük performans gösterecektir. Başka bir ifadeyle test verisi üzerinden elde edilen performanslar düşük çıkacaktır. Aşırı öğrenmenin ortadan kaldırılması için bir takım yöntemler bulunmaktadır.

2.1.13 Model Performansı Değerlendirme Yöntem ve Ölçüleri

Makine öğrenmesi algoritmaları üzerinden gerçekleştirilen modellerin performansının ölçülmesi çeşitli yöntemler bulunmaktadır. Model performans değerlendirme ölçütleri kullanılarak uygulanan modelin ne ölçüde performans sağladığı görülebilmektedir. Bu sayede;

- Geliştirilen modelin uygulamaya geçirilip geçirilmeyeceğine,
- Birden fazla modelin geliştirildiği durumlar için, hangi modelin daha başarılı olduğuna karar verilebilmektedir [42].

Bir modelin uygulamaya konulup konulmayacağı model başarısı ile doğrudan ilişkilidir. Bu noktada model başarısının değerlendirilmesi önem arz etmektedir. Bu bağlamda sınıflandırma modellerinin başarısının ölçümünde hata matrisi(Confision Matrix) ile çeşitli performansa değerlendirme ölçütleri geliştirilmiştir. Bu ölçütler doğruluk, kesinlik, duyarlılık, F-ölçütü ve ROC eğrisi olarak sayılabilir.

Hata matrisi, bir sınıflandırıcının farklı sınıf etiketlerini ne ölçüde sınıflandırabildiğini gösteren bir analiz aracıdır. Bir veri setindeki sınıf etiketi sayısı k adet ise hata matrisi k*k boyutundan oluşan bir matris olarak düşünülebilir [14].

Tablo 2.3 Hata Matrisi(Confision Matrix)

		TAHMİN	
		0	**1**
GERÇEK	**0**	(TN) TRUE POSİTİVE	(FP) FALSE POSİTİVE
	1	(FN) FALSE NEGATİVE	(TP) TRUE POSİTİVE

Doğruluk(Accuracy): Doğru tahmin edilen örneklerin sayısının tüm örnek sayılarının toplamına oranı. (Doğruluk=TN+TP/TOPLAM)

Kesinlik(Precision): Doğru tahmin edilen pozitif örnek sayısının, pozitif olarak tahmin edilen örnek sayısına oranı. (Kesinlik=TP/TP+FP)

Duyarlılık(Recall): Doğru tahmin edilen pozitif örnek sayısının, gerçekte pozitif olan örnek sayısına oranı.(Duyarlılık= TP/TP+FN)

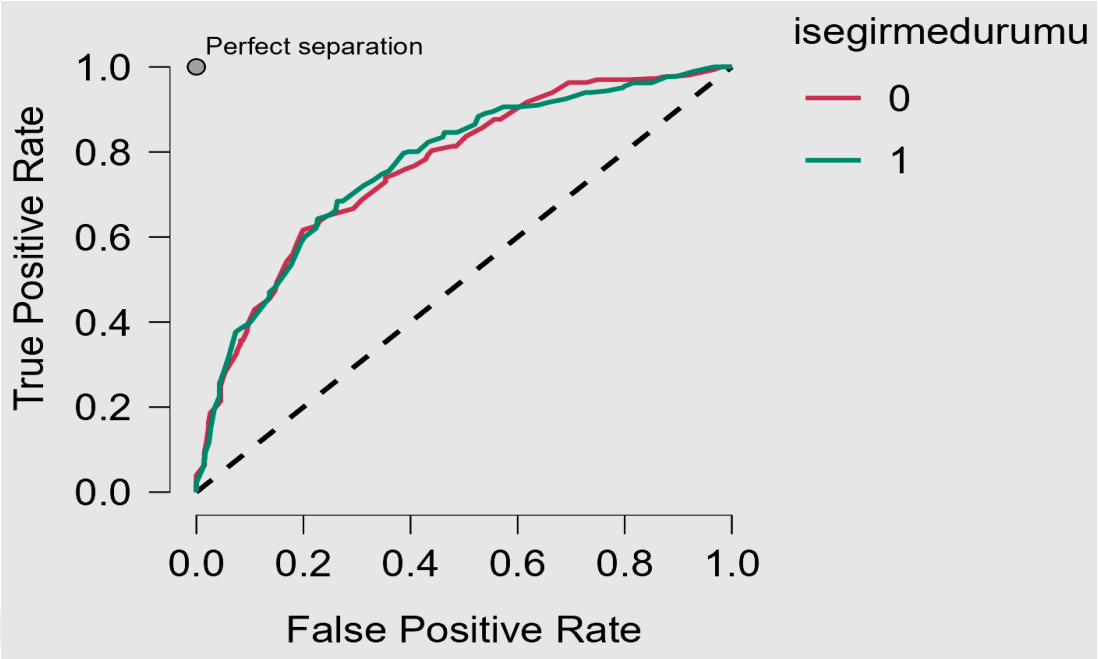
F-Ölçütü: Kesinlik ve duyarlılık ölçütlerinin harmonik ortalaması.

(F-Ölçütü= 2(*Precision*Recall)/(Precision + Recall))

Makine öğrenmesi ile oluşturulan sınıflandırma modelinin performans değerlendirilmesinde doğruluk ölçütü büyük önem arz etmektedir. Yüksek bir doğruluk sınıflaması gösteren bir model uygulamaya konulmak için büyük yol kat etmiştir. Ancak bu durum her zaman böyle olmaz. Diğer bir ifadeyle her yüksek doğruluk performansı iyi kabul edilmez. Örneğin veri setinin dengesiz dağıldığı durumda model yüksek bir sınıflama performansı göstermekle beraber aynı zamanda yanlı bir sonuçta göstermiştir. Bununla birlikte kimi zaman yüksek doğruluk performansı yüksek maliyeti de beraberinde getirilebilir. Böyle bir durumda yüksek doğruluk performansından ziyade maliyete göre optimum doğruluk performansı daha fazla değer kazandırabilir.

Ayrıca aşır öğrenme (overfitting) durumunda model eğitim verisinde yüksek doğruluk performansı göstermesine rağmen test verisinde düşük doğruluk performansı göstermektedir. Bazı modellemelerde pozitif örnek sayısının(Örneğin işe yerleşenlerin sayısı ya da kanserli hasta sayısı veya siber saldırı sayısı) doğru tahmin edilmesi daha fazla önem kazanabilir. Bu durumda ise kesinlik, duyarlık veya F-ölçütü doğruluk performansından daha fazla önem arz edebilir. Dolayısıyla tek başına doğruluk performansı üzerinden model performans değerlendirmesi yapmak yeterli değildir. Model performans değerlendirmesinde hata matrisinden elde edilen diğer ölçütlerle beraber model türü ve maliyeti de değerlendirmeye alınması gerekmektedir.

Elde edilen model değerlendirme ölçümlerine ek olarak başka bir işlem de bu süreçte karşımıza çıkmaktadır. Bunun adı çapraz doğrulamadır(cross-validation). Çapraz doğrulama, özellikle eğitim verisinin yetersiz olduğu durumda yeniden örnekleme yöntemi yardımıyla modelin eğitildiği ve değerlendirildiği sürecin adıdır [1]. Başka bir ifadeyle çapraz doğrulama ile hem model eğitilebilir hem doğruluk performans artışı sağlanabilir hem de aşırı öğrenme sorununa karşı önlem alınabilir. Ancak burada dikkat edilmesi gereken önemli husus veri setindeki kurallı/sistematik sıralanışa karşı çapraz doğrulama yönteminin aşırı duyarlı olmasıdır. Bu duyarlılık nedeniyle eğitilen model düşük performans göstermektedir. Böyle bir durumun aşılması için veri setindeki düzenli sıralanışın ortadan kaldırılması gerekmektedir. Diğer bir ifadeyle veri setindeki tesadüfilik arttırılmalıdır. Aksi takdirde çapraz doğrulama model performansını arttırmadığı gibi azaltacaktır.



Şekil 2.5 ROC Eğrisi Örneği

Model performans değerlendirmesinde bir diğer hata matrisi ölçütü ROC(Receiver Operating Characteristic) eğrisi analizidir. ROC eğrisine göre iki tür performans değerlendirmesi yapılır. Bunlardan birisi eğrinin sol üst köşeye yakın olmasıdır. ROC eğrisinde arzulanan görünür aslında eğrinin grafiğin sol üstüne yakın olmasıdır [14]. ROC eğrisinde yorum yapılırken bir diğer kullanılabilecek ölçüt ise eğrinin altında kalan alandır(area under the curve, AUC). Bu alan ne kadar büyükse performans o kadar iyidir yorumu yapılabilir. AUC değeri 0 ile 1 arasında değişirken, arzulanan ise AUC değerinin 1'e yakın olmasıdır. Örneğin AUC değerinin 0,74 olarak bulunması, kanser ve kanser değil sınıflarının %74 oranında sınıflandırıcı tarafından doğru sınıflandırılabildiği sonucunu ortaya koymaktadır [14]. Dengeli dağılıma sahip olmayan veri setlerinde yapılacak modelleme işlemlerinde etkin bir performans değerlendirme ölçütüdür.

2.1.14 Model Performans İyileştirme/Güncelleme

Makine öğrenmesi algoritmaları üzerinden gerçekleştirilen modellerin performansının arttırılması yâda iyileştirilmesi model devamlılığı için önem arz etmektedir. Model performans iyileştirme yetersiz performans durumunda yapılacağı gibi performans artırımı içinde yapılır. Yine veri setinde meydana gelen gelişimler nedeniyle de model güncelleme veya iyileştirmeye gidilebilir.

Model iyileştirme/Güncelleme faaliyetleri/süreçleri.

- ✓ Veri seti değişimi(Yığın veya örneklem veri seti ile modelleme).
- ✓ Veri artırımı(İlave veriler ile modelin güncellenmesi).
- ✓ Veri dengeleme(Veri dengeleme metotlar ile verinin dengelenmesi).
- ✓ Veri niteliğinde değişim(Sürekli verinin kategorik hale getirilmesi).
- ✓ Veri standardizasyonu(Veri ölçü biriminde tek düzeliğin sağlanması)
- ✓ Veri normalizasyonu(Veri dağılım fonksiyonunda tek düzeliğin sağlanması)
- ✓ Model algoritmaların da farklılık (Farklı tür makine öğrenmesi algoritmalarının kullanılması).
- ✓ Model geçerleme ve doğrulama tekniklerinde farklılaşma (Holdout, k çarpan vb.).
- ✓ Eğitim ve test verisi oranlarında değişime gitme(Eğitim/%80/Eğitim/%70).
- ✓ Doğrulama sayısında farklılaşma(10-kat yâda 5-kat çapraz doğrulama).
- ✓ Hiperparametrelerde değişim(gini/entropy, yakın komşu, uzaklık ölçümü vb.)
- ✓ Boyut indirgeme(Öznitelik sayısında değişim veya faktörleme)
- ✓ Eğitim verisini arttırma [1].
- ✓ Özellik değerlerinin yeniden ağırlıklandırılması [1].
- ✓ Özellik setini güncelleme [1].
- ✓ Model birleştirme [1]. Birden fazla makine öğrenmesi modeli ile daha iyi bir model oluşturma. Topluluk öğrenmesi
- ✓ Yeni veriler ile model güncelleme.
- ✓ Model tekniğinde değişim.(Makine öğrenmesi algoritmaları yerine çok değişkenli istatistiksel yöntemler ya da fuzzy gibi matematiksel yöntemler kullanılabilir).

Yukarıda yer alan işlemlerin biri ya da birden fazlası model amacına göre gerçekleştirilebilir. Çalışmalar bir geri besleme mantığı içerisinde en iyi sonuca ulaşana kadar devam ettirilmelidir [1]. Başka bir ifadeyle model iyileştirme/güncelleme işlemi bir sistem olup içerisinde iyileştirme faaliyetleri/süreçleri barındırır.

2.1.15 Makine Öğrenmesi Araçları

Makine öğrenmesi algoritmaları ile modelleme işlemini gerçekleştirmek için birçok makine öğrenmesi aracı bulunmaktadır. Bu araçların bir kısmı paket program şeklinde iken bir kısmı ise programlama dili şeklindedir. SPSS, RAPİD MİNER, WEKA, JASP, TANAGRA, ORANGE, MATLAB gibi paket programlar ile makine öğrenmesi

modellemeleri gerçekleştirilebilir. Bu programlar bir kısmı ücretli iken bir kısmı ücretsiz ve açık lisans kodludur. PHTYON, R, C gibi yazılım dilleri ile de makine öğrenmesi modellemeleri gerçekleştirilebilir. Makine öğrenmesi modellemelerine imkân tanıyan bu programların ve program dillerinin birbirine üstünlük gösteren tarafları olduğu gibi benzerlik ve farklılık gösteren tarafları da bulunmaktadır. Makine öğrenmesine araçlarının kullanımında özellikle PHTYON ve R dilleri son yıllarda oldukça yaygın kullanılmaktadır.

2.1.16 Çok Değişkenli İstatistiksel Analizler

Veri madenciliğinde modelleme tekniklerinden biriside çok değişkenli istatistiksel analizlerdir. İstatistiksel analizler tek değişkenli ve çok değişkenli istatistiksel analizler olarak iki parçaya ayrılabilir. TDK'ye göre istatistik(sayılmama) *"Bir sonuç çıkarmak için verileri yöntemli bir biçimde toplayıp sayı olarak belirtme işi"* ve *"İlkelerini olasılık kuramlarından alarak eldeki verileri grafik ve sayı biçiminde değerlendirmeye dayandıran matematiğin uygulamalı dalı; sayım bilimi* [8]*."* olarak tanımlamaktadır. TDK tarafından yapılan bu sözlük tanımı oldukça terimsel görünmektedir. İstatistiksel analizler ise en basit tanımıyla verinin bilgiye dönüşüm sürecidir. Bu tanım itibariyle istatistiksel analizler ile veri madenciliği arasındaki yüksek benzerlik göze çarpmaktadır. Her iki yaklaşım da verilerin anlamını çözmek ile ilgilenir. Her iki araç belirsizliklerin üstesinden gelmek ve gelecekteki olaylar hakkında bilgi vermek için bulunmuştur. Veri madenciliği ve istatistiğin her ikisi de bir olayı etkileyen önemli faktörleri belirlemek ve türetilen modeller ile gelecekteki olayları daha iyi öngörmek ile ilgilenmektedir [15]. Ancak önemle belirtmek gerekmektedir ki veri madenciliği uygulamalarında istatistiksel analizler bir model tekniği olarak kullanıldığından veri madenciliği süreçleri daha büyük ve kapsayıcı bir yapıya sahiptir. Bu bağlamda veri madencisi, veri bilimci ve veri analisti gibi meslekler istatistikçi mesleğinin kariyer uzmanlıkları olarak da değerlendirilebilir.

Verinin istatistiksel tekniklerle anlamlı bilgiye dönüştürülme sürecinde bağımlı ve bağımsız değişken sayısının türü istatistiksel tekniğin tek değişkenli mi yoksa çok değişkenli mi olduğunu ortaya koymaktadır. Örneğin bir bağımsız değişkenin bir bağımlı değişkeni açıklaması basit doğrusal regresyon analizi iken birden fazla bağımsız değişkenin açıklanması ise çoklu doğrusal regresyon analizidir.

Hayatın olağan akışında bir olay ve sonuç birden fazla faktör ve değişkenin etkisinde meydana geldiğinden çok değişkenli istatistiksel analizler daha fazla önem kazanmaktadır. Ancak kimi zaman karmaşık olmayan problemlerin çözümünde tek değişkenli analizlere de ihtiyaç duyulmaktadır. Hem tek değişkenli hem de çok değişkenli istatistiksel analizler betimsel ve çıkarımsal istatistiksel analizler olarak iki ana gruba ayrılmaktadır. Betimsel istatistiksel analizler veri madenciliğindeki veri keşfine, çıkarımsal istatistiksel analizler ise veri madenciliğindeki modelleme sürecine benzemektedir. Betimsel istatistiksel analizler ile veri görselleştirilir, merkezi eğilim ve yaylım ölçülerine göre tanımlanır ve veri hakkında önemli bilgiler elde edilir. Betimsel analizler ile ayrıca veri çıkarımsal analizlere uygun hale getirilmesi için ön hazırlık gerçekleştirilir. Çıkarımsal istatistiksel analizlerde ise bağımlı ve bağımsız değişkenler arasındaki ilişkiler ortaya konulur, veri setindeki gizli kalan bilgiler açığa çıkartılır ve geleceğe ilişkin tahminlerde bulunulur. Bununla birlikte çıkarımsal istatistiksel analizlerle gözlemler arası kümeleme ve ayırma işlemleri de yapılabilir.

Çok değişkenli istatistiksel analizler, analizin amacı ve değişken yapısına göre farklı türlere ayrılmaktadır. Sıklıkla kullanılan ve bilinen çok değişkenli istatistiksel analizler şöyle sıralanabilir.

- Kümeleme Analizi
- Faktör Analizi
- Path Analizi
- Çok Değişkenli Regresyon Analizi
- Çok Değişkenli Lojistik Regresyon Analizi
- Temel Bileşenler Analizi
- Diskriminant(Ayırma Analizi)
- Manova ve Mancova Analizi
- Uyum Analizi

İstatistiksel analizlerle birlikte az da olsa bazı kaynaklarda istatistiksel öğrenmeden[6] bahsedilmiştir. İstatistik; veri toplama, düzenleme, görselleştirme, analiz etme, çıkarsama, tahminleme ve betimleme gibi çok geniş kavramları kapsarken, istatistiksel

[6] İstatistiksel öğrenme teorisi, verilere dayalı bir tahmin fonksiyonu bulmanın istatistiksel çıkarım problemiyle ilgilenir. İstatistiksel öğrenme teorisi bilgisayarlı görme, konuşma tanıma ve biyoenformatik gibi alanlarda başarılı uygulamalara yol açmıştır [43].

öğrenme çok daha dar bir kavramı yani çıkarsama(inference) ve tahminlemeyi(estimation/prediction) içerir. Başka bir deyişle istatistiksel öğrenme, stokastik bir model kurma ve bu modeli geçerleme durumudur [44]. Bu tanımdan da anlaşılacağı üzere istatistiksel öğrenmeden anlaşılan çıkarımsal istatistiksel analizlerdir. Bu çıkarıma göre lojistik regresyon, doğrusal regresyon, kümeleme ve ayrım analizi gibi analizler istatistiksel öğrenme metotları olarak değerlendirilebilir. Bazı anlatımlarda karar ağacı, k en yakın komşu, desek vektör makinaları gibi makine öğrenmesi algoritmalarının istatistiksel öğrenme şeklinde yer verildiği tespit edilmiştir. Bu noktadan hareketle istatistiksel öğrenmenin makine öğrenmesi ile karıştırıldığı tespit edilmiştir. Esasında çok değişkenli istatistiksel analizlerde bir öğrenme söz konusu değildir. Öğrenmenin olduğu yöntemlerde veri seti eğitim ve test verisi şeklinde ayrılmakta iken istatistiksel yöntemlerde böyle bir ayrım söz konusu olmayıp veri olduğu gibi analize dâhil edilmektedir. Dolayısıyla istatistiksel öğrenme ifadesinin zorlama bir kavram olduğu tarafımızca değerlendirilmektedir. Çok değişkenli istatistiksel yöntemler ile makine öğrenmesi modelleri arasındaki yüksek ilişki bu teknikler arasındaki benzerlik ve farklılıkların daha net bir şekilde ortaya konulması gerekliliğini de birlikte getirmektedir.

Çok Değişkenli Lojistik Regresyon Analizi/Modellemesi

Regresyon analizi değişkenler arasındaki matematiksel ilişkiyi modellemek ve incelemek amacıyla kullanılan bir istatistiksel yöntemdir. Hastaların iyileşme süresi ile tedavide kullanılan ilaç dozu arasında matematiksel bir bağıntı var mıdır? Eğer bu iki değişken arasında matematiksel bir bağıntı varsa, bunun biçimi nasıldır? Regresyonda değişkenler arasındaki bağıntının biçimi model için model ifadesi kullanılır [45].

Regresyon analizinin amacı, bağımsız değişkenlerdeki(x) değişime sayalı olarak bağımlı değişkendeki(y) değişimin açıklanmasıdır. Amacı giriş ve çıkış değişkenleri arasındaki ilişkiyi yakalamak olan regresyon analizi için nedenselliği(sebep-sonuç ilişkisi) açıklayan bir algoritmadır [1]. Çoğu kez bir bağımlı değişken ile yapılan modelleme yeterli olmayabilir. Eğer bir bağımsız değişken le yapılan modelleme yeterli olmazsa modelleme işlemine başka bağımsız değişkenlerde katılmalıdır. Bu durumda yapılan işlem çoklu regresyon(multiple regresyon) olarak adlandırılır [45].

Bilindiği gibi basit ve çoklu doğrusal regresyonda bağımlı değişken sayısal veri türündedir. Buna karşılık, çalışmalarda bağımlı değişkenin kategorik/niteliksel veri türünde olması durumu ile sıklıkla karşılaşılır. Bu durumda, doğrusal regresyonda parametre kestirimlerini hesaplamak için kullanılan en küçük kareler yönteminden yararlanmak bu yöntemle ilgili varsayımlar sağlanmadığı için uygun olmamaktadır. Bu nedenle, bağımlı değişken iki ya da ikiden çok kategorili niteliksel veri türünde olduğunda lojistik regresyon yöntemi ile çözümleme gerçekleştirilmektedir. Değişkenlerin türü ve dağılım ile ilgili varsayımların az olması ve sonuçların kolaylıkla yorumlanabilmesi, vb. nedenlerle lojistik regresyon yöntemi son yıllarda sıklıkla kullanılan bir regresyon yöntemi konumuna gelmiştir [46].

Bağımlı değişken sayısının iki kategorili olması durumunda binary lojistik regresyon, iki kategoriden fazla olması durumunda çok kategorili(multinominal) lojistik regresyon ve bağımlı değişkenin sıralı olması durumunda ise ordinal lojistik regresyon modelleri bulunmaktadır.

Lojistik regresyon analizinde aşağıda ye alan işlemler sıralı şekilde gerçekleştirilir.

1. Doğrusallık, hata terimlerinin bağımsızlığı; çoklu doğrusal bağlantı ve beklenen frekans sayısı gibi model varsayımlarının sağlanması,
2. Bağımlı değişken ile açıklayıcı bağımsız değişkenler arasında ilişkinin test edilmesi(ki-kare analizi, korelasyon analizi ve tek değişkenli logit model),
3. Bağımlı değişken ile istatistiksel olarak yeterli ilişkiye sahip bağımsız değişkenlerin seçilmesi(p değeri 0,25'ten küçük bağımsız değişkenlerin seçilmesi),
4. Belirlenen bağımsız değişkenler arasında çoklu doğrusal bağlantı sorunu olan değişkenlerin belirlenerek ve uygun olmayanların çıkartılması(0,60 veya 0,70 üzerinden korelasyona sahip değişkenler),
5. Sadece sabit değerin ve bağımlı değişkenini yer aldığı basit başlangıç modeli oluşturulması(basit lojistik regresyon modeli),
6. Sabit değer ve ilişki analizlerine göre modele alınması uygun görülen bağımsız değişkenlere göre ana modellenin oluşturulması ve ilk oluşturulan basit model ile son oluşturulan ana model karşılaştırılarak iki model arasındaki farkın test edilmesi(olabilirlik oran testi),

7. Ana modelde yer almasına rağmen istatistiksel anlamlığa sahip olmayan bağımsız değişkenlerin modelden çıkartılması (enter, ileriye doğru ve geriye doğru çıkarma yöntemleri),
8. Bağımsız değişken çıkarma işlemi ile amaçlanan modele ulaşılıp ulaşılmadığı test edilmesi(olabilirlik oran testi),
9. Model parametrelerinin istatistikse anlamlılığının test edilmesi(wald testi ve skor testi),
10. Modelin veri setini ne kadar iyi temsil ettiğinin uyum iyiliği ile test edilmesi(hosmer-lemoshow testi),
11. Oluşturulan amaçlanan modele ilişkin model parametrelerinin yorumlanması(ODDS[7] değeri üzerinden),
12. Modelin yorumlanması(Açıklama katsayısı üzerinden(R^2).

Lojistik regresyon analizini istatistiksel yöntemlerle gerçekleştirilmesi durumunda yukarıda yer alan işlemlerin sıralı bir şekilde gerçekleştirilmesi gerekmektedir. Ancak literatürde lojistik regresyon modellemesi bir yapay zekâ makine öğrenmesi algoritması şeklinde de uygulandığında ve söz konusu istatistiksel işlemlerin gerçekleştirilmemektedir.

Lojistik Regresyon Avantajları [47],

- Lojistik regresyonun uygulanması, yorumlanması kolaydır.
- Veri seti doğrusal olarak ayrılabiliyorsa oldukça iyi performans gösterir.
- Overfitting daha az meyillidir ama büyük veri setlerinde overfit olabilir.

Lojistik Regresyon Dezavantajları [47],

- Gözlem sayısı özellik sayısından azsa, Lojistik Regresyon kullanılmamalıdır, aksi takdirde overfit olabilir.
- Lojistik regresyonun ayrım yapabilmesi için veri setinin doğrusal olarak ayrılabiliyor olması lazım) Bir takım varsayımların sağlanması gerekmektedir.

[7] ODDS değeri(Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı):$\frac{P(Y)}{1-P(Y)}$ şeklinde gösterilir.

2.1.17 Makine Öğrenmesi Algoritmaları İle Çok Değişkenli İstatistiksel Yöntemler Karşılaştırması

Veri madenciliği kapsamında hem makine öğrenmesi algoritmaları hem de çok değişkenli istatistiksel yöntemler modelleme tekniği olarak kullanılmaktadır. Daha öncede belirtildiği üzere bu modelleme teknikleri kimi zaman birbiriyle de karıştırılmakta veya birbiri yerine de kullanılmaktadır.

Makine öğrenmesi yapay zekânın bir uygulama alanı iken çok değişkenli istatistiksel yöntemler istatistiğin bir uygulama alanıdır. Makine öğrenmesi algoritmaları ağırlıkla matematiksel tabanlı olup model oluşturulması için herhangi bir varsayımın kabul edilmesi gibi bir ön şart taşımamaktadır. Çok değişkenli istatistiksel yöntemler ise istatistiksel teknikler ile oluşturulmuş olup çoğu zaman, normallik, doğrusallık ve yeterli örneklem gibi sağlanması gereken varsayımlara sahiptir. Makine öğrenmesi algoritmaları ile tahmin işlemi yapılırken çok değişkenli istatistiksel yöntemler ile hem tahmin hem de çıkarım yapılır. Makine öğrenmesi algoritmaları ile oluşturulan modeller ve model parametreleri bir teste tabi tutulmadan uygulamaya konulma imkânı varken çok değişkenli istatistiksel yöntemler ile oluşturulan modeller kısmi ve tümel F testi gibi testlere tabi tutulduğu gibi modelin veriye uygunluğu da uyum iyiliği testleri ile sağlanır. Makine öğrenmesin de çeşitli doğruluma ve geçerleme yöntemleri ile veri seti eğitim ve test olmak üzere parçalara ayrılır ve makine öğrenme işlemi gerçekleştirilir. Çok değişkenli istatistiksel yöntemlerde ise veri herhangi bir parçaya ayrılmadan bütün şekilde analize alınır ve öğrenme işlemi yerine veriye göre model uydurma işlemi gerçekleşir.

Makine öğrenmesi modellerinde yüksek performans hedeflenirken çok değişkenli istatistiksel yöntemlerde yüksek performanstan ziyade modelin açıklama oranın yüksek olması ve modelin veri setine olan uyum iyiliği hedeflenmektedir. Model performans değerlendirmesinde her iki yöntem için de hata matrisi ve ROC eğrisi kullanılabilir. Makine öğrenmesi modelleri büyük veriler için kullanımı daha elverişli olup veri ön işleme süreçlerine duyarlıdır. Çok değişkenli istatistiksel yöntemler ise daha küçük veri setlerinde çoğunlukla örneklem veri seti ile kullanılır ve daha elverişlidir ayrıca veri öne işleme sürecine çok duyarlıdır.

Tablo 2.4 Çok Değişkenli İstatistiksel Yöntemler İle Makine Öğrenmesi Algoritmaları Arasındaki Benzerlikler ve Farklılıklar

	Çok Değişkenli İstatistiksel Yöntemler	Makine Öğrenmesi Algoritmaları
Kaynak	Ağırlıkla olasılık ve istatistik temellidir. İstatistiğin alanından sayılır.	Ağırlıkla matematik temelli olup bazı algoritmalar olasılık ve istatistik temellidir. Yapay zekânın alanından sayılır.
Hedef	Uyum iyiliğine ve yüksek açıklama oranına sahip model amaçlanmaktadır.	Performans göstergeleri yüksek bir model amaçlanmaktadır.
Amaç	Değişkenler/Öznitelikler arasındaki ilişkinin açıklanması yapar ve tahminde bulunmaktadır Çıkarım (inference) yapar. (Parametre testi, Güven aralıkları tahmini, uyum iyiliği, Hata incelemesi vs. gibi) [44]	Bilinen bilgilerden öğrenme yoluyla bilinmeyen anlamlı verilerin çıkartılması ve tahmin edilmesi işlemi yapılır. Ancak elde edilen anlamlı bilgiyi belirleyen ve etkileyen faktörler ile ilgilenmez. Çıkarım yapılamaz. Performans metriklerine bakılarak modelin gücüne karar verilir. (Confusion matrix: Accuracy, recall, sensitivity, roc-auc gibi) [44]
Yöntem	İstatistiksel yöntemlere tabidir.	Optimizasyon yöntemleri ön plana çıkar.(gradient descent, newton raphsongibi) [44]
Varsayım	Uygulanması için bir takım istatistiksel varsayımların gerçekleşmesi gerekmektedir.	Uygulama için varsayımlar ve ön savlara gerek yoktur.
Uyum İyiliği	Hazırlanan model için uyum iyiliği testi uygulanır.	Oluşturulan model bir teste tabi tutulmaz.
Eğitim Öğrenme	Yöntemin eğitilmesine gerek yoktur. Veri setinin tamamı model oluşturma aşamasında kullanılır.	Veri seti eğitim ve test verisi olarak ikili bir ayrıma tabi tutulur ve öğrenme işlemi sağlanır.
Öznitelik Boyutlandırma	Çok Değişkenli İstatistiksel modeller aynı zamanda bir öz nitelik çıkarma ve seçme aracı olarak ta kullanılmaya çok elverişlidirler.	Makine öğrenmesi algoritmaları öz nitelik boyutlandırma için çok elverişli olmayıp, başkaca boyutlandırma yöntemlerine ihtiyaç duymaktadır. Ancak denetimsiz öğrenme yöntemi algoritmaları ile öznitelik boyutlandırılabilir.
Gelişim	Açıklama oranı sayesinde modelde yer almayan ancak hedef değişkeni açıklayabileceği öz niteliklerin araştırılmasına ve modelin geliştirilmesine ve doğruluk oranın arttırılmasına olanak sağlamaktadır.	Farklı tip algoritma kullanımı ve değişik tip model geçerleme yöntemleri ile modelin performans değerleri arttırılabilmektedir.
Veriye Duyarlılık1	Veri ön işlemeye çok duyarlıdır	Veri işlemeye duyarlıdır.
Veriye Duyarlılık2	Küçük veriler için hem uygun hem de yeterlidir.	Daha büyük ve çok büyük veriler için uygundur.

Makine öğrenmesi algoritmaları ve çok değişkenli istatistiksel yöntemlere ilişkin ortaya koyulan benzerlikler ve farklılıklar birlikte değerlendirildiğinde her iki modelleme tekniğinin avantaj ve dezavantajları şu şekilde açıklanabilir.

Makine öğrenmesi modellerinin avantajları;

- Modelleme için bir varsayıma tabi olmama,

- Büyük veri ile çalışmaya elverişli olma,
- Yüksek model performansı sergileme,
- Performans iyileştirme ve güncellemeye olanak tanıma,
- Veri ön işlemeye daha az duyarlılık,
- Model için uyum iyiliğine gerek duymama,
- Model parametreleri için teste tabi olmama,

Makine öğrenmesi modellerinin dezavantajları;

- Model açıklaması ve çıkarımda yetersizlik,
- Boyut indirgemede etiksizlik,
- Büyük veri temini, işlemesi ve modellemesi kaynaklı çeşitli maliyetlerdir.

Makine öğrenmesi modellerinin avantaj ve dezavantajları birlikte değerlendirildiğinde makine öğrenmesi modellerinin avantajları dezavantajlara üstün gelmektedir.

Çok değişkenli istatistiksel yöntemlerin avantajları;

- Çıkarımsal bilgilerin elde edilmesine olanak sağlar,
- Model ile tahmin gerçekleştirilebilir,
- Değişkenler arasındaki nedensel ve nedensel olmayan ilişkiyi tespit eder,
- Yükseğe yakın bir model performansı sergiler,
- Boyut indirgemede çok etkilidir,
- Örneklem ile veri seti indirgemeye olanak sağlar,
- Veri temini ve modelleme maliyeti daha düşüktür

Çok değişkenli istatistiksel yöntemlerin dezavantajları;

- Varsayımlara tabidir,
- Model ve model parametreleri teste tabidir,
- Büyük veride performans düşüklüğüne neden olur,
- Veri ön işlemeye daha duyarlıdır,

Çok değişkenli istatistiksel yöntemlerinde makine öğrenmesinde olduğu gibi avantajları dezavantajlara üstün gelmektedir. Her iki modelleme tekniğinden hangisinin seçileceği sorusu çok net olarak araştırmanın problemi ya da işin ve veri setinin durumu ile ilgilidir. Eğer yüksek bir performans isteniyor ise makine öğrenmesi modelleri tercih edilebilir. Ancak her iki yöntemde birlikte de kullanılabilir.

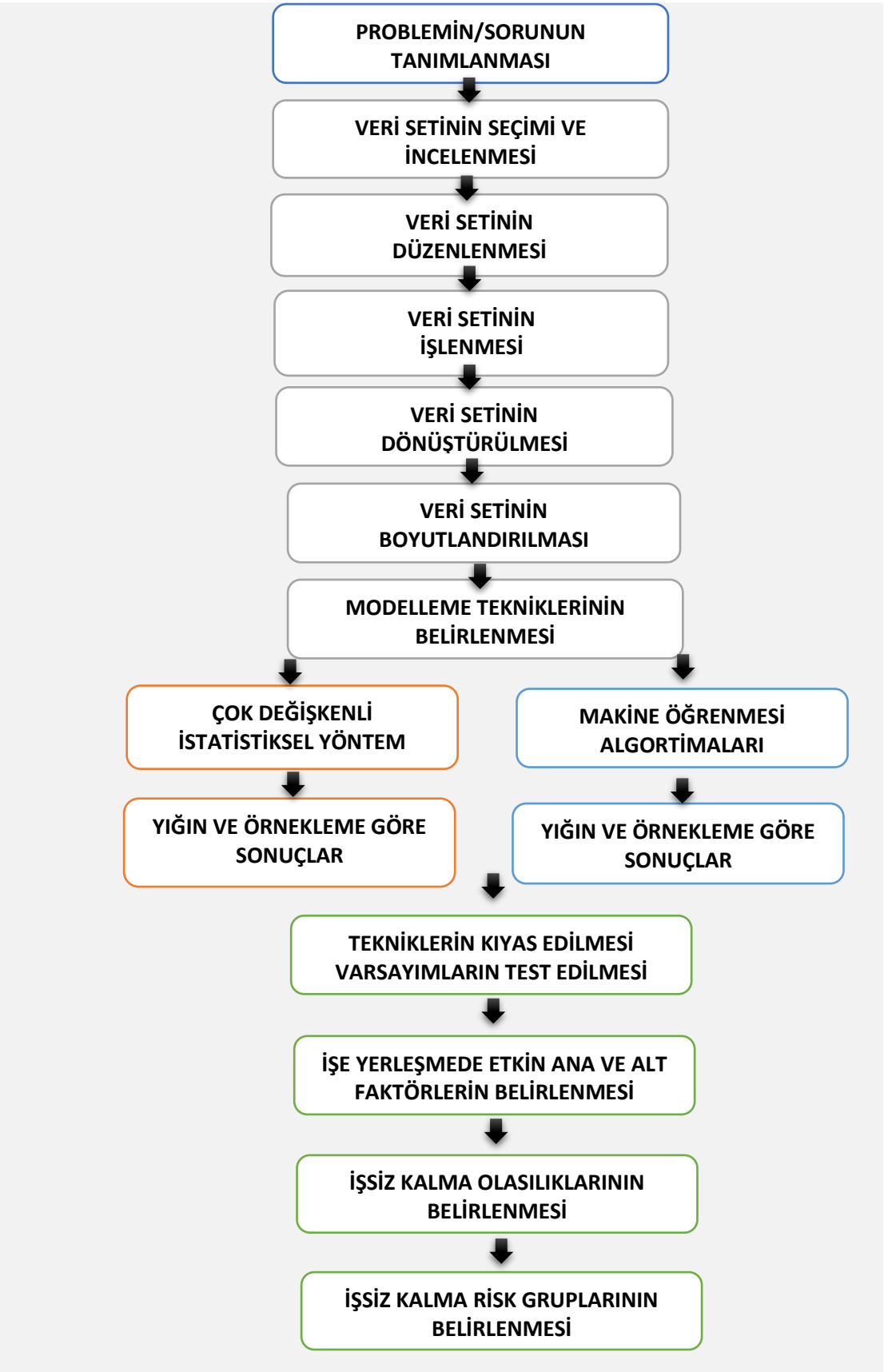
3. BULGULAR

Çalışma kapsamında öncelikle çalışma sistematiğinin oluşturulması ve şemalaştırılması işlemi yapılmıştır. Daha sonra çalışma problemin tanımlanması yapılmıştır. Sonrasında ise İŞKUR'a 2022 Eylül, Ekim ve Kasım aylarında iş için kayıt yaptıran iş arayanlara ait İŞKUR'dan resmi olarak elde edilen veri setinin veri madenciliği sürecine hazır hale getirilmesi ve tanımlayıcı istatistikler ile veri keşfi yapılması, veri madenciliği ile verilerin ön işlem tabi tutularak modellenmeye ve sınıflandırılmaya uygun hale getirilmesi, işsizlerin kayıt sonrası bir yıl içerisinde işe yerleşme/yerleşmeme durumuna göre modellenmesi, sınıflandırılması, işe yerleşmede etkin ana ve alt faktörlerin belirlenmesi, iş arayanın kişisel, demografik ve işgücüsel özelliklerine göre işe yerleşme olasılıklarının hesaplanması ve bu olasılık hesabına göre işsiz kalma risk gruplarının belirlenmesi, çok değişkenli istatistiksel sınıflama modeli ile makine öğrenmesi algoritmalarının birbirine üstünlüğü ve model geçerleme /doğrulama yöntemlerine göre doğruluk performansında meydana gelen değişimler, yığın ve örneklem verilerine göre makine öğrenmesi algoritmalarının doğruluk performansında meydana gelen değişimler ve çalışma sorusunun test edilmesi işlemleri gerçekleştirilmiştir. Yapılan işlemlere dair bulgulara ve bulgulara dayalı yorumlara yer verilmiştir.

Çalışmaya ilişkin istatistiksel testler ve analizler GRETL, TURCOSA, JASP TANAGRA gibi istatistiksel analizlere imkân veren programlarda hazırlanmıştır. Yapay zekâ makine öğrenmesine ilişkin modellemeler ise Google Colab platformu kullanılarak PYTHON programı ile yapılmıştır. Bununla birlikte veri seti incelemesi, düzenlenmesi ve keşfi ise EXCEL programı ile gerçekleştirilmiştir.

3.1 Çalışma Sistematiğinin Oluşturulması ve Şemalaştırılması

Tez çalışması esasında altı ana aşamadan oluşmaktadır. Bu aşamalar şu şekilde sıralanmaktadır. Veri madenciliği ile veri setinin modellenmeye uygun hale getirilmesi, çok değişkenli istatistiksel yöntem ile modellenmenin ve sınıflandırılmanın sağlanması, makine öğrenmesi algoritması ile sınıflandırmaların sağlanması, teknik, yöntem ve algoritmaların kıyas edilmesi ve varsayımların test edilmesi, işe yerleşmede etkin ana ve alt faktörlerin tespiti ve işsiz kalma olasılığı ve bu olasılığa göre işsiz risk gruplarının belirlenmesi.



Şekil 3.1 Çalışma Sistematiğine İlişkin Akış Şeması

3.2 Problemin/Sorunun Tanımlanması

Tez çalışması iki ana problemden/sorudan oluşmaktadır.

Birincisi; "İşsizlerin zamana göre işsiz kalma risklerinin tespitinde yapay zekâ makine öğrenmesi yöntemleri klasik istatistiksel yöntemlere göre daha mı etkili?"

İkincisi; "İşsizlerin işe yerleşmesinde etkin faktörlerin/değişkenlerin neler olduğu?"

Esasında çalışmada yer alan her iki problem birbirini tamamlayıcı ve geliştirici niteliktedir. Özellikle daha sonra yapılacak çalışmalarda daha iyi modellemelerin ve sınıflamaların yapılması için daha yüksek performans tekniğin belirlenmesi ve bu tekniğin daha yüksek açıklama ve doğruluk oranına sahip olması içinde işe yerleşmede etkin değişkenlerin belirlenmesi gerekmektedir. Dolayısıyla çalışmada yer alan her iki problem ve probleme ilişkin sonuçları önem arz etmektedir. Ayrıca tez çalışmasının ana soruları altında makine öğrenmesi algoritmalarının ve model geçerleme yöntemlerinin, yığın ve örneklem verilerinin de birbirine üstünlüğü alt problemler olarak irdelenmiştir. Böylece performans değerlendirilmesi seçilen teknik, algoritma, eğitim yöntemi ve veri setine göre de ayrı ayrı yapılmıştır.

3.3 Veri Setinin Seçimi, İncelenmesi ve Düzenlenmesi

Çalışma kapsamında İŞKUR tarafından verilen ham veriler üç parçadan oluşmaktadır. Veri setinde toplam 13.459 satırdan yani iş arayana ait bilgiden oluşmaktadır. Yine veri setinde iş arayanın kişisel, demografik ve işgücüsel bilgilerine ait 19 sütun bulunmaktadır. Verilerin birinci parçasında 2022 Eylül, Ekim ve Kasım döneminde iş arayanların cinsiyet, yaş, doğum yılı, doğum ayı, öğrenim, medeni durum gibi bilgiler yer almaktadır. Verilerin ikinci parçasında iş arayanların iş arama kaydı sonrası işe başlama tarihleri, üçüncü parçasında ise iş arayanların kayıt yaptırdıkları son meslek bilgisi yer almaktadır. Ham veri seti bu haliyle bir analiz ve modelleme yapmaya elverişli değildir. Dolayısıyla veri setindeki bu parçalı yapı kişiPK numarası[8] kullanılarak Excel ortamında ortadan kaldırılmış ve üç veri seti parçası tek parça şeklinde birleştirilmiştir. Birleştirilen veri seti incelendiğinde iş arayan mesleğinin ana meslek gruplarına göre ayrılmadığı, iş arayanın doğum yılına yer verildiği ancak yaş

[8] **KişiPK numarası**; Kayıtlı iş arayan özgü bir numara olup belirli bir kişiyi ifade etmektedir.

bilgisinin olmadığı, Sygm[9] ve İÖ[10] başlama tarihlerinde "NULL[11]" ifadesinin yer aldığı gözlemlenmiştir. Bu noktada Türk Meslekler Sözlüğü kullanılarak iş arayan meslekleri meslek gruplarına ayrılmıştır. Ayrıca Excel'de iş arayanların doğum yılı üzerinden yaşları hesaplanmış ve yaşlar en uygun şekilde kategorilere ayrılarak dönüştürülmüştür. Bununla birlikte "NULL" ifadesi çıkartılarak sütün içeriğine uygun ifadeler yazılmıştır. Ayrıca Sivas ilinde ikamet etmeyen iş arayan veri setinden çıkartılmıştır böylece veri setindeki iş arayan sayısı 13.457' düşmüştür. Tüm bu inceleme ve düzenleme işlemleri ile veriler, veri keşfine ve ön işlemeye hazır hale getirilmiştir.

Tablo 3.1 Birleştirilmiş Örnek Ham Veri Seti(20 İş Arayana Ait)[12]

	İŞE GİRME DURUMU	İŞE GİRİŞ TARİHİ	CINSIYET	BASVURU_TARIH	MESLEK	OGRENIM_DURUM	BASVURU_TUR	SOSYAL_DURUM	Sygm Baslama Tarihi	IO_Baslama Tarihi
1	1	20221126	Erkek	20221124	BÜRO MEMURU (GENEL)	Ortaöğretim (Lise ve	Çalışırken işsiz kalan	Engelli	NULL	NULL
2	0		Erkek	20221128	TEMİZLİK GÖREVLİSİ	Ortaöğretim (Lise ve	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
3	0		Erkek	20221118	KASAP	Ortaöğretim (Lise ve	Çalışırken işsiz kalan	Normal	NULL	NULL
4	0		Erkek	20221121	MOTOR İMALAT BAKIM VE ONAR	Lisans	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
5	1	20221201	Erkek	20221024	MEKANİK BAKIM ONARIMCISI	Ortaöğretim (Lise ve	İlk kez iş hayatına atılan	Normal	NULL	20221015
6	0		Kadın	20221014	MAKİNECİ (DİKİŞ)	İlköğretim	Çalışırken işsiz kalan	Normal	NULL	NULL
7	0		Erkek	20221129	TEMİZLİK GÖREVLİSİ	Ortaöğretim (Lise ve	Çalışırken işsiz kalan	Normal	NULL	NULL
8	0		Erkek	20221017	FIRINCI USTASI-UNLU MAMULLER	İlköğretim	Çalışırken işsiz kalan	Engelli	NULL	NULL
9	0		Erkek	20221004	BEDEN İŞÇİSİ (İNŞAAT)	İlkokul	Çalışırken işsiz kalan	Normal	NULL	NULL
10	0		Erkek	20221130	TEMİZLİK GÖREVLİSİ	İlköğretim	Daha iyi şartlarda iş arayan	Normal	20171124	NULL
11	0		Erkek	20221124	BÜRO MEMURU (GENEL)	Ortaöğretim (Lise ve	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
12	0		Erkek	20221017	BEDEN İŞÇİSİ (GENEL)	İlköğretim	Daha iyi şartlarda iş arayan	Normal	NULL	20221017
13	0		Erkek	20221020	ORTACI/AYAKÇI (TEKSTİL)	İlkokul	İlk kez iş hayatına atılan	Normal	NULL	20221014
14	0		Erkek	20221024	AŞÇI (YÖRESEL MUTFAK)	İlköğretim	İlk kez iş hayatına atılan	Normal	NULL	20220925
15	0		Erkek	20221125	TEMİZLİK GÖREVLİSİ	Ortaöğretim (Lise ve	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
16	0		Kadın	20221017	SATIŞ DANIŞMANI / UZMANI	Lisans	Çalışırken işsiz kalan	Normal	NULL	NULL
17	1	20230101	Erkek	20221128	ODA GÖREVLİSİ-TURİZM	İlköğretim	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
18	0		Erkek	20221028	DİĞER BÜRO MEMURLARI	Ortaöğretim (Lise ve	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
19	0		Erkek	20221005	OVERLOK MAKİNESİ OPERATÖRÜ	Önlisans	Daha iyi şartlarda iş arayan	Normal	NULL	NULL
20	0		Erkek	20221103	TEMİZLİK GÖREVLİSİ	Ortaöğretim (Lise ve	Çalışırken işsiz kalan	Engelli	NULL	NULL

3.4 Veri Keşfi ve Tanımlayıcı İstatistikler

Birleştirilmiş ham veri setine göre iş arayanların %99,9 Türkiye Cumhuriyeti uyruklu, %78,4'ü Sivas merkezde kalanı ilçelerde, %97,5'i normal statüde kalanı engelli ve eski hükümlü statüsünde, %55'i erkek ve %45'i kadın, %80,4'ü 15-39 yaş aralığında kalanı 40 ve üzeri yaş aralığın da, %56,3'ü bekâr, %39,3 evli ve kalan diğer medeni hale sahip, %99,9'u sosyal yardım almıyor, %18,2 ilköğretim ve altı, %39,1 orta öğretim(lise), %13,4'ü ön lisans ve %15,9'u lisans ve üzeri örenime sahip, %46,7'si nitelik gerektirmeyen mesleklere sahip iken kalanı nitelikli mesleklere sahiptir, %42,5'u bir işi olan ve daha iyi şartlarda iş ararken %57,'i işsiz olarak iş aramaktadır,

[9] **Sygm başlama tarihi;** İş Arayanın sosyal yardım almaya başlama tarihini ifade etmektedir.
[10] **İÖ başlama tarihi;** İş Arayanın işsizlik ödeneği almaya başlama tarihini ifade etmektedir.
[11] **NULL ifadesi;** Boş/geçersiz anlamına gelip iş arayanın sosyal yardım ve işsizlik ödeneği almadığını ifade etmektedir.
[12] İş arayana ait doğum ay, gün, yılı ve doğum yeri, ikamet edilen ilçe bilgilerine yer verilmemiştir.

%94,6'sı işsizlik ödeneği almıyorken kalanı almakta ve iş arayanların %89,6'sı İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşmemişken kalanı işe yerleşmiştir.

Veri setindeki iş arayanların demografik, kişisel ve iş aramaya ilişkin özellikleri esasında Sivas ili işgücü piyasasında iş arayanların bir mikyasta özelliklerini de yansıtmaktadır. Bu yansımaya göre Sivas ili işgücü piyasası ağırlıkla niteliksiz veya daha az nitelikli meslek ve öğrenime sahip, normal statüde daha iyi şartlarda iş arayanların ve işsizlerin iş aradığı bir yapı görünümündedir. Buna ek olarak iş arayanlar arasında cinsiyet, medeni durum ve yaş gibi değişkenler arasında il nüfusu kaynaklı bir dengeliliğin olduğu değerlendirilmektedir. Ayrıca veri seti il işgücü piyasasının istihdam edilebilirliği hakkında da önemli bilgiler vermektedir. Başka bir ifadeyle başvuru sonrası ilde istihdam edilebilirlik oldukça düşük olup istihdam edilmeme sorununa işaret etmektedir. Malum istihdam edilmeme sorunu da işsizliğe ve işsiz kalma sürelerinin uzunluğuna işaret etmektedir. Tüm bu tanımlayıcı istatistikler birlikte değerlendirildiğinde Sivas ilinde işgücü piyasası gelişimine ilişkin bir sorundan da bahsedilebilir ve bu sorunun çözümünde bu çalışma önemli bir yer teşkil edebilir.

Veri seti incelendiğinde az sayıda Türkiye Cumhuriyet uyruğuna sahip iş arayanların olduğu yine iş arayanların ağrılıkla merkezde ikamet etmesine karşın Sivas ilinin geniş bir coğrafyaya yayılması nedeni ile ilçede ikamet eden iş arayanların homojen bir yapıda olmadığı, sosyal durumu eski hükümlü olan, medeni durumu dul ve evliliğin iptali olan, öğrenimi okuryazar olmayan, okuryazar, yüksek lisans ve doktora olan, meslek grubu nitelikli tarım ve ormancılık, yönetici ve diğer meslek mensupları olan ve sosyal yardım alan iş arayan sayısının çok az olduğu tespit edilmiştir. Bununla birlikte öğrenci olmasına karşın iş arayanlarda veri setinde yer almaktadır. Veri setinde bazı kategorilerde iş arayan sayısının az olması özellikle çok değişkenli istatistiksel modeli doğrudan etkileyerek yanlı sonuçlara sebebiyet verme ihtimali bulunmaktadır. Ayrıca bilindiği üzere kategori sayısının artması modeli açıklama gücünü büyüterek modelin gerçeklikle olan ilişkisini azaltmaktadır.

Modelin gerçeklikle olan ilişkisini azaltan ve modelin yüksek doğru sınıflama yapmasına sebebiyet veren nedenlerden bir tanesi de verinin dengeli dağılmamasıdır. Tanımlayıcı istatistiklere göre iş arayanların ancak %10,4'ü İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşmiştir. Bu oran oldukça az olup veri seti bağımlı/hedef

değişken yönünden dengeli dağılmamıştır. Başka bir ifadeyle dengesiz bu veri seti işe yerleşmemiş olanları yüksek doğruluk oranı ile sınıflayacak ancak işe yerleşenler için aynı sınıflamayı başaramayacaktır. Dolayısıyla veri setinde hedef öznitelik/değişken yönünden veri dengelemeye ihtiyaç bulunmaktadır. Aksi takdirde iyi bir modelleme ve sınıflama gerçekleştirilmeyecektir. Tüm bu bilgiler birlikte değerlendirildiğinde veri setinin bu haliyle analiz edilmesi, modellenmesi ve sınıflandırılması ancak veri madenciliği kapsamında iyi bir ön işleme süreci ile mümkündür.

Tablo 3.2 Öznitelikler ve Kategorilere Göre Tanımlayıcı İstatistikler

Öz Nitelikler	Kategoriler	İş Arayan Sayısı	Yüzde
Uyruk	Afganistan İslam Cumhuriyeti	2	0,0%
	Almanya Federal Cumhuriyeti	2	0,0%
	İran İslam Cumhuriyeti	1	0,0%
	Suriye Arap Cumhuriyeti	3	0,0%
	Türkiye Cumhuriyeti	13448	99,9%
	Uganda Cumhuriyeti	1	0,0%
İkamet İl	Sivas	13457	100,0%
İkamet İlçe	Akıncılar	40	0,3%
	Altınyayla/Sivas	166	1,2%
	Divriği	135	1,0%
	Doğanşar	17	0,1%
	Gemerek	203	1,5%
	Gölova	9	0,1%
	Gürün	253	1,9%
	Hafik	78	0,6%
	İmranlı	67	0,5%
	Kangal	359	2,7%
	Koyulhisar	71	0,5%
	Sivas Merkez	10553	78,4%
	Suşehri	330	2,5%
	Şarkışla	420	3,1%
	Ulaş	148	1,1%
	Yıldızeli	387	2,9%
	Zara	221	1,6%
Sosyal Durum	Engelli	278	2,1%
	Eski Hükümlü	53	0,4%
	Normal	13126	97,5%
Cinsiyet	Erkek	7406	55,0%
	Kadın	6051	45,0%
Yaş	15-19	1319	9,8%
	20-24	3486	25,9%
	25-29	2621	19,5%
	30-34	1826	13,6%
	35-39	1565	11,6%

	40-44	1324	9,8%
	45+	1316	9,8%
Medeni Durum	Bekâr	7574	56,3%
	Boşanmış	527	3,9%
	Dul	66	0,5%
	Evli	5289	39,3%
	Evliliğin İptali	1	0,0%
Sosyal Yardım Alma Durumu	Sosyal Yardım Almıyor	13437	99,9%
	Sosyal Yardım Alıyor	20	0,1%
Öğrenim	Doktora	2	0,0%
	İlkokul	208	1,5%
	İlköğretim	3854	28,6%
	Lisans	2054	15,3%
	Okur Yazar	141	1,0%
	Okur Yazar Olmayan	45	0,3%
	Ortaöğretim (Lise Ve Dengi)	5256	39,1%
	Ön lisans	1807	13,4%
	Yüksek Lisans	90	0,7%
Meslek	Büro Hizmetlerinde Çalışan Elemanlar	1365	10,1%
	Öğrenci	115	0,9%
	Hizmet Ve Satış Elemanları	1710	12,7%
	Nitelik Gerektirmeyen İşlerde Çal.	6284	46,7%
	Nitelikli Tarım, Ormancılık Ve Su Ürünleri Çalışanları	68	0,5%
	Profesyonel Meslek Mensupları	838	6,2%
	Sanatkârlar Ve İlgili İşlerde Çal.	1329	9,9%
	Teknisyenler, Teknikerler Ve Yardımcı Profesyonel Meslek Mensupları	838	6,2%
	Tesis Ve Makine Operatörleri Ve Montajcılar	825	6,1%
	Yönetici Ve Diğer Meslek Mensupları	85	0,6%
Başvuru Türü	Çalışırken İşsiz Kalan	3378	25,1%
	Daha İyi Şartlarda İş Arayan	5717	42,5%
	Emekli Olup İş Arayan	3	0,0%
	İlk Kez İş Hayatına Atılan	4359	32,4%
İşsizlik Ödeneği Alma Durumu	İşsizlik Ödeneği Almıyor	12730	94,6%
	İşsizlik Ödeneği Alıyor	727	5,4%
İşe Yerleşme Durumu	İşe Yerleşmemiş	12054	89,6%
	İşe Yerleşmiş	1403	10,4%

3.5 Veri Madenciliği Ön İşleme Tekniklerinin Uygulanması

Verinin anlamlı bilgiye dönüşümü süreci ancak değerli verilerin değersiz ve gereksiz verilerden ayıklanmasına olanak sağlayan iyi bir veri ön işleme ve hazırlama işlemi ile mümkündür. Aksi takdirde veri öncelikle iyi bir modele akabinde ise gerçek bir bilgiye dönüşmeyecektir. Bu noktada veri madenciliği ön işleme teknikleri önem arz etmektedir. Veri madenciliği ön işleme ile ham veriler eksik, aykırı, yanlış, gürültülü ve gereksiz bilgilerden ayıklanarak model hazırlamaya ve sınıflandırılmaya hazır hale getirilmektedir. Veri ön işlemeye tabi tutulmayan verilerden gerçekleştirilen modeller yanlı, tutarsız ve gerçek dışı bilgilere dönüşmesi kuvvetle muhtemeldir. Ayrıca veri ön işleme teknikleri ile verinin yığın veya örneklem şeklinde analize dâhil edilip edilmemesi kararı da veriler. Buna ek olarak modelde ve sınıflandırma da yer alacak öz niteliklerin seçimi ve çıkarımı işlemi de bir veri ön işleme tekniği olup özellikle büyük veri yığınlarındaki maliyeti en aza indirmek için uygulanması kaçınılmaz bir tekniktir. Tüm bu bilgilere göre verilere veri madenciliği ön işleme teknikleri uygulanmış ve gerekli iyileştirmeler aşağıda yer aldığı üzere sıralı bir şekilde yapılmıştır.

Veri ön işleme işlemi kapsamında veri setinde aşağıdaki işlemler gerçekleştirilmiştir.

1. Türkiye Cumhuriyeti Uyruğu dışında %0,1 oranında başka ülke uyruğuna sahip iş arayanlar veri setinden tamamen çıkartılmış ve veri setinde sadece Türkiye Cumhuriyeti Uyruğuna sahip iş arayanlara yer verilmiştir.
2. Sivas ili dışında ikamet eden iki iş arayan veri setinden çıkartılmış ve sadece Sivas ilinde ikamet eden iş arayanlara yer verilmiştir.
3. İlçede ikamet eden iş arayan sayılarının homojen dağılmaması, ilçe sayısının fazla olması ve bazı ilçeler için iş arayan sayısının yetersiz olması nedeniyle ikamet değişkeni “Merkez” ve “İlçe” şeklinde ikili kategorize edilmiştir.
4. Sosyal durumu eski hükümlü olanlar bu statüde iş arayan sayısının azlığı nedeniyle veri setinden çıkartılmış ve sosyal durumu değişkeni “Normal” ve “Engelli” olmak üzere iki kategoriye ayrılmıştır.
5. Medeni durumu evliliğin iptali şeklinde bir iş arayan olduğu için bu iş arayan veri setinden çıkartılmış ve kalan medeni durum türleri “Bekâr”, “Boşanmış”, “Dul” ve “Evli” olarak dört kategoriye ayrılmıştır.

6. Okuryazar, okuryazar olmayan ve ilkokul öğrenimine sahip iş arayanların sayısı yetersiz olduğundan bu öğrenime sahip tüm iş arayanlar "İlkokul ve Altı" şeklinde birleştirilmiştir. Bununla birlikte yüksek lisans ve doktora türünden öğrenime sahip iş arayanların sayısı az olduğundan bu adaylar lisans öğrenim türü altında toplanarak "Lisans ve Üzeri" şekline birleştirilmiştir. Tüm bu işlemlerden sonra öğrenim değişkeni, "İlkokul ve Altı", "İlköğretim", "Orta Öğretim(Lise)", "Ön Lisans" ve "Lisans ve Üzeri" şeklinde kategorize edilmiştir
7. Meslek grubu öğrenci olanlar veri setinden tamamen çıkartılmıştır. Yönetici ve diğer meslek grubundaki iş arayan sayısı az olduğundan bu meslek grubundaki iş arayanlar profesyonel meslek mensupları grubundaki iş arayanlar birleştirilmiştir. Nitelikli tarım, ormancılık ve su ürünleri çalışanları meslek grubunda iş arayan sayısının az olması ve İŞKUR'un iş arayanlara yönelik hizmet sunumunun tamamının hizmet ve sanayi ana sektörlerinden olması nedeniyle veri setinden bu meslek grubundaki iş arayanlar çıkartılmıştır. Tüm bu işlemlerden sonra veri seti TMS'ye uygun olarak "Büro Hizmetlerinde Çalışan Elemanlar", "Hizmet Ve Satış Elemanları", "Nitelik Gerektirmeyen İşlerde Çalışanlar", "Profesyonel Meslek Mensupları", "Sanatkârlar Ve İlgili İşlerde Çalışanlar", "Teknisyenler, Teknikerler Ve Yardımcı Profesyonel Meslek Mensupları" ve "Tesis Ve Makine Operatörleri Ve Montajcılar" olmak üzere yedi kategoriye ayrılmıştır.
8. Çalışmanın konusu işsizlerin zamana göre işsiz kalma risklerinin tespit edilmesi olduğu için daha iyi şartlarda iş arayan ve emekli olup iş arayan başvuru türüne sahip iş arayanlar veri setinden çıkartılmıştır. Başvuru türü değişkeni/öz niteliği "Çalışırken İşsiz Kalan" ve "ilk kez İş Hayatına Atılan" olmak üzere iki kategoriye ayrılmıştır. Böylece veri setindeki ayrık/uç değerlere ilişkin veri ön işleme gerçekleştirilmiştir.
9. Veri setinde eksik bir veri olmadığı için veri tamamlama işlemi yapılmamıştır.

Gerçekleştirilen veri düzenlemesi ve veri ön işleme adımları ile hem öznitelik içi hem de öznitelikler arası bir uygunluk ve ahenk sağlanarak ham veri seti modellenebilir bir yapıya erişmesi sağlanmıştır. Ancak hâlihazırda ön işlemeye tabi tutulan verilerin bir takım dönüşüm, çıkarım ve seçim aşamalarında da geçmesi gerekmektedir.

3.6 Veri Madenciliği Dönüştürme Tekniklerinin Uygulanması

Ham verinin anlamlı, değerli ve gerçekçi bilgiye dönüşümü sürecinde veri hazırlama ve ön işleme aşamasından sonra verinin dönüşümü gelmektedir. Bir veri setinde öznitelikler birden farklı yapıda olabilirler. Bu olabirlik özniteliklerin ölçümü ve ölçüm birimi ile alakalı bir durumdur. Dolayısıyla özniteliğin ölçüm birimindeki farklılık hem analiz hem modelleme hem de sınıflama performanslarını olumsuz etkileyen bir yapı arz edebilir. Böyle bir olumsuzluktan kaçınmak için tüm özniteliklerin aynı ölçüm birimine getirilmesi gerekmektedir. Başka bir ifadeyle öznitelik birimlerinin standartlaştırılması gerekmektedir. Ölçüm birimi standartlaştırılması üzerine birden fazla farklı metot bulunmaktadır. Bununla birlikte bir veri setinde öznitelikler sürekli sayısal veriler olabildiği gibi kategorik verilerde olabilir. Bazı durumlarda sürekli verilerin kategorik hale getirilmesi daha iyi performans sonuçları verebilmektedir. Aynı zamanda doğrusal olmayan ve normal dağılmayan verilerin özellikle istatistiksel analiz varsayımlarının sağlanması için çeşitli teknikler ile doğrusallığı ve normalliği sağlanabilir.

Veri dönüşümünde bir diğer önemli konu ise veri dengelemedir. Bazı durumlarda veri setinde özellikle hedef öznitelikte dengesizlikler olabilir. Örneğin olağan durumlarda bir web sitesine normal giriş çok yüksek iken siber saldırı amaçlı giriş oldukça azdır. Yine gelişmemiş veya az gelişmiş bir işgücü piyasasında bir yıldan az sürede yani kısa ve orta vadede işe yerleşmeme sayısı çok yüksek iken işe yerleşme sayısı ise azdır. Her iki örnekten elden edilen veri seti dengesiz bir yapı arz ettiğinden bu veri setine göre gerçekleştirilen hem istatistiksel analiz, hem modelleme hem sınıflama yanlı, tutarsız ve gerçek dışı sonuçlar vererek araştırmacıyı yanıltacaktır. Böyle bir durumda veriler arasında dengenin sağlanması gerekmektedir. Veri dengelemeye ilişkin farklıca yöntemler bulunmaktadır.

Veri dönüşümünde bir diğer husus ise hedef değişken ile ilişkisiz özniteliklerin çıkarılması, birbiriyle yüksek bir ilişki ile çoklu bağlantılı özniteliklerin seçimi ve öznitelik sayısının çok olduğu durumlarda özniteliklerin grup içi yüksek gruplar arası ise düşük ilişkiye sahip faktörlere dönüştürme işlemi yapılmasıdır. Bu işlemelere öznitelik çıkarımı ve seçimi ve daha genel ifadeyle boyutlandırma/boyut küçültme denilmektedir. Boyutlandırma işlemi için uygulanan birçok teknik bulunmaktadır.

Veri dönüşümü kapsamında veri setinde aşağıdaki işlemler gerçekleştirilmiştir.

1. Veri setinde doğrudan bir yaş bilgisi olmadığı için öncelikle iş arayanların doğum tarihinden yaş bilgisi hesaplanmış ve sürekli bir veriye dönüştürülmüştür. Ancak iş arayanların yaşı dışındaki tüm veriler kategorik olması nedeniyle yaş bilgisi kategorik bir yapıya dönüştürülmüştür.
2. Veri setindeki tüm özniteliklerin kategorik olması/dönüştürülmesi nedeniyle tüm öznitelikler aynı ölçü birimine sahip olduğundan öznitelikler arası bir standartlaştırma işlemi gerçekleşmemiştir.
3. Tez çalışması kapsamında gerçekleştirilen istatistiksel analizlerde uyum iyiliği yüksek bir model oluşturmadan ziyade doğruluk oranı yüksek bir sınıflama amaçlandığı için doğrusallık ve normallik varsayımını gerçekleştirilmesine ilişkin bir veri dönüşümü yapılmamıştır.
4. Veri setinde yer alan iş arayanların İŞKUR'a başvuru sonrası bir yıl içerinde işe yerleşmeyenlerin sayısı 12.054 ve işe yerleşenlerin sayısı ise 1.403'tür. Anlaşıldığı üzere veri seti hedef değişken üzerinden dengelilik arz etmemektedir. Veri setindeki bu dengesizliği gidermek için öncelikle veriler özniteliklere göre sistematik şekilde sırlanmıştır. Sonrasında oversampling(az sayıda örnek çoğaltarak dengeleme) dengeleme yöntemi ile işe yerleşenlerin sayısı işe yerleşmeyenlerin sayısına yakın değerde işe yerleşen iş arayanlara ait satır özelliklerine göre arttırılmıştır. Böylece veri seti dengeli hale getirilerek işe yerleşenlerin sayısı 7.050 ve işe yerleşemeyenlerin sayısı 7.096 seviyesine ve toplam satır sayısı/iş arayan sayısı 14.146 seviyesine getirilmiştir.
5. Veri setine boyutlandırma işlemi, öznitelik seçme tekniği kullanılarak filter ve sarmal metot kullanılarak gerçekleştirilmiştir. Filter metotta istatistiksel korelasyon ve ki-kare analiz kullanılmış, sarmal metotta lojistik regresyona analizi geri arama tekniği kullanılmıştır. Yapılan işlemlere ilişkin detaylı açıklama çalışmanın ileriki başlıklarında yer almaktadır. Boyutlandırma yöntemi ile bazı öznitelikler veri madenciliği sürecinden çıkartılarak modele dâhil edilmemiştir

Gerçekleştirilen veri dönüşüm işlemleri ile ön işlemeye tabi tutulan veri seti tutarlı, gerçekçi ve yüksek performans sağlayacak bir yapıya dönüştürülmüştür. Veri setinin analize ve modellemeye hazır hale getirilmiştir.

Tablo 3.3 Veri Hazırlama, Ön İşleme ve Dönüşüm Sonrası Tanımlayıcı İstatistikler

Öz Nitelikler	Kategoriler	İş Arayan Sayısı	Yüzde
Uyruk	Türkiye Cumhuriyeti	14146	100,0%
İkamet İl	Sivas	14146	100,0%
İkamet İlçe	İlçe	3019	21,3%
	Merkez	11127	78,7%
Sosyal Durum	Engelli	425	3,0%
	Normal	13721	97,0%
Cinsiyet	Erkek	6866	48,5%
	Kadın	7280	51,5%
Yaş	15-19	2299	16,3%
	20-24	4863	34,4%
	25-29	2366	16,7%
	30-34	1362	9,6%
	35-39	1019	7,2%
	40-44	1189	8,4%
	45+	1048	7,4%
Medeni Durum	Bekâr	9176	64,9%
	Boşanmış	579	4,1%
	Dul	65	0,5%
	Evli	4326	30,6%
Sosyal Yardım Alma Durumu	Sosyal Yardım Almıyor	14102	99,7%
	Sosyal Yardım Alıyor	44	0,3%
Öğrenim	İlkokul ve Altı	412	2,9%
	İlköğretim	3785	26,8%
	Ortaöğretim (Lise ve Dengi)	5685	40,2%
	Ön lisans	1967	13,9%
	Lisans ve Üzeri	2297	16,2%
Meslek	Büro Hizmetlerinde Çalışan Elem.	1620	11,5%
	Hizmet Ve Satış Elemanları	1668	11,8%
	Nitelik Gerektirmeyen İşlerde Çal.	6435	45,5%
	Profesyonel Meslek Mensupları	898	6,3%
	Sanatkârlar Ve İlgili İşlerde Çal.	1427	10,1%
	Teknisyenler, Teknikerler Ve Yar. Prof. Mes. Men.	1149	8,1%
	Tesis Ve Makine Oper. Ve Mon.	949	6,7%
Başvuru Türü	Çalışırken İşsiz Kalan	6611	46,7%
	İlk Kez İş Hayatına Atılan	7535	53,3%
İşsizlik Ödeneği Alma Durumu	İşsizlik Ödeneği Almıyor	13439	95,0%
	İşsizlik Ödeneği Alıyor	707	5,0%
İşe Yerleşme Durumu	İşe Yerleşmemiş	7096	50,2%
	İşe Yerleşmiş	7050	49,8%

İşe yerleşmenin gerçekleşmesi iş arayan, açık iş ve işgücü piyasası ana faktörleri ve ana faktörler altında yer alan alt faktörlere bağlı olduğunu daha önceki satırlarda belirtilmişti. Çalışmada iş arayan ana faktörünün işe yerleştirmede etkisinin ortaya konulması amaçlanmıştır. Bu bağlamda iş arayan işsizlerin öznitelikleri ve bu özniteliklerin açıklaması, türü ve modeldeki durumu aşağıda yer alan tablodaki şekilde belirlenmiştir.

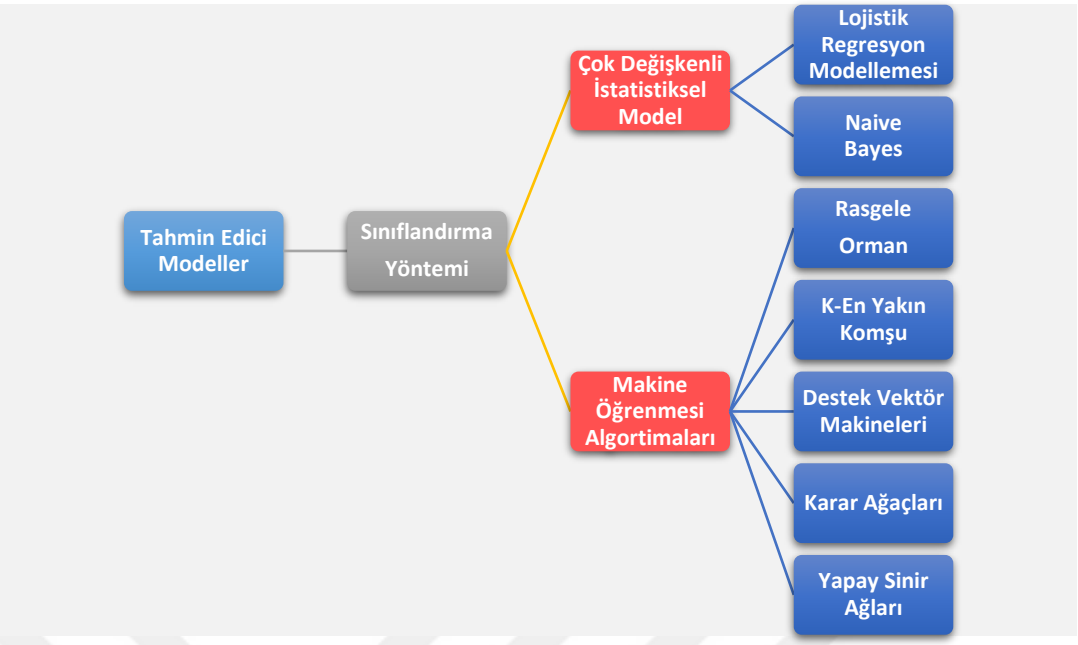
Öznitelik tablosu incelendiğinde özniteliklerin tamamının kategorik hale getirildiği görülmektedir. Bu veri dönüşümünün model performanslarını olumlu etkileyeceği değerlendirilmektedir.

Tablo 3.4 Özniteliklerin Türü, Açıklaması ve Modeldeki Durumu

Öznitelikler	**Nitelik Açıklaması**	**Nitelik Türü**	**Modeldeki Durumu**
İkamet İlçe	İşsizin ikamet ettiği ilçe	Nominal	Açıklayıcı Öznitelik
Sosyal Durum	İşsizin iş arama sosyal durumu	Nominal	Açıklayıcı Öznitelik
Cinsiyet	İşsizin cinsiyeti	Nominal	Açıklayıcı Öznitelik
Yaş	İşsizin yaşı	Ordinal	Açıklayıcı Öznitelik
Medeni Durum	İşsizin medeni durumu	Nominal	Açıklayıcı Öznitelik
Sosyal Yardım Alma Durumu	İşsizin sosyal yardım alma durumu	Nominal	Açıklayıcı Öznitelik
Öğrenim	İşsizin öğrenimi	Ordinal	Açıklayıcı Öznitelik
Meslek Grup	İşsizin sahip olduğu mesleğin meslek grubu	Nominal	Açıklayıcı Öznitelik
Başvuru Türü	İşsizin iş arama başvuru türü(Çalışırken İşsiz/İlk kez)	Nominal	Açıklayıcı Öznitelik
İşsizlik Ödeneği Alma Durumu	İşsizin işsizlik ödeneği alma durumu	Nominal	Açıklayıcı Öznitelik
İşe Yerleşme Durumu	İşsizin başvuru sonrası bir yıl içerisinde işe yerleşme durumu	(Binary)	Açıklanan/Hedef Değişken/Target

3.7 Veri Madenciliği Modelleme Yöntemlerinin Seçimi

Veri hazırlama, ön işleme ve dönüştürme işlemlerinden sonra veri madenciliği sürecindeki bir sonraki aşama hazır hale getirilmiş veri setine uygun modelin belirlenmesidir. Veri madenciliğinde iki ana modelleme yöntemi bulunmaktadır. Bunlardan birincisi tahmin edici modellemeler diğer ise açıklayıcı/tanımlayıcı modellerdir. Tahmin edici modeller sınıflama modelleri ve regresyon modeli olmak üzere iki gruba ayrılmaktadır.



Şekil 3.2 Veri Madenciliği Kapsamında Uygulanacak Modeller

Çalışma kapsamında tahmin edici model grubunda yer alan sınıflama modeli İŞKUR'a kayıt yaptıran işsizlerin kayıt sonrası işe yerleşme(1)/yerleşmeme(0) durumu sınıflandırılmıştır. Sınıflandırma işlemi çok değişkenli istatistiksel modelleme yöntemi ve makine öğrenmesi algoritmaları kullanılarak iki şekilde yapılmıştır. Verilerin tamamının kategorik olması nedeniyle veri yapısına uygun çok değişkenli istatistiksel analiz yöntemi olan lojistik regresyon[13] modellemesi kullanılmıştır. Yine veri yapısına uygun olarak denetimli öğrenme algoritmalarından Rasgele Orman, K-En Yakın Komşu, Naive Bayes, Karar Ağaçları ve Destek Vektör Makineleri, Yapay Sinir Ağları algoritmaları kullanılmıştır.

3.8 Çok Değişkenli İstatistiksel Modelleme ve Boyutlandırma

Açıklayıcı özniteliklerin veya açıklanan hedef özniteliğin çoklu bir yapı arz etmesi durum çok değişkenli istatistiksel analizler ve modellemelere uygunluk göstermektedir. Ancak bu uygunluk veri tipinin sayısal ve kategorik olmasına göre farklılık arz etmektedir. Örneğin tamamı sürekli sayısal verilerden oluşan bir veri seti çok değişkenli doğrusal istatistiksel analize uygun iken kategorik verilerden oluşan bir veri seti de lojistik regresyona analizine uygundur.

[13] Her ne kadar lojistik regresyon modellemesi bazı kaynaklarda bir denetimli makine öğrenmesi algoritması olarak yer verilse de, aslında lojistik regresyon modellemesi çok değişkenli istatistiksel analiz yöntemi olup kendine özgü varsayımları ve istatistiksel testleri yapısında bulundurmaktadır.

Çalışmadaki hem açıklayan hem de açıklanan özniteliklerdeki tüm verilerin kategorik olması/kategorik hale getirilmesi ve işsizlerin işe yerleşip yerleşmemelerine göre sınıflandırılması amacıyla çalışmada lojistik regresyon modellemesi/analizi uygulanmıştır. Esasında lojistik regresyon analizinin gerçekleştirilmesi için doğrusallık, hata terimlerinin bağımsızlığı, çoklu doğrusal bağlantı, beklenen frekans sayısı şeklinde varsayımlar bulunmaktadır. Ancak çalışmanın uyum iyiliğine sahip bir model oluşturmaktan daha ziyade doğru sınıflandırmaya odaklanması, ayrıca verinin kategorik olması ve örneklem sayısının fazla olması ile merkezi limit teoremine[14] uygunluğu gibi nedenlerden dolayı analize ilişkin varsayım uygunluğu yapılmamıştır.

Çoklu lojistik regresyon analizi uygulanmadan önce çalışmada kullanılan bağımsız değişkenlerin tek değişkenli analizlerle incelenerek ilgisiz olanların modele alınmaması önerilmektedir. Bu amaçla, iki yaklaşımdan yararlanılır. Birinci yaklaşımda bağımlı/yanıt/sonuç değişkeni ile ilgili bağımsız değişkenler arasında bir ilişki olup olmadığı ki-kare, iki ortalama arasındaki farkın önemlilik testi, MannWhitney U testi gibi testlerden yararlanılarak incelenir. Diğer yaklaşımda ise her bir bağımsız değişkenle tek değişkenli lojistik regresyon analizi yapılması ve p değerlerinin incelenmesi gerekir. Her iki yaklaşım sonucunda(yani tek değişkenli analizler sonucunda) p değeri 0,25 altında bulunan değişkenlerin çok değişkenli çözümlemede dikkate alınması önerilmektedir. Böylece çok değişkenli çözümlemede daha az sayıda değişken ile ilgilenilir. Özellikle gözlem sayısının az olduğu çalışmalarda, bu yaklaşım çoklu bağlantı gibi sorunların ortaya çıkmasını önleyebilir [46]. Çalışmada birinci yaklaşım tercih edilerek öznitelikler/değişkenler arasındaki ilişki ki-kare analizi ile test edilerek p değeri 0,25 üstünde bulunan öznitelikler çok değişkenli çözümlemede dikkate alınmamış ve model alınmamıştır.

Lojistik regresyon modellemesi açıklanan hedef değişkendeki kategorik yapıya göre binary, ordinal, multinominal gibi birden fazla türü bulunmaktadır. Çalışmada açıklanan değişken işe yerleşme(1) ve işe yerleşmeme(0) olarak iki farklı değer aldığı için Binary Lojistik Regresyon Modellemesi kullanılmıştır. Binary lojistik regresyon modellemesine göre yordanan değişkenler 0 ile 1 arasında olasılıklı sonuçlar

[14] Yığının dağılım normal değilse, n arttıkça $\bar{x}$ örnekleme dağılımının normale yaklaştığı söylenebilir [49].

vermektedir. Sonuçların 0,50'den büyük olması durumunda işsizler işe yerleşmiş, 0,50'den küçük olması durumunda ise yerleşmemiş olacaklardır.

Lojistik regresyon analizinde modele alınacak öznitelikler bir dizi yönteme tabi tutulurlar. Bu yöntemler aşağıda sıralı bir şekilde yer verilmiştir.

1. Öncelikle modele alınacak veri setinin veri madenciliği veri hazırlama, işleme ve dönüştürme süreçlerinden geçirilerek modellemeye hazır hale getirilmelidir.
2. Sonrasında modele verinin tamamının mı yoksa bir örnekleminin mi alınması kararı verilir. Eğer sadece bir sınıflama işlemi yapılacaksa verinin tamamı tıpkı makine öğrenmesi yöntemlerindeki gibi alınabilir. Ancak ileriye yönelik bir kestirim yapılacaksa veriden örneklem çekilmesi gerekmektedir.
3. Modelde alınacak veri tipi belirlendikten sonra açıklanan hedef öznitelik ile açıklayıcı öznitelikler arasında doğrusal veya doğrusal olmayan ilişkinin tespitidir. Bu tespitler ki-kare analizi, korelasyon analizi ve tek değişkenli logit model analizidir. Bu analizlere göre hedef öznitelik ile istatistiksel olarak yeterli ilişkiye sahip öznitelikler seçilir.
4. Belirlenen öznitelikler arasında çoklu doğrusal bağlantı sorunu olan öznitelikler belirlenir ve uygun olmayan öznitelik çıkarılır.
5. Sadece sabit değerin ve hedef özniteliğin yer aldığı basit başlangıç modeli oluşturulur.
6. Sabit değer ve ilişki analizlerine göre modele alınması uygun görülen öznitelikler için ana modelleme yapılır. İlk oluşturulan basit model ile son oluşturulan ana model karşılaştırılır ve iki model arasındaki fark test edilir.
7. Ana modelde yer almasına rağmen istatistiksel anlamlığa sahip olmayan öznitelikleri modelden çıkarmak için enter, ileriye doğru ve geriye doğru çıkarma yöntemlerinden biri yapılarak öznitelik çıkarımı işlemi gerçekleştirilir. Öznitelik çıkarma işlemi ile amaçlanan modele ulaşılıp ulaşılmadığı test edilir.
8. Oluşturulan amaçlanan modele ilişkin model parametreleri ODDS değeri(Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı) üzerinden yorumlanır.
9. Ki-kare analizi ve ODDS oranlarına göre risk grupları belirlenir.
10. Son olarak oluşturulan amaçlanan modele ilişkin etraflıca bir genel değerlendirme yapılır.

Lojistik Regresyon Uygulaması

Lojistik regresyon uygulaması kapsamında yukarıda belirtilen işlemler aşağıda sıralı bir şekilde uygulanmıştır.

1. Veri hazırlama, ön işleme ve dönüşüm

Veri seti daha önce yapılan işlemlerle modellemeye hazır hale getirildiğinden tekrar bir işlem yapmaya gerek kalmamıştır.

2. Veri Seti Tipinin Belirlenmesi

Lojistik regresyon modellemesinde kestirim yapılacağı için veri setini en iyi temsil edilen örneklemin belirlenmesi gerekmektedir. Örneklemin belirlenmesi için çeşitli yöntemler[15] bulunmaktadır. Genelde, çok değişkenli analizlerden iyi sonuç elde edilebilmesi için değişken sayısının 10 katı gözlem ile çalışmanın yürütülmesi önerilmektedir [46]. Bu çalışmada öncelikle veri yığını özniteliklere göre sıralanmış ve örneklem sayısına göre tabakalı ve sistematik örneklem kullanılarak seçim yapılmıştır.

Örneklem sayısının belirlenmesinde bir takım formüller ve tablolar bulunmaktadır. Çalışmada işe yerleşme/işe yerleşmeme gibi bir oran hesabının yapılması belirtilen formül[16] ve tablolar üzerinden[17] hesaplanan örneklem sayısının çok düşük olmasına(n=374) neden olduğu gibi örnek sayısının en yüksek n=1000 olmasına da neden olmaktadır. Bu durum ise örneklememizin yığını çok iyi temsil etmemesine veya iyi bir model oluşturulamamasına neden olmaktadır. Bu nedenden dolayı veri yığınından %10, %15 ve %20 oranında örnekler çekilmiş ve ODDS[18] oranındaki

[15] Tabakalı, sistematik, basit tesadüfi, kota, küme örneklemesi örnekleme teknikleridir.

[16] $n_0 = \left\lfloor \frac{(p \times q \times z^2)}{(d^2)} \right\rfloor, n = \frac{n_0}{\left(1+\frac{n_0}{N}\right)}$

$p = Gerçekleşme\ sıklığı$(İş$e\ Yerleşme\ Oranı$)

$q = Gerçekleşmeme\ sıklığı$(İş$e\ Yerleşmeme\ Oranı$)

$z = Güven\ düzeyi$

$d = Hata\ payı$

$n = Yığından\ çekilmesi\ gereken$ örnek hacmi

$N = Yığındaki$ İssiz Sayısı

[17] Yığın sayısının 14.146 olduğu bir veri setinden örnekleme hatası(d=0,05), p=0,5, q=0,5 ve %95 güven düzeyinde(z=1,96) n=374 örnek çekilmektedir.

[18] Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı.

sapmalar hesaplanmıştır. 14.146 işsizin yer aldığı yığından %10 oranında örneklem çekildiğinde 1.415 işsiz, %15 oranında örneklem çekildiğinde 2.021 işsiz ve %20 oranında örnekleme çekildiğinde ise 2.830 işsiz belirlenmektedir.

Gerçekleştirilen hesaplamalara göre ODDS Hata ve Oransal Hata ortalaması en düşük olan %20 oranındaki örneklem ile analiz ve modelleme yapılmasına karar verilmiştir. %20 oranında yığından çekilen örneklem, öznitelik ve kategori düzeyinde yığından sadece %0,5 oranında farklılık arz etmektedir. Bu haliyle model oluşumuna kaynaklık edecek örneklem verisi yığın verisi ile neredeyse bire bir benzerlik arz ederek iyi bir analizin yapılmasına ve modelin oluşturulmasına olanak sağlamaktadır.

Tablo 3.5 Örneklem Oranına Göre Ortalama ODDS ve Oransal Hata Oranı

	%10 Örneklem		**%15 Örneklem**		**%20 Örneklem**	
Örneklem Hacmi	1.415		2.021		2.830	
Hata Türleri	ODDS Hata	Oransal Hata	ODDS Hata	Oransal Hata	ODDS Hata	Oransal Hata
Ortalama	9,7%	0,9%	4,4%	0,7%	1,2%	0,5%

3. İlişki Analizi İle Öznitelik Seçimi

Öznitelikler/Değişkenler arasındaki doğrusal/doğrusal olmayan, nedensel/nedensel olmayan ve neden/sonuç ilişkinin belirlenmesinde birden farklı türde analizler bulunmaktadır. Bu analizlerden en bilinenleri korelasyon analizi, ki-kare analizi, regresyon analizidir. Bununla birlikte path analizi, yapısal eşitlik modeli, kanonik regresyon gibi daha az bilinen analizlerde bulunmaktadır. Ancak bu analizlerin tamamı bütün veri tipleri için uygun değildir. Verinin sayısal ve kategorik olmasına göre ilişki analizi tipide değişmektedir. Örneğin sayısal veriler için korelasyon ve regresyon analizinin uygulanması uygun iken, kategorik veriler için ki-kare analizi ve lojistik regresyona analizi daha uygundur. Bu çalışmada hem açıklayıcı hem de açıklanan değişkenlerin kategorik olması nedeniyle ki-kare analizi tercih edilmiştir. Bununla birlikte kategorilerin kodlanması ile elde edilen sayısal verilere de korelasyon analizi uygulanarak değişkenler arasındaki hem doğrusal hem de doğrusal olmayan ilişkiler ortaya koyularak birbiriyle kıyas edilmiştir.

Ki-kare Analizi Sonuçlarına Göre İlişkiler

Ki-kare analizi ile işsizlerin İŞKUR'a başvuru sonrası işe yerleşip yerleşmeme durumu ile işsizin öznitelikleri arasında bir ilişki olup olmadığı test edilmiştir. Bu test işlemi hâlihazırda tez çalışmasının varsayımlarından biridir. Analiz sonuçlarına göre işsizin

işe yerleşmesi durumu ile işsizin sosyal yardım alma durumu dışındaki tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel olarak anlamlı bir ilişki bulunmaktadır. Başka bir ifadeyle işsizin işe yerleşmesi durumu özniteliklerden bağımsız değildir.

H_0=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki yoktur.
H_1=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki vardır.

Kontenjans katsayısı ile ilişki gücü[19] incelendiğinde işsizin işe yerleşmesi ile tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. İlişki değerlerine göre en yüksek ilişki 0,18 ile işsizin yaşı ile 0,17 ile medeni durumu ve mesleği iledir. En düşük ilişki ise 0,01 ile işsizin ikameti ve 0,05 korelasyon ile işsizlik ödeneği alma durumu ve 0,07 korelasyon başvuru türü ve sosyal durumu iledir. Her ne kadar işsizin başvuru sonrası işe yerleşmesi ile demografik, kişisel ve iş arama öznitelikleri arasında çok zayıf ilişki bulunsa da bu ilişkilerin bir modelde birlikte değerlendirilmesi daha güçlü bir yapı ortaya koyabilir. Zaten istatistiksel modellemenin bir amacı da düşük ilişkilerden açıklayacılığı yüksek bir model oluşturmaktır. İşsizin işe yerleşmesi ile öznitelikleri arasındaki ilişkinin varlığı bu özniteliklerin işe yerleşme durumunun hedef değişken olarak yer aldığı bir modele girebilmesine olanak sağlamaktadır. Ki-kare analizi sonuçlarına göre p değeri 0,25'ten büyük olan sosyal yardım alma durumuna ilişkin özniteliğin dışında tüm öznitelikler modele girmeye adaydırlar.

Tablo 3.6 İşe Yerleşme ile İşsizin Özellikleri Arasındaki İlişkiler(Örneklem, Ki-Kare)

ÖZNİTELİKLER	Test istatistiği	S. d	P Değeri	İlişki Varlığı	İlişki Gücü	İlişki Derecesi
İkamet	4,8	1	0,028	Var	0,014	Çok Zayıf
Sosyal Durum	15,1	1	<0,001	Var	0,073	Çok Zayıf
Cinsiyet	24,2	1	<0,001	Var	0,092	Çok Zayıf
Yaş1	89,0	6	<0,001	Var	0,175	Çok Zayıf
Yaş2	87,6	4	<0,001	Var	0,173	Çok Zayıf
Medeni Durum1	76,0	3	<0,001	Var	0,162	Çok Zayıf
Medeni Durum2	72,0	1	<0,001	Var	0,158	Çok Zayıf
Sosyal Yardım Alma Durumu	0,4	1	0,519	Yok	0,012	-
Öğrenim	16,6	4	0,002	Var	0,076	Çok Zayıf
Meslek	79,3	6	<0,001	Var	0,165	Çok Zayıf
Başvuru Türü	14,7	1	<0,001	Var	0,072	Çok Zayıf
İşsizlik ödeneği Alma Durumu	8,2	1	0,004	Var	0,054	Çok Zayıf

[19] **Pearson'un Kontenjans Katsayısı (Contingency Coefficient):** Kontenjans katsayısı, katsayısının IxJ boyutlu tablolardaki iki değişken arasındaki ilişkinin büyüklüğünü ölçen biçimidir [48].

Yaş2 değişkeni ile Yaş1 değişkeninin kategori sayısı ODDS oranlarına göre azaltılmış ancak ilişki gücünde belirgin bir artış gerçekleşmediği için Yaş1 değişkenin modelde yer almasına karar verilmiştir. Bununla birlikte Medeni Durum1 değişkeninin de "Dul" kategorisindeki örneklem sayısının çok düşük olması ve "Bekâr" ve "Boşanmış" kategorisindeki işsizlerin işe yerleşmesi ODDS oranlarının birbirine yakın olması nedeniyle bekâr, boşanmış ve dul olan işsizler için "Evli Olmayan" şeklinde yeni bir kategori oluşturulmuş ve evli olan işsizler için de ve "Evli Olan" şeklinde bir kategori oluşturarak iki kategorili Medeni Durum2 değişkeni meydana getirilmiştir. Analiz sonuçlarına göre ilişki derecesi yönünden Medeni Durum1 değişkeni ile Medeni Durum2 değişkeni arasında belirgin bir fark olmadığı için Medeni Durum2 değişkenin modele girmesine karar verilmiştir. Ancak önemle belirtmek gerekir ki ki-kare analizi sonuçları özniteliklerin tek başına modele alınması için yeterli olmayıp, özniteliklerin modele girip girmeyeceğine korelasyon analizi ve çoklu bağlantı değerlendirilmesi sonrasında karar verilecektir.

Korelasyon Analizi Sonucuna Göre İlişkiler

Korelasyon analizi[20] ile işsizin İŞKUR'a başvuru sonrası işe yerleşip yerleşmeme durumu ile işsizin öznitelikleri arasında doğrusal bir ilişki olup olmadığı test edilmiştir.

Tablo 3.7 İşe Yerleşme ile İşsizin Özellikleri Arasındaki İlişkiler(Örneklem, Korelasyon)

Öznitelikler	P Değeri	İlişki Varlığı	İlişki Gücü	İlişki Derecesi
İkamet	0,028	Var	0,041	Çok Zayıf
Sosyal Durum	< ,001	Var	-0,073	Çok Zayıf
Cinsiyet	< ,001	Var	-0,093	Çok Zayıf
Yaş1	< ,001	Var	-0,128	Çok Zayıf
Yaş2	< ,001	Var	-0,118	Çok Zayıf
Medeni Durum1	< ,001	Var	-0,163	Çok Zayıf
Medeni Durum2	< ,001	Var	-0,159	Çok Zayıf
Sosyal Yardım Alma Durumu	0,519	Yok	-0,012	-
Öğrenim	0,322	Yok	-0,019	-
Meslek	< ,001	Var	0,094	Çok Zayıf
Başvuru Türü	< ,001	Var	0,072	Çok Zayıf
İşsizlik Ödeneği	0,004	Var	-0,054	Çok Zayıf

[20] Korelasyon katsayısı iki değişken arasındaki doğrusal ilişkinin yönünü ve derecesini belirtir. Değişkenler arasında doğrusal olmayan çok sıkı ilişkiler olsa bile, korelasyon katsayısı sıfır ya da sıfıra yakın çıkabilir [49]

Analiz sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin sosyal yardım alma ve öğrenim durumu dışındaki tabloda yer alan tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel olarak anlamlı doğrusal bir ilişki bulunmaktadır.

H_0=İşsizin öznitelikleri ile işe yerleşmesi arasında doğrusal ilişki yoktur.
H_1=İşsizin öznitelikleri ile işe yerleşmesi arasında doğrusal ilişki vardır.

İlişki gücü incelendiğinde işsizin işe yerleşmesi ile doğrusal ilişkili tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. İlişki değerlerine göre en yüksek doğrusal ilişki 0,16 korelasyon ile işsizin medeni durumu ile 0,13 ile yaşı ve 0,09 ile mesleği ve cinsiyeti iledir. En düşük doğrusal ilişki ise 0,04 korelasyon ile işsizin ikameti ve 0,05 ile işsizlik ödeneği alma durumu ve 0,07 ile başvuru türü ve sosyal durumu iledir. İşsizin başvuru sonrası işe yerleşmesi ile demografik, kişisel ve iş arama öznitelikleri arasında çok zayıf doğrusal ilişki bulunsa da bu ilişkilerin bir modelde birlikte değerlendirilmesi daha güçlü bir yapı ortaya koyabilir. Korelasyon analizi sonuçlarına göre işsizin sosyal yardım alma durumu ve öğrenimi özniteliğin dışındaki tüm öznitelikleri modele girmeye adaydırlar.

Yaş1 özniteliği ile Yaş2 özniteliği arasında ilişkinin gücü yönünden bir farklılık olmadığı için Yaş1 özniteliğinin modelde yer alması daha uygun görülmüştür. Yine aynı durum Medeni Durum1 ve Medeni Durum2 özniteliğinde de görüldüğü için Medeni Durum2 özniteliğinin modelde yer alması daha uygun görülmüştür. Hem ki-kare hem de korelasyon analizi sonuçlarında Yaş1 ve Medeni Durum2 özniteliğinin modelde yer almasının uygun görülmesinde ODDS hata ve kategori içerikleri dikkate alınmıştır.

Ki-kare analizi ve korelasyon analizi sonuçlarına göre işsiz bireyin işe yerleşmesi ile öznitelikleri arasında hem doğrusal hem de doğrusal olmayan ilişki varlığı, gücü ve derecesi arasında yüksek benzerlikler ortaya çıkmıştır. Bundan sonraki aşamada işsizin öznitelikleri arasındaki çoklu doğrusal bağlantı (yüksek ilişki(r>0,60)) kontrol edilerek yüksek ilişkili değişkenlerden biri modele dâhil edilmeyecektir.

4. Öznitelikler Arasındaki Çoklu Doğrusal Bağlantı Sorunu

Öznitelikler arasında 0,60 ve üzeri korelasyon olması çoklu doğrusal bağlantıyı işaret etmektedir. Bazı çalışmalarda çoklu doğrusal bağlantı oran 0,70 hatta 0,80 olarak ta kabul edilmektedir. Öznitelikler arasındaki korelasyon incelendiğinde öznitelikler

arasında sadece Yaş1 özniteliği ile Medeni Durum2 özniteliği arasında 0,607 değerinde orta düzeyde doğrusal ilişki olup çoklu doğrusal bağlantı sorunu bulunmaktadır. Ancak hem korelasyonun orta düzeyde ve sınırda olması hem de daha yüksek oranlar için çoklu doğrusal bağlantı olması gerekliliği yönündeki görüşler nedeniyle söz konusu çoklu doğrusal bağlantı değerlendirmeye alınmayarak hem Yaş1 özniteliği hem de Medeni Durum2 özniteliğinin modelde yer alması uygun görülmüştür.

Ki-kare analizi, korelasyon analizi ve çoklu doğrusal bağlantı sorunu birlikte değerlendirildiğinde işsizin sosyal yardım alma durumu özniteliği dışındaki tüm özniteliklerinin modele alınması uygun görülmüştür. İşsizin sosyal yardım alma durumu ile işsizin işe yerleşmesi arasında istatistiksel olarak anlamlı bir ilişki olmamasında sosyal yardım alan işsiz sayısının çok az olmasının kaynaklık ettiği değerlendirilmektedir. Esasında işsiz bireyin sosyal yardım alma durumu işsizin gelir ile ilişkili bir durumdur. Gelir ise işsizin işe ihtiyacını, iş arama ve çalışma isteğini doğrudan ve dolaylı olarak etkileyen bir özniteliktir. Ancak bu çalışmadaki sosyal yardım alma durumuna ilişkin veri yapısı işsizin geliri ile olan bu ilişkiyi ortaya koymak için oldukça yetersizdir. Başka çalışmalarda ve daha yeterli veriler ile bu öznitelik tekrar değerlendirilebilir.

Tablo 3.8 Öznitelikler Arasındaki Korelasyon

Öznitelikler	İkamet	Sosyal Durum	Cinsiyet	Yaş1	Medeni Durum2	Sosyal Yardım Alma Durumu	Öğrenim	Meslek	Başvuru Türü	İşsizlik Ödeneği Alma Durumu
İkamet	—									
Sosyal Durum	0,025	—								
Cinsiyet	0,065	0,111	—							
Yaş1	0,016	-0,14	0,041	—						
Medeni Durum2	0,004	-0,018	0,156	0,607	—					
Sosyal Yardım Alma Durumu	0,042	-0,01	0,002	-0,003	-0,012	—				
Öğrenim	0,002	0,078	0,118	-0,258	-0,294	-0,02	—			
Meslek	0,003	0,06	-0,134	0,044	0,013	-0,003	-0,094	—		
Başvuru Türü	0,041	-0,056	-0,198	0,271	0,1	-0,016	0,052	0,042	—	
İşsizlik Ödeneği Alma Durumu	-0,048	-0,04	0,099	-0,23	-0,152	-0,013	0,123	-0,089	-0,062	—

5. Basit Başlangıç Lojistik Regresyon Modelinin Oluşturulması

Sadece sabit değerin ve hedef özniteliğin yer aldığı model basit başlangıç modeli olup amaçlanan model ile en iyi modele ulaşılıp ulaşılmadığı test edilmesi için kullanılacaktır.

Tablo 3.9 Basit Başlangıç Lojistik Regresyon Modellemesi(Örneklem)

	Katsayı	Ölç. Hata	Z	P-değeri
Const(Sabit Değer)	-0,007	0,038	-0,188	0,851

Özniteliklerin yer almadığı basit başlangıç modelinde sabit değer için p değeri 0,05'ten büyük olması nedeniyle sabit değer %95 güven düzeyinde istatistiksel anlamlılık arz etmemektedir.

Tablo 3.10 Basit Başlangıç Lojistik Regresyon Model Özeti (Örneklem)

Sonuçlar	Değerler
McFadden R-kare	0,000
Log-olabilirlik	-1961,589
'Doğru kestirilen' durum oranı	50,2%

Model özetine göre oluşturulan basit model henüz işe yerleşme değişkenini açıklamamaktadır. Doğru kestirilen durum sayısı %50,2 oran ile orta seviyededir.

6. Ana Modelinin Oluşturulması

Sabit değer ve ilişki analizlerine göre modele alınması uygun görülen özniteliklerin yer aldığı ana modele oluşturulmuştur.

H_0=Başlangıç modelle ana model arasında fark yoktur.

H_1=Başlangıç modelle ana model arasında fark vardır.

Tablo 3.11 Ana Lojistik Regresyon Modeli (Örneklem)

	Katsayı	Ölç. Hata	Z	P-değeri
Const(Sabit Değer)	4,336	0,721	6,015	<0,0001
İkamet	0,202	0,096	2,105	0,035
Sosyal Durum	- 1,225	0,259	- 4,723	<0,0001
Cinsiyet	- 0,040	0,084	- 0,483	0,629
Yaş1	- 0,163	0,029	- 5,532	<0,0001
Medeni Durum2	- 0,557	0,111	- 5,018	<0,0001
Öğrenim	- 0,136	0,040	- 3,443	0,001
Meslek	0,114	0,024	4,679	<0,0001
Başvuru Türü	0,466	0,085	5,468	<0,0001
İşsizlik Ödeneği Alma Dur	- 0,841	0,197	- 4,268	<0,0001

Oluşturulan modele göre cinsiyet ve ikamet özniteliği dışındaki tüm öznitelikler için p değeri 0,05'ten küçük olduğu için %95 güven düzeyinde istatistiksel olarak anlamlıdır. Bununla birlikte ikamet özniteliği %90 güven düzeyinde anlamlı olup cinsiyet özniteliği bu düzeyde de anlamlı değildir. Bu bilgilere göre cinsiyet özniteliği modelden çıkarılmaya aday görülmektedir.

Tablo 3.12 Ana Lojistik Regresyon Model Özeti (Örneklem)

Sonuçlar	Değerler
McFadden R-kare	0,050701 (5,1%)
Log-olabilirlik	-1862,134
'Doğru kestirilen' durum oranı	59,2%
Olabilirlik Oran Sınaması Kikare(9)	198,911 [0,0000]

Model özeti verilerine göre log-olabilirlik değerinin basit modele göre biraz düştüğü, olabilirlik oran sınamasında[21] p değerinin de 0,05'ten küçük bir değer alarak H_0 hipotezi reddedilir ve H_1 hipotezi kabul edilir. Başka bir ifadeyle oluşturulan ana model ilk oluşturulan basit modelden farklılık arz etmektedir. Her ne kadar bu farklılık olsa da modelde istatistiksel olarak anlamsız öznitelikler olduğu için bu özniteliklerin olmadığı yeni bir model oluşturulabilir.

Oluşturulan ana modeldeki McFadden R-kare(Açıklama Katsayısı)'ye göre modelde yer alan öznitelikler model hedef değişkeni olan işsizin işe yerleşme durumunu yaklaşık %5,1 oranında açıklamaktadır. Başka bir deyişle işsizin temel kişisel, demografik ve iş arama bilgileri işsizin İŞKUR'a başvuru sonrası bir yıl sürede işe yerleşmesini en fazla %5,1 oranında açıklamaktadır. Bu bilgi işsizin başvuru sonrası işe yerleşmesine etki edecek %95 oranında ana ve alt faktörlerin olduğunu göstermektedir.

Tablo 3.13 Ana Modele İlişkin Sınıflandırma Tablosu (Örneklem)

		Kestirilen	
		İşe Yerleşmemiş(0)	**İşe Yerleşmiş(1)**
Gözlenen	**İşe Yerleşmemiş(0)**	802	618
	İşe Yerleşmiş(1)	537	873

[21] **Olabilirlik oran testi;** Modeldeki tüm değişkenlerin birlikte test edilmesi mantığına dayanmaktadır. Modelde sadece sabit değer varken elde edilen indirgenmiş log-olabilirliği ile tüm modelin log-olabilirliği birlikte değerlendirilir**.**

Ana modele ilişkin sınıflandırma tablosuna göre doğru sınıflandırma oranı %59,2'dir. Bu oran yüksek bir oran olmayıp modelde yer alan öznitelik sayısının düşüklüğünden kaynaklandığı değerlendirilmektedir. İstatistiksel çok değişkenli modellemede amaç en az değişkenle en yüksek açıklama oranının yakalamaktır. Bu amaca ulaşılabilmesi içinde modelin olgunlaştırılması gerekmektedir. Oluşturulan modeldeki istatistiksel olarak anlamsız özniteliklerin çıkartılarak modelin daha olgunlaşması kararının verilmesi çalışmanın amacı ile ilintili bir meseledir. Tez çalışmasında daha yüksek bir açıklama ve doğru sınama oranı elde edilmesi amaçlandığı için modelde yer alan anlamsız özniteliklerin çıkartılması gerçekleştirilecektir. Bu öznitelik çıkartma işlemi ile veri setinde boyut düşürme/boyutlandırma işlemi de gerçekleştirilmiş olacaktır.

7. Amaçlanan Modelin Oluşturulması

Ana modelde yer almasına rağmen istatistiksel anlamlığa sahip olmayan öznitelikleri modelden çıkarmak için geriye doğru çıkarma yöntemi yapılarak öznitelik çıkarımı işlemi gerçekleştirilmiştir. Öznitelik çıkarma işleminde %95 güven düzeyi tercih edildiği için modelden sadece cinsiyet özniteliği çıkartılmış ve amaçlanan modele ulaşılmıştır. Amaçlanan modeldeki tüm öznitelikler %95 güven düzeyinde istatistiksel olarak anlamlıdır ve işsizin başvuru sonrası işe yerleşme durumunu açıklamaktadır.

H_0=Başlangıç modelle amaçlanan model arasında fark yoktur.

H_1=Başlangıç modelle amaçlanan model arasında fark vardır.

Tablo 3.14 Amaçlanan Lojistik Regresyon Modeli (Örneklem)

	Katsayı	Ölç. Hata	Z	P-değeri
Const(Sabit Değer)	4,339	0,721	6,015	<0,0001
Sosyal Durum	-1,239	0,258	-4,804	<0,0001
Yaş1	-0,164	0,029	-5,569	<0,0001
Medeni Durum2	-0,566	0,109	-5,182	<0,0001
Öğrenim	-0,140	0,039	-3,587	0,000
Meslek	0,116	0,024	4,769	<0,0001
Başvuru Türü	0,476	0,083	5,738	<0,0001
İkamet	0,198	0,096	2,074	0,038
İşsizlik Ödeneği Alma Durumu	-0,853	0,196	-4,357	<0,0001

Oluşturulan amaçlanan modeldeki McFadden R-kare(Açıklama Katsayısı)'ye göre modelde yer alan öznitelikler model hedef değişkeni olan işsizin işe yerleşme durumunu yaklaşık %5,1 oranında açıklamaktadır. Ancak bir önceki model olan ana

modelde çok küçük bir düşüş yaşanmıştır. Bu düşüş öznitelik çıkarma işleminden kaynaklanmaktadır. Hâlihazırda öznitelik arttıkça açıklama oranın artması beklenen bir gerçekliktir. Ancak önemle belirtmek gerekir ki öznitelik çıkarma işlemi en az öznitelik ile en iyi modele ulaşma amacına uygun düşmektedir. Modellemenin başlangıcından itibaren 3 adet öznitelik çıkartılarak 8 özniteliğin olduğu amaçlanan modele ulaşılmıştır.

Tablo 3.15 Amaçlanan Lojistik Regresyon Model Özeti (Örneklem)

Sonuçlar	Değerler
McFadden R-kare	0,050642 (5,1%)
Log-olabilirlik	-1862,250
'Doğru kestirilen' durum oranı	59%
Olabilirlik Oran Sınaması Kikare(9)	198,678 [0,0000]

Amaçlanan model özeti verilerine göre log-olabilirlik değerinin basit modele göre biraz düştüğü, olabilirlik oran sınamasında p değerinin de 0,05'ten küçük bir değer alarak H_0 hipotezi reddedilir ve H_1 hipotezi kabul edilir ve son oluşturulan amaçlanan modelimizde istatistiksel olarak anlamlıdır.

Tablo 3.16 Amaçlanan Modele İlişkin Sınıflandırma Tablosu (Örneklem)

		Kestirilen	
		İşe Yerleşmemiş(0)	**İşe Yerleşmiş(1)**
Gözlenen	**İşe Yerleşmemiş(0)**	805	615
	İşe Yerleşmiş(1)	546	864

Oluşturulan modele ilişkin sınıflandırma tablosuna göre doğru sınıflandırma oranı %59'dur. Amaçlanan modeldeki doğru sınıflandırma oranı bir önceki ana modele göre %0,2 azalış göstermiştir. Düşük sınıflandırma oranı işsizin temel kişisel, demografik ve iş arama bilgilerinin İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşme bilgisinin etkili bir şekilde tespit edilemeyeceğini de ortaya koymaktadır. Oluşturulan amaçlanan modeldeki açıklama oranın ve doğru sınıflandırma oranın düşüklüğü iki şeyin önemini ortaya koymakta ve gerekli kılmaktadır. Bunlardan birincisi başvuru sonrası işsizin işe yerleşmesinde etkin ve açıklama oranını yükseltecek diğer ana ve alt faktörlerin belirlenmesi, ikincisi ise daha yüksek sınıflandırma oranına sahip sınıflandırma tekniklerinin kullanılmasıdır.

8. Model Parametrelerinin[22] Yorumlanması

P(Y)=P(İşe Yerleşme)=$\frac{1}{1+e^{-z}}$

Elde edilen sonuç P(Y) eşitliğine konulduğunda 0 ile 1 arasında bir değer elde edilir. P(Y)<0,50 ise Y=0, P(Y)>0,50 ise Y=1 kabul edilir. Y=0 olduğunda işsizin İŞKUR'a başvuru sonrası işe yerleşmediği, Y=1 olduğunda ise yerleştiği kabul edilir. 0,05 anlamlılık seviyesinde model aşağıdaki şekilde oluşturulur.

Tablo 3.17 Amaçlanan Modele İlişkin ODDS Ratio(OR) Oranları (Örneklem V.S.)

Öznitelikler	**Kategoriler**	**ODDS Ratio (OR)[23]**
Sosyal Durum	Engelli	Referans[24]
	Normal	2,48
Yaş1	15-24	Referans
	25-34	1,74
	35-39	2,99
	40-44	1,18
	45+	2,28
Medeni Durum2	Evli Olmayan	Referans
	Evli	2,02
Öğrenim	İlkokul ve Altı	2,25
	İlköğretim	1
	Ortaöğretim(Lise)	Referans
	Yükseköğretim	1,23
Meslek	Büro Hizmetlerinde Çalışan Elemanlar	2,69
	Hizmet Ve Satış Elemanları	3,07
	Nitelik Gerektirmeyen İşlerde Çalışanlar	3,36
	Profesyonel Meslek Mensupları	5,75
	Sanatkârlar Ve İlgili İşlerde Çalışanlar	2,18
	Teknisyenler, Teknikerler Ve Yar. Prof. Mes.Men.	2,4
	Tesis Ve Makine Operatörleri Ve Montajcılar	Referans
Başvuru Türü	Çalışırken İşsiz Kalanlar	Referans
	İlk Kez İş Hayatına Atılan	1,15
İşsizlik Ödeneği Alma Durumu	İşsizlik Ödeneği Alanlar	Referans
	İşsizlik Ödeneği Almayanlar	1,66

[22] Özniteliklere ilişkin ağırlıklar/katsayılar örnekten elde edilerek kestirildiği için parametre olarak adlandırılmaktadır.

[23] ODDS RATİO(O)= $\frac{ODDS\ açıklanmak\ istenen}{ODDS\ referans}$

[24] Referans kategorisi ODDS oranı en düşük olan kategori olarak belirlenmiştir.

***Z**=4,339-1,239(SosyalDurum)-0,164(Yaş1)-0,566(MedeniDurum2)-0,140(Öğrenim)+0,116(Meslek)+0,476(BaşvuruTürü)+0,198(İkamet)+0,967(İşsizlik Ödeneği Alma Durumu)*

Modele göre işsizin başvuru sonrası işe yerleşme durumu ile işsizin mesleği, başvuru türü ve işsizlik ödeneği alma durumu arasında pozitif ilişki varken işsizin sosyal durumu yaşı, medeni durumu ve öğrenim arasında negatif ilişki vardır.

ODDS Ratio oranlarına göre İŞKUR 'a başvuru sonrası bir yıl içerisinde işe yerleşmemede(işsiz kalmada) daha fazla riskli olanlar aşağıda sırala bir şeklide gösterilmiştir.

- Sosyal durumu engelli olanlara göre normal olanlar yaklaşık 2,5 kat daha fazla risklidir.
- Yaşı 15-24 grubunda[25] olanlara göre 25-35 grubunda[26] olanlar yaklaşık 1,7 kat, 35-39 grubunda olanlar 3 kat, 40-44 grubunda olanlar 1,2 kat ve 45+ grubunda olanlar 2,3 kat daha fazla risklidir.
- Medeni durumu evli olmayanlara göre evli olanlar yaklaşık 2 daha fazla risklidir.
- Öğrenimi ortaöğretim olanlara göre ilkokul ve alt öğrenimde olanlar yaklaşık 2,3 kat, yükseköğretimde olanlar ise 1,2 kat daha fazla risklidir. Ortaöğretim düzeyinde öğrenimde olanlar ilköğretim düzeyinde öğretimde olanlar ile benzer riske sahiptir.
- Meslek grubu Tesis Ve Makine Operatörleri Ve Montajcılar olanlara göre Büro Hizmetlerinde Çalışan Elemanlar yaklaşık 2,7 kat, Hizmet Ve Satış Elemanları 3,1 kat, Nitelik Gerektirmeyen İşlerde Çalışanlar 3,4 kat, Profesyonel Meslek Mensupları 5,8 kat, Sanatkârlar Ve İlgili İşlerde Çalışanlar 2,2 kat, Teknisyenler, Teknikerler Ve Yardımcı Profesyonel Meslek Mensupları ise 2,4 kat daha fazla risklidir.
- Başvuru türü çalışırken işsiz kalanlara göre ilk kez iş hayatına atılanlar yaklaşık 1,3 kat daha fazla risklidir.
- İşsizlik ödeneği alanlara göre almayanlar yaklaşık 1,7 kat daha fazla risklidir.

[25] 15-19 ve 15-24 yaş gruplarının ODDS değerleri birbirine çok yakın olması ve bu iki yaş grubunun **genç yaş grubunu** oluşturması nedeniyle bu yaş grupları 15-24 yaş grubu şeklinde birleştirilmiştir

[26] 25-29 ve 30-34 yaş gruplarının ODDS değerleri birbirine çok yakın olması ve bu iki yaş grubunun **orta yaş grubunu** oluşturması nedeniyle bu yaş grupları 15-24 yaş grubu şeklinde birleştirilmiştir

9. Özniteliklere Göre Riskli Gruplarının Belirlenmesi

Çalışmanın en önemli amaçlarından biri de işsizlerin İŞKUR'a başvuru sonrası bir yıl içerisinde işsiz kalma risklerinin belirlenmesidir. Başka bir ifadeyle başvuru sonrası işsizleri işe yerleşmeme risklerinin tespit edilmesidir. İşsizlerin işsiz kalma risk tespiti kadar önemli bir diğer konu özniteliklere göre risk gruplarının belirlenmesidir. Özniteliklere göre risk tespiti işsizin özniteliği ile işe yerleşmesi/yerleşmemesi arasında ki-kare analizi sonuçlarına göre istatistiksel ilişkinin olması ve ODDS sonuçlarına göre irdelenmesi ile ortaya konulur. Hâlihazırda risk kestiriminde[27] bir göreceli risk ölçütü olan ODDS oranı tercih edilir. Ki-kare analizi sonuçlarına göre işsizin sosyal yardım alma özniteliği dışındaki tüm öznitelikleri ile başvuru sonrası bir yıl içerisinde işe yerleşmesi arasında istatistiksel anlamlı bir ilişki olduğu için söz konusu öznitelik dışındaki tüm öznitelikler için risk tespiti yapılmıştır. Risk tespitinde kategoriler "riskli" ve "daha riskli" olmak üzere iki gruba ayrılmıştır. Grupların tasnifi, işe yerleşmeme ODDS oranı düşük olan kategoriler "riskli", ODDS oranı yüksek olan kategoriler ise "daha riskli" olarak gerçekleştirilmiştir.

Tablo 3.18 Özniteliklere Göre İşsizlerin Başvuru Sonrası İşsiz Kalma Risk Grupları

	Riskli	**Daha Riskli**
İkamet	Merkez	İlçe
Sosyal Durum	Engelli	Normal
Cinsiyet	Erkek	Kadın
Yaş1	15-24(genç) Ve 40-44(İleri)	25-39(Orta) Ve 45+(Daha İleri)
Medeni Durum2	Evli Olmayanlar	Evliler
Öğrenim	İlköğretim Ve Üstü Öğrenime Sahip Olanlar	İlkokul Ve Altı Öğrenime Sahip Olanlar
Meslek	Tesis Ve Makine Operatörleri Ve Montajcılar	Diğer Meslek Grupları Mensupları
Başvuru Türü	Çalışırken İşsiz Kalanlar	İlk Kez İş Hayatına Atılanlar
İşsizlik Ödeneği Alma Durumu	İşsizlik Ödeneği Alanlar	İşsizlik Ödeneği Almayanlar

Risk gruplandırmasına göre ilçe de oturanlar, normal statüde olanlar, kadınlar, orta ve daha ileri yaşta olanlar, evliler, ilkokul ve altı eğitime sahip olanlar, tesis ve makine operatörü ve montajcısı dışında mesleğe sahip olanlar, ilk kez iş hayatına atılanalar ve işsizlik ödeneği almayanlar İŞKUR'a başvuru sonrası bir yıl içerisinde işsiz kalma riski daha yüksektir. Ancak önemle belirtmek gerekir ki sadece bu risk gruplaması

[27] **Risk Kestirimi**, genel risk ölçüleri ve göreceli risk ölçüleri olarak ele alınır [46].

üzerinden bir işsizin işsiz kalıp kalmayacağının değerlendirilmesi doğru sonuçlar vermeyebilir. Örneğin çalışırken işsiz kalan bir işsiz ilk kez iş hayatına atılan bir işsize göre daha az işsiz kalma riski taşımasına rağmen bu işsiz etkili bir iş arama davranışı sergilemediğinde etkili bir iş arama davranışı sergileyen ilk kez iş arayan işsizden daha fazla işsiz kalma riski taşıyacaktır. Dolayısıyla işsiz kalma risk hesabının çok yönlü ve çoklu öznitelikler ile yapılacak hesaplamalara göre belirlenmesi önem arz etmekte olup yukarıda yer verilen tablo temel risk gruplandırması olarak kabul edilmesi gerekmektedir.

Risk gruplandırmasına dair daha detaylı inceleme aşağıda yer verilmiştir.

- Merkezde açık iş ilanlarının ve iş kurabilme fırsatlarının ilçeye göre daha çok olması merkezde ikamet edenlerin ilçe de ikamet edenlere göre daha az işsiz kalma riski oluşturduğu,
- 4857 sayılı İş Kanunun "engelli ve eski hükümlü çalıştırma zorunluluğu" başlıklı 30. Maddesi gereğince hem kamu hem de özel sektör işyerleri belirli oranlarda engelli çalıştırma zorunluluğu bulunmaktadır. Bu pozitif ayrımcılığın pratikteki yansıması kapsamında engelli işsizler normal işsizlere göre daha az işsiz kalma riski taşıdığı,
- Erkek işsizlerin kadın işsizlere göre daha az işsiz kalma riski taşımasında ana sebebin ise hem toplum hem de il işgücü piyasası olarak erkek merkezli bir yapının oluşması ve mevcut işgücü piyasasının kadın istihdamına nitelik olarak çok uygun olmadığı,
- Esasında genç işsizlik ülke düzeyinde bir sorun iken orta yaş işsizlere göre daha az risk taşımasında ana sebebin il işgücü piyasasının az gelişmişliği ve daha az nitelikli işgücünün daha kolay ve hızlı işe girmesi olması,
- Evli bireylerin bakmakla yükümlü olduğu aile bireylerinin olması onların işe olan ihtiyaçlarını arttırdığı gibi iş arama ve çalışma isteklini arttırarak daha yüksek oranda ve daha hızlı işe yerleşerek daha az işsiz kalma riski taşımaları beklenirken evli olmayan bireylerin daha az işsiz kalma riski taşımlarında evli bireylerin iş tercih ve beklentilerinin daha üst seviyede olmasından kaynaklandığı,
- İlkokul ve altı öğrenime sahip olan işsizlerin daha üst öğrenime sahip olanlardan daha fazla işsiz kalma riski taşıması beklenir bir durum iken yükseköğretim mezunlarının daha alt öğrenime sahip işsizlere göre daha fazla işsiz kalma riski

taşıması yine il işgücü piyasasının az gelişmişliği başka bir ifadeyle daha az nitelikli işgücünden oluşması,

- Tesis ve makine operatörleri ve montajcıları meslek grubunda yer alan işsizlerin diğer meslek gruplarındaki işsizlere göre daha az işsiz kalma riski taşımasında özellikle dikiş makinesi operatörü mesleğine sahip işsizlerin istihdam edildiği ildeki tekstil faaliyetinin yüksek olmasından kaynaklandığı ayrıca nitelik gerektirmeyen meslekler, hizmet ve satış elemanları, büro hizmetinde çalışanlar, teknisyenler, teknikerler ve yardımcı profesyonel meslek mensupları, sanatkârlar ve ilgili işlerde çalışanlar ve profesyonel meslek mensuplarına göre daha az işsiz kalma riski taşımasında yine il işgücü piyasasının az gelişmişliğinden kaynaklandığı,
- Çalışırken işsiz kalanların bir iş arama ve çalışma tecrübesine sahip olması ve işgücü piyasası deneyimleri, onları ilk kez iş hayatına atılan işsizlere göre daha az işsiz kalma riski taşımalarına neden olmaktadır. Buna mukabil ilk kez iş hayatına atılan işsizlerin hem iş ve iş arama tecrübelerinin olmaması hem de iş beklenti ve tercihlerinin yüksek olması onları daha fazla işsiz kalma riski taşımlarına neden olduğu,
- İşsizlik ödeneği alan işsiz bireylerin daha az işsiz kalma riski taşımasında işsizlik ödeneği alma şartlarının ağırlığının neden olduğu,
- İşsizin yaşı ve medeni durumuna göre işsiz kalma risk sınıflamasında işsizin beklenti ve tercihlerini etkin iken diğer özniteliklerine göre işsiz kalma risk sınıflamasında işgücü piyasasının durumu ve gelişmişliğinin etkin olduğu değerlendirilmektedir.

10. Genel Değerlendirme

Amaçlanan model için elde edilen tüm bilgiler birlikte değerlendirildiğinde,

- ✓ Veri yığınından çekilen %20 oranındaki örneklem ile analiz ve modelleme yapıldığında ODDS Hata ve Oransal Hata ortalamasının minimize edildiği, ayrıca daha az maliyetli bir seçenek olarak %15 oranında da örnekte çekilebileceği,
- ✓ Ki-kare analizi sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin sosyal yardım alma durumu dışındaki çalışmada yer alan tüm öznitelikleri arasında istatistiksel olarak anlamlı bir ilişki bulunduğu ancak bu ilişki gücünün çok zayıf olduğu,

- ✓ Korelasyon analizi sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin sosyal yardım alma ve öğrenim durumu dışındaki çalışmada yer alan tüm öznitelikleri arasında istatistiksel olarak anlamlı doğrusal bir ilişki bulunduğu ancak bu doğrusal ilişki gücünün çok zayıf olduğu,
- ✓ Korelasyon ve ki-kare sonuçlarında yüksek benzerlikler olması nedeniyle her iki yöntemin öznitelik boyutlandırma işleminde ayrı ayrı veya birlikte kullanılabileceği,
- ✓ Lojistik regresyon modellemesi ile oluşturulan amaçlanan modelin çok düşük seviyede bir açıklama oranına(%5,1) sahip olduğu ve modelin istatistiksel olarak anlamlı doymuş bir model olduğu,
- ✓ Çok düşük açıklama oranın modelde yer alan öznitelik sayısının hem yetersiz olması hem de işe yerleşmede etkin değişkenler olmamasından kaynaklandığı,
- ✓ Çok düşük açıklama oranı nedeniyle işsizin başvuru sonrası işe yerleşmesine etki edecek %95 oranında başkaca ana ve alt faktörlerin olduğu,
- ✓ İşsiz bireyin başvuru sonrası işe yerleşmesinde işsizin temel kişisel, demografik ve iş arama bilgilerinin çok düşük bir etkisinin olduğu ve işsizin temel bilgileri/öznitelikleri dışında başkaca ana(iş arayan, açık iş ve işgücü piyasası) ve alt faktörlerin etkisinin de ölçülmesi gerekliliği,
- ✓ Amaçlanan çok değişkenli lojistik regresyon modelin orta seviyede ve yetersiz bir doğru sınıflandırma oranına(%59) sahip olduğu,
- ✓ Orta düzey doğru sınıflandırma oranına olanak sağlayan işsize ait temel kişisel, demografik ve iş arama bilgileri ile işsizin İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşme bilgisinin yüksek bir isabetle tespit edilemeyeceği ve başkaca bilgi ve tekniğe başvurulması gerekliliği,
- ✓ Tutarsız ve yanlış bir işe yerleşme bilgisi üzerine iş arayan danışmanlık hizmetlerinin bina edilemeyeceği ve doğru bir danışmanlık belirlenmesi için işsizin işe olan ihtiyacı, iş arama donanımı ve iş arama davranışı gibi bilgilerinin tespit edilmesi ve daha kapsamlı bir iş profillemenin yapılması gerekliliği,
- ✓ Amaçlanan modelde daha yüksek bir açıklama ve doğru sınıflandırma oranın elde edilmesi için işsizin işe yerleşmesinde etkin ana ve alt faktörlerin belirlenmesi, modele alınmasını ve daha yüksek sınıflandırma oranına imkân tanıyan yapay zekâ ve makine öğrenmesi algoritmalarının kullanılmasını gerekliliği

- ✓ Öznitelik seçme ve çıkarma işlemi ile yapılan boyut düşürmenin hem açıklama oranında hem de doğru sınıflandırma oranında düşük bir etkisinin olduğu, boyutlandırmadaki bu düşük etkide özniteliğin sayısı ve türünün belirleyici olduğu
- ✓ Risk gruplandırmasına göre ilçe de oturanlar, normal statüde olanlar, kadınlar, orta ve daha ileri yaşta olanlar, evliler, ilkokul ve altı eğitime sahip olanlar, tesis ve makine operatörü ve montajcısı dışında mesleğe sahip olanlar, ilk kez iş hayatına atılanalar ve işsizlik ödeneği almayanlar İŞKUR'a başvuru sonrası bir yıl içerisinde işsiz kalma riski daha yüksek olduğu ancak tek başına bu risk gruplaması üzerinden bir işsizin işsiz kalıp kalmayacağının değerlendirilmesi doğru sonuçlar vermeyeceği, işsizin yaşı ve medeni durumuna göre işsiz kalma risk gruplandırmasında işsizin beklenti ve tercihleri ön plana çıkmakta iken diğer özniteliklerde ise işgücü piyasasının ön plana çıktığı,
- ✓ Çok değişkenli analizlerin hem modelleme hem sınıflandırma hem de boyutlandırma işlemi gerçekleştirmeye imkân sağladığı ve öznitelik çıkarma ve seçme de etkin bir şekilde kullanılabileceği ancak yüksek bir doğru sınıflandırma gerçekleştirmediği için daha yüksek sınıflandırma oranı imkân sunan başkaca yöntemlere ihtiyaç bıraktığı,
- ✓ Genel olarak gerçekleştirilen çok değişkenli istatistiksel analizler ve testler ile işsizin İŞKUR'a başvuru sonrasında bir yıl içerisinde işe yerleşmesinin daha yüksek açıklama oranında tespit edilmesinin başkaca ana ve alt faktörlerin yer aldığı bir modelle sağlanacağı ve daha yüksek doğru sınıflandırma oranı içinde yapay zekâ ve makine öğrenmesi algoritmalarının denenmesi gerekliliği tespit edilmiş ve değerlendirilmiştir.

3.9 Makine Öğrenmesi Sistem Tasarımı

Yüksek bir açıklama ve doğru sınıflandırma oranın için çok değişkenli istatistiksel modellemenin dışında başkaca modelleme tekniklerinin kullanılması gerektiği daha önceki satırlarda belirtilmiştir. Bu modelleme tekniğinin en bilineni ve en çok tercih edileni yapay zekâ makine öğrenmesi tekniğidir. Bu tekniğin uygulanarak verinin bilgiye dönüştürülmesi için bir takım şartların yerine getirilmesi gerekmektedir. Esasında bu şartlar bütünü veri madenciliği süreci olarak da adlandırılmaktadır. Bu süreç üç ana aşamadan oluşmaktadır. Bu aşamalardan birincisi araştırma probleminin belirlenmesidir. İkincisi araştırma problemine uygun olarak ham verinin seçilip, düzenlenip, işlenip, dönüştürülüp ve en nihayetinde boyutlandırılarak modellemeye

hazır hale getirilmesi aşamasıdır. Üçüncü aşama da ise modellenmeye hazır hale getirilen verinin bir takım modelleme teknik ve yöntemleriyle modellenerek bilgiye dönüştürülmesidir. Üçüncü aşamada makine öğrenmesi tekniği tercih edilmesi makine öğrenmesi sistem tasarımını da beraberinde getirmektedir.

Makine öğrenmesi sistem tasarımı problemin tanımlanması ile başlayıp, model seçimi, model geçerleme yöntemlerinin belirlenmesi, özellik seçimi, model eğitimi ve inşası, model performans değerlendirme ve yorumlama, model performans iyileştirme, nihai modelin seçimi ile devam edip modelin yeni veri ile test edilmesi ve yeni veriye ilişkin tahmin yapılması ile son bulmaktadır. Çalışma kapsamında makine öğrenmesi sistem tasarımına ilişkin gerçekleştirilecek işlemler detaylı bir şekilde aşağıda özet şekilde sıralanmıştır.

Makine Öğrenmesi Sistem Tasarımında Gerçekleştirilen İşlemler

Tablo 3.19 Makine Öğrenmesi Sistem Tasarımı Aşamaları

	Aşamalar	**Açıklama**
1	**Problem 'in Tanımlanması**	İŞKUR'a kayıt yaptıran işsizlerin kayıt sonrası bir yıl içerisinde işe yerleşme(1)/yerleşmeme(0)'sinin Tahmin Edilmesi
2	**Model Tekniği**	Makine Öğrenmesi Teknikleri
3	**Model Seçimi**	Tahmin Edici Modeller
4	**Model Türü**	Denetimli Öğrenme(Sınıflandırma)
5	**Model Algoritmaları**	Rasgele Orman, K-En Yakın Komşu, Naive Bayes, Karar Ağaçları ve Destek Vektör Makineleri, Yapay Sinir Ağları
6	**Model Geçerleme /Doğrulama Yöntemi**	Hold Out, K-Çapraz Doğrulama
7	**Veri Seti Seçimi**	Yığın ve Örneklem
8	**Özellik Seçimi**	İkamet, Sosyal Durum, Cinsiyet, Yaş1, Medeni Durum(1 ve 2), Öğrenim, Meslek, Başvuru Türü, İşsizlik Ödeneği Alma Durumu
9	**Model Eğitimi ve İnşası**	(Eğitim=%70, Test=%30), (Eğitim=%80, Test=%20) K-5 Ve K-10 Çapraz Doğrulama
10	**Model Performans Değerlendirme**	Doğruluk(Accuracy),Kesinlik(Precision), Duyarlılık(Recall), F-Ölçütü
11	**Model Performans İyileştirme**	Veri Seti Değişimi, Algoritma Değişimi, Geçerleme Değişimi, Özellik Değişimi, Eğitim Değişimi, Algoritma Parametre Değişimi
12	**Nihai Modelin Seçimi**	Model Performans Değerleri, En İyi Model
13	**Model Veri Tahmini**	Yeni Veri İle Nihai Model Üzerinden Tahmin Yapılması

1) Çalışmanın ana problemlerinden biri daha önceki satırlarda "İşsizlerin zamana göre işsiz kalma risklerinin tespitinde yapay zekâ makine öğrenmesi yöntemleri klasik istatistiksel yöntemlere göre daha mı etkili?" şeklinde olduğu belirtilmişti. Makine öğrenmesi sistem tasarımına ilişki ara problem ise "İŞKUR'a kayıt yaptıran işsizlerin kayıt sonrası bir yıl içerisinde işe yerleşme(1)/yerleşmeme(0)'sinin tahmin edilmesi" şeklinde belirlenmiştir.
2) Model Tekniği olarak araştırma problemine uygun olarak daha önce çok değişkenli lojistik regresyon modellemesi yapılmış olup bu kısımda ise makine öğrenmesi tekniği kullanılmıştır.
3) Veri madenciliğinde iki ana modelleme bulunmaktadır. Bunlardan birincisi tahmin edici modellemeler diğer ise açıklayıcı/tanımlayıcı modeller olup çalışma probleminin yapısına uygun olarak tahmin edici modeller seçilmiştir.
4) Çalışma sorusu bir sınıflandırma işlemini gerektirdiğinden denetimli makine öğrenmesi türü kullanılmıştır.
5) Denetimli makine öğrenmesi kapsamında sınıflama işlemini gerçekleştiren Rasgele Orman, K-En Yakın Komşu, Naive Bayes, Karar Ağaçları ve Destek Vektör Makineleri, Yapay Sinir Ağları algoritmaları kullanılmıştır.
6) Model geçerleme yöntemi olarak hem Hold Out hem de K-Çapraz Doğrulama yöntemi kullanılarak model performanslarının kıyas edilmesi sağlanmıştır.
7) Çalışmada iki farklı veri seti bulunmaktadır. Bunlardan birincisi tüm işsizlerin yer aldığı yığın ve diğer ise yığından %20 oranda çekilen örneklemdir. Çalışma kapsamında yığın ve örneklem verilerine göre elde edilen model performansları kıyas edilmiştir.
8) Öznitelik seçme ve çıkarma işlemine göre indirgenmiş veri seti ile indirgenmemiş veri seti için makine öğrenmesi algoritmaları uygulanmış ve her iki veri setinin model performansları kıyas edilmiştir.
9) Modelin eğitimi dört farklı şekilde gerçekleştirilmiştir. Birincisinde veri setinin %70 eğitim, %30'u test, ikincisinde veri setinin %80'i eğitim ve %20'si test, üçüncüsünde K-5 çapraz doğrulama ve dördüncüsünde K-10 çapraz doğrulamadır. Böylece makine öğrenme işleminin öğrenme türü ve şekline göre model performansları değerlendirilmiştir.
10) Model performansları Doğruluk, Kesinlik, F-Ölçütü ve Duyarlılık gibi ölçütler üzerinden değerlendirilmiştir. Model algoritması, geçerleme yöntemi, veri seti,

özellik seçimi, eğitim şekline göre model performans değerlendirmeye tabi tutulmuştur.

11) Model performansının iyileştirilmesi için model algoritmalarının parametrelerinde değişime gidilmiştir.
12) Performans değerlerine göre en iyi model nihai model olarak seçilmiştir.
13) Yeni veya veri setinde yer alan bir veri üzerinden nihai model ile sınıflandırma tahmin işlemi gerçekleştirilerek çalışma probleminin çözümü için gerekli olan bilgiye ulaşılmıştır.

3.10 Makine Öğrenmesi Uygulaması ve Bulgular

İŞKUR'a kayıt yaptıran işsizlerin kayıt sonrası bir yıl içerisinde işe yerleşme(1)/yerleşmeme(0)'sinin tahmin edilmesi için makine öğrenmesi tekniklerinden bir tahmin edici model türü olan denetimli öğrenme ile model algoritmalarının veri seti, model geçerleme yöntemi, özellik seçimi ve model eğitim şeklinde göre performansları aşağıda tespit edilmiştir. Öncelikle yığın veri setine göre söz konusu uygulamalar gerçekleştirilmiş sonrasında aynı uygulamalar örneklem veri seti içinde yapılarak her iki veri setindeki performans değerleri karşılaştırılmıştır. Tüm işlemler Google Colab platformunda Python yazılım dilinde yapılmıştır.

Yığın Veri Setinde Öznitelik Boyutlandırma

Veri madenciliği sürecinin en önemli aşamalarından biride modelleme işlemi için belirlenen veri setindeki özniteliklerin boyutlandırılmasıdır. Öznitelik boyutlandırma öznitelik seçimi ve öznitelik çıkarımı olmak üzere iki şekilde yapılmaktadır. Daha önceki konu başlığında yığın veri setinden seçilen %20 oranında örneklem için boyutlandırma işlemi gerçekleştirilmiştir. Bu başlık altında ise yığın veri seti için boyutlandırma işlemi yapılmıştır.

Boyutlandırma işleminde sırasıyla,

1. Öznitelikler ile İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşme arasındaki ilişkileri ki-analizi ile tespit edilmiştir.
2. Daha sonra korelasyon analizi ile ilişkiler arasındaki çoklu doğrusal bağlantının varlığı tespit edilmiştir.
3. En sonunda lojistik regresyon modellemesi ile modelden çıkartılacak öznitelik belirlenmiştir.

Ki-Kare Analizi Sonuçlarına Göre İlişkiler

Ki-Kare analizi ile işsizin İŞKUR'a başvuru sonrası işe yerleşip yerleşmeme durumu ile işsizin öznitelikleri arasında bir ilişki olup olmadığı yığın veri seti üzerinde test edilmiştir. Analiz sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel olarak anlamlı bir ilişki bulunmaktadır.

Kontenjans katsayısı ile ilişki gücü[28] incelendiğinde işsizin işe yerleşmesi ile tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. Bu ilişki sonuçlarına göre işsizin tüm öznitelikleri modele girmeye aday konumundadır. Ancak işsizin sosyal yardım alma durumu özniteliğinin ilişki derecesinin ve test istatistiğinin düşük olması ve bu özniteliği ait veri sayısının az olması nedeniyle bu öznitelik modele alınmamış olup diğer özniteliklerin tamamı model girme potansiyeline sahiptirler. Ancak önemle belirtmek gerekir ki ki-kare analizi sonuçları özniteliklerin tek başına modele alınması için yeterli olmayıp, özniteliklerin modele girip girmeyeceğine korelasyon analizi ve çoklu bağlantı değerlendirilmesi sonrasında karar verilecektir.

Tablo 3.20 İşe Yerleşme ile İşsizin Özellikleri Arasındaki İlişkiler(Yığın, Ki-Kare)

ÖZNİTELİKLER	Test istatistiği	S. d	P Değeri	İlişki Varlığı	İlişki Gücü	İlişki Derecesi
İkamet	26,15	1	<0.001	Var	0,043	Çok Zayıf
Sosyal Durum	75,47	1	<0.001	Var	0,0728	Çok Zayıf
Cinsiyet	121,91	1	<0.001	Var	0,0924	Çok Zayıf
Yaş1	439,93	6	<0.001	Var	0,1737	Çok Zayıf
Medeni Durum1	388,64	3	<0.001	Var	0,1635	Çok Zayıf
Sosyal Yardım Alma Durumu	5,94	1	0.015	Var	0,0205	Çok Zayıf
Öğrenim	81,21	4	<0.001	Var	0,0756	Çok Zayıf
Meslek	378,04	6	<0.001	Var	0,1613	Çok Zayıf
Başvuru Türü	68,32	1	<0.001	Var	0,0693	Çok Zayıf
İşsizlik ödeneği Alma Durumu	27,25	1	<0.001	Var	0,0439	Çok Zayıf

Korelasyon Analizi Sonuçlarına Göre İlişkiler

Korelasyon analizi ile işsizin İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşip yerleşmeme durumu ile işsizin öznitelikleri arasında doğrusal bir ilişki olup olmadığı

[28] **Pearson'un Kontenjans Katsayısı (Contingency Coefficient):** Kontenjans katsayısı, katsayısının IxJ boyutlu tablolardaki iki değişken arasındaki ilişkinin büyüklüğünü ölçen biçimidir [48].

test edilmiştir. Analiz sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin tabloda yer alan tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel olarak anlamlı doğrusal bir ilişki bulunmaktadır.

İlişki gücü incelendiğinde işsizin işe yerleşmesi ile doğrusal ilişkili tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. İşsizin başvuru sonrası işe yerleşmesi ile demografik, kişisel ve iş arama öznitelikleri arasında çok zayıf doğrusal ilişki bulunsa da bu ilişkilerin bir modelde birlikte değerlendirilmesi daha güçlü bir yapı ortaya koyabilir. Bu ilişki sonuçlarına göre işsizin tüm öznitelikleri modele girmeye aday konumundadır.

Tablo 3.21 İşe Yerleşme ile İşsizin Özellikleri Arasındaki İlişkiler(Yığın, Korelasyon)

Öznitelikler	P Değeri	İlişki Varlığı	İlişki Gücü	İlişki Derecesi
İkamet	< ,001	Var	0,043	Çok Zayıf
Sosyal Durum	< ,001	Var	-0,073	Çok Zayıf
Cinsiyet	< ,001	Var	-0,093	Çok Zayıf
Yaş1	< ,001	Var	-0,128	Çok Zayıf
Medeni Durum1	< ,001	Var	-0,162	Çok Zayıf
Öğrenim	0,021	Var	-0,019	Çok Zayıf
Meslek	< ,001	Var	0,089	Çok Zayıf
Başvuru Türü	< ,001	Var	0,069	Çok Zayıf
İşsizlik Ödeneği Alma Durumu	< ,001	Var	0,044	Çok Zayıf

Ki-kare analizi ve korelasyon analizi sonuçlarına göre işsizin işe yerleşmesi ile öznitelikleri arasında hem doğrusal hem de doğrusal olmayan ilişki varlığı, gücü ve derecesi arasında yüksek benzerlikler ortaya çıkmıştır. Bundan sonraki aşamada işsizin öznitelikleri arasındaki çoklu doğrusal bağlantı (yüksek ilişki(r>0,60)) kontrol edilerek yüksek ilişkili değişkenlerden biri modele dâhil edilmeyecektir.

Öznitelikler Arasındaki Çoklu Doğrusal Bağlantı Sorunu

Öznitelikler arasında 0,60 ve üzerinde değerde korelasyon olması çoklu doğrusal bağlantıyı işaret etmektedir. Bazı çalışmalarda çoklu doğrusal bağlantı oran 0,70 hatta 0,80 olarak ta kabul edilmektedir. Öznitelikler arasındaki korelasyon tablosu incelendiğinde öznitelikler arasında 0,60 ve daha yukarı oranda bir korelasyonun sadece yaş1 ile medeni durum1 özniteliği arasında görüldüğü tespit edilmiştir. Ancak hem korelasyonun çoklu doğrusal bağlantı sınır değerine yakın olması hem de daha

yüksek oranlar için çoklu doğrusal bağlantı olması gerekliliği yönündeki görüşler nedeniyle ihtiyatlı davranılarak söz konusu çoklu doğrusal bağlantı değerlendirmeye alınmayarak hem Yaş1 özniteliği hem de Medeni Durum1 özniteliğinin modelde yer alması uygun görülmüştür. Burada araştırmacının bilgi birikimi ve tecrübesiyle konuya hâkim olması ve özniteliklerden hangisinin modele alınıp hangisinin alınmayacağına karar vermesi büyük önem arz etmektedir.

Tablo 3.22 Öznitelikler Arasındaki Korelasyon(Yığın)

	Cinsiyet	**Sosyal Durum**	**Yaş1**	**Öğrenim**	**Başvuru Türü**	**Meslek**	**Medeni Durum1**	**İkamet**	**İşsizlik Ödeneği Alma Durumu**
Cinsiyet	—								
Sosyal Durum	0,108	—							
Yaş1	0,041	-0,14	—						
Öğrenim	0,115	0,074	-0,258	—					
Başvuru Türü	-0,193	-0,058	0,27	0,051	—				
Meslek	-0,132	0,056	0,049	-0,093	0,046	—			
Medeni Durum1	0,174	-0,027	0,648	-0,324	0,114	0,029	—		
İkamet	0,058	0,021	0,021	-0,007	0,036	-0,001	0,024	—	
İşsizlik Ödeneği Alma Durumu	-0,1	0,037	0,231	-0,11	0,075	0,098	0,157	0,035	—

Ki-kare analizi, korelasyon analizi ve çoklu doğrusal bağlantı sorunu birlikte değerlendirildiğinde işsizin sosyal yardım alma durumu özniteliği dışındaki tüm özniteliklerinin modele alınması uygun görülmüştür. Analiz sonuçlarına göre yığın veri seti ile örneklem veri seti arasında çok benzer sonuçlar çıkmıştır. Bu çok yüksek benzerlik yığından seçilen %20 oranındaki örneklemin yığını çok iyi temsil ettiğini göstermektedir. Bu iyi temsil ise yığın ve örneklemden elde edilecek modellemelerin, sınıflandırmaların ve performans değerlerinin de yüksek benzerlikler göstereceğini işaret etmektedir.

Lojistik Regresyon Modeli İle Özniteliklerin Kesin Olarak Belirlenmesi

İlişki analizleri sonucunda modele alınacak öznitelikler seçilmiştir. Sırada ise lojistik regresyon modellemesi ana modelin oluşturulması ve özniteliklerin ikinci derecede boyutlandırılması bulunmaktadır. Burada gerçekleştirilen çok değişkenli lojistik regresyon modellemesi istatistiksel analiz amaçlı olmayıp özniteliklerin seçimi/çıkarımı amaçlıdır. Dolayısıyla çok değişkenli lojistik regresyon modellemesinin tüm aşamalarına yer verilmeyecektir. Hâlihazırda bu aşamalar örneklem verisinin analizi ve boyutlandırılmasında gerçekleştirilmiştir.

Sabit değer ve ilişki analizlerine göre modele alınması uygun görülen özniteliklerin yer aldığı ana model oluşturulmuştur. Oluşturulan modele göre cinsiyet dışındaki tüm öznitelikler için p değeri 0,05'ten küçük olduğu için %95 güven düzeyinde istatistiksel olarak anlamlıdır. Bu bilgilere göre cinsiyet özniteliği modelden çıkarılmaya aday görülmektedir.

Tablo 3.23 Ana Lojistik Regresyon Modeli(Yığın)

	Katsayı	**Ölç. Hata**	**Z**	**P-değeri**
Const(Sabit Değer)	2,241	0,254	8,814	<0,0001
İkamet	0,233	0,043	5,461	<0,0001
Sosyal Durum	-1,191	0,116	-10,3	<0,0001
Cinsiyet	-0,037	0,038	-0,974	0,33
Yaş1	-0,143	0,014	-10,51	<0,0001
Medeni Durum2	-0,21	0,018	-11,87	<0,0001
Öğrenim	-0,152	0,018	-8,508	<0,0001
Meslek	0,11	0,011	10,1	<0,0001
Başvuru Türü	0,449	0,038	11,84	<0,0001
İşsizlik Ödeneği Alma Durumu	0,693	0,085	8,153	<0,0001

Oluşturulan ana modeldeki istatistiksel olarak anlamsız özniteliklerin çıkartılarak modelin daha olgunlaşması gerekmektedir. Bu öznitelik çıkartma işlemi ile veri setinde boyut düşürme/boyutlandırma işlemi de gerçekleştirilmiş olacaktır. Ana modelde yer almasına rağmen istatistiksel anlamlığa sahip olmayan öznitelikleri modelden çıkarmak için geriye doğru çıkarma yöntemi yapılarak öznitelik çıkarımı işlemi gerçekleştirilmiştir. Öznitelik çıkarma işleminde %95 güven düzeyi tercih edilmiş ve modelden sadece cinsiyet çıkartılmış ve amaçlanan modele ulaşılmıştır. Oluşturulan amaçlanan modeldeki tüm öznitelikler %95 güven düzeyinde istatistiksel olarak anlamlıdır ve işsizin başvuru sonrası işe yerleşme durumunu açıklamaktadır.

Tablo 3.24 Amaçlanan Lojistik Regresyon Modeli(Yığın)

	Katsayı	Ölç. Hata	Z	P-değeri
Const(Sabit Değer)	2,215	0,253	8,757	<0,0001
İkamet	0,230	0,043	5,411	<0,0001
Sosyal Durum	-1,203	0,115	-10,460	<0,0001
Yaş1	-0,143	0,014	- 10,530	<0,0001
Medeni Durum2	-0,214	0,017	-12,350	<0,0001
Öğrenim	-0,155	0,018	-8,859	<0,0001
Meslek	0,111	0,011	10,290	<0,0001
Başvuru Türü	0,458	0,037	12,370	<0,0001
İşsizlik Ödeneği Alma Durumu	0,703	0,084	8,320	<0,0001

Oluşturulan amaçlanan model McFadden R-kare(Açıklama Katsayısı)'ye göre modelde yer alan öznitelikler model hedef değişkeni olan işsizin işe yerleşme durumunu yaklaşık %5 oranında açıklamakta olup doğruluk oranı da %59,7'dir. Aynı zamanda model ve model parametreleri istatistiksel olarak anlamalıdır. Hem açıklama oranı hem de doğru sınıflama oranı hem de modelde yer alan öznitelikler yönünden yığın ve örneklem veri seti çok yüksek benzerlikler arz etmektedir. Bu çok yüksek benzerlikler ile örneklemin yığını iyi temsil etmesi yığın veri seti kaynaklı maliyetlerin yüksek olduğu durumlarda örneklem veri setinin kullanılması gerekliliğini net bir şekilde ortaya koymaktadır.

İlişki analizleri ile başlayıp amaçlanan lojistik regresyon modellemesi ile son bulan öznitelik seçme/çıkarma işleminde sosyal yardım alma durumu ve cinsiyet öznitelikleri modelden çıkartılmış ve sekiz özniteliğin olduğu bir yapıya indirgenmiştir. Bundan sonraki aşamalarda indirgenmemiş ve indirgenmiş yığın ve örneklem veri seti üzerinden makine öğrenmesi algoritmaları ile modeller oluşturmak ve bu modellerin performans değerlerini ortaya koymak ve gerekli performans iyileştirmeleri ile en iyi modele ulaşmak işlemleri gerçekleştirilecektir.

Yığın Veri Seti Kapsamında Gerçekleştirilen Makine Öğrenmesi İşlemleri

Bu konu başlığı altında hem indirgenmemiş hem de indirgenmiş yığın veri seti, hold out model geçerleme yöntemine(%70 eğitim-%30 test ve %80 eğitim-%20 test) ve k çapraz doğrulama yöntemine(5-kat çapraz doğrulama ve 10 kat çapraz doğrulama) göre denetimli öğrenme algoritmaları ile modellenmiş ve geliştirilen modeller veri seti, algoritma ve model geçerlemeye/doğrulama türlerine göre performans değerleri

karşılaştırılmıştır. Makine öğrenmesi algoritmalarında uygulanan parametreler ilişkin detaylı bilgi aşağıda yer almaktadır.

- K-en yakın komşu algoritması için k(n_neighbors)=1 ve uzaklık ölçütü olarak "minkowski" uzaklığı tercih edilmiştir.
- Rasgele orman algoritması için ağaç sayısı parametresi(n_estimators)=10 ve bölünme kriteri olarak(criterion=entropy) parametresi seçilmiştir.
- Destek vektör makineleri için çekirdek parametresi "linear" olarak seçilmiştir.
- Naive Bayes algoritması için GaussianNB parametresi seçilmiştir.
- Lojistik regresyon algoritması için varsayılan parametre seçilmiştir.
- Karar ağaçları algoritması için bölünme kriteri olarak(criterion=gini) parametresi seçilmiştir.

Model performans göstergeleri tez çalışması kapsamında aşağıda yer verilen şekilde yorumlanmıştır.

Doğruluk (Accuracy): İŞKUR'a başvuru sonrası bir yıl süre içerisinde işe yerleşmiş/yerleşmemiş olarak yapılan doğru tahmin oranın ifade etmektedir. Modelin doğruluk performans göstergeleri model için çok büyük bir anlam ifade etmekle beraber tek başına yeterli değildir. Dolayısıyla model performansı kesinlik ve duyarlılık gibi diğer göstergelerle birlikte değerlendirilmiştir.

Kesinlik (Precision): İŞKUR'a başvuru sonrası bir yıl süre içerisinde işe yerleşmiş olarak tahmin edilenlerin içerisindeki gerçekten işe yerleşmiş olanların oranını ifade etmektedir.

Duyarlılık (Recall): İŞKUR'a başvuru sonrası bir yıl süre içerisinde gerçekte işe yerleşmiş olanların doğru tahmin edilme oranını ifade etmektedir.

F1-Skor: Kesinlik ve duyarlılığın harmonik ortalamasıdır. Veri setinin dengesiz olduğu durumlarda etkilidir. Çalışmada veri dengeleme işlemi yapıldığı için diğer performans göstergeleri ile birlikte değerlendirilmiştir.

İndirgenmemiş Yığın Veri Seti İçin Hold Out-%70 Eğitim-%30 Test Model Geçerlemeye Göre Modellerin Performans Değerleri

İndirgenmemiş yığın veri setinde 14.146 satır veri ve 9 öznitelik modele alınmıştır. Veri setindeki verilerin 9.902'si eğitim ve 4.244'ü test verisi olarak kullanılmıştır.

İndirgenmemiş Yığın Veri Seti için Hold Out Model geçerleme yöntemi ve %70 eğitim ve %30 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.25 İndirgenmemiş Yığın Veri Seti İçin Hold Out-%70 Eğitim-%30 Test Modellerin Performans Değerleri

	Doğruluk (Accuracy)	**Kesinlik (Precision)**	**Duyarlılık (Recall)**	**F1 Skor**
K En Yakın Komşu	0,724	0,701	0,767	0,733
Lojistik Regresyon	0,606	0,595	0,626	0,610
Naive Bayes	0,583	0,639	0,354	0,455
Karar Ağaçları	0,777	0,738	0,848	0,789
Rasgele Orman	0,784	0,745	0,853	0,795
Destek Vektör Makinaları	0,602	0,571	0,772	0,656
Yapay Sinir Ağları	0,676	0,642	0,773	0,702

Doğruluk oranı en yüksek olan modeller, rasgele orman, karar ağaçları ve k en yakın komşu algoritmaları ile en düşük olan modeller ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Kesinlik oranı en yüksek olan modeller rasgele orman, karar ağaçları ve k en yakın komşu algoritmaları ile en düşük olan modeller naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Duyarlılık oranı en yüksek olan modeller, rasgele orman ve karar ağaçları algoritmaları ile en düşük olan modeller ise naive bayes ve lojistik regresyon algoritmaları ile oluşturulmuştur. İyi performansı rasgele orman, karar ağaçları, k en yakın komşu algoritmaları göstermiş olup performans düzeyi yüksek seviyededir. Kötü performans gösteren algoritmalar ise lojistik regresyon, naive bayes, destek vektör makinaları, yapay sinir ağları algoritması olup orta seviyede performans göstermişlerdir.

Rasgele orman algoritması %78,4 doğruluk oranı ile en yüksek performansı gösterirken, naive bayes algoritması ise %58,3 doğruluk oranı ile en düşük performansı göstermiştir. Genel olarak karar ağaçları ve rasgele orman algoritması diğer algoritmalardan yüksek performans ile belirgin olarak ayrılmaktadır. Lojistik regresyon ve naive bayes gibi istatistiksel algoritmalar diğer makine öğrenmesi algoritmalarına göre belirgin olarak daha düşük bir performans ortaya koymuştur.

İndirgenmemiş Yığın Veri Seti İçin Hold Out-%80 Eğitim-%20 Test Model Geçerlemeye Göre Modellerin Performans Değerleri

İndirgenmemiş yığın veri setinde 14.146 satır veri ve 9 öznitelik modele alınmıştır. Veri setindeki verilerin 11.316'si eğitim ve 2.830'u test verisi olarak kullanılmıştır. İndirgenmemiş yığın Veri Seti için Hold Out Model geçerleme yöntemi ve %80 eğitim ve %20 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.26 İndirgenmemiş Yığın Veri Seti İçin Hold Out-%80 Eğitim-%20 Test Modellerin Performans Değerleri

	Doğruluk (Accuracy)	**Kesinlik (Precision)**	**Duyarlılık (Recall)**	**F1 Skor**
K En Yakın Komşu	0,731	0,728	0,731	0,729
Lojistik Regresyon	0,607	0,600	0,626	0,613
Naive Bayes	0,584	0,644	0,361	0,463
Karar Ağaçları	0,781	0,745	0,848	0,793
Rasgele Orman	0,788	0,752	0,855	0,800
Destek Vektör Makinaları	0,593	0,567	0,761	0,650
Yapay Sinir Ağları	0,686	0,658	0,763	0,706

Doğruluk oranı en yüksek olan modeller, rasgele orman, karar ağaçları ve k en yakın komşu algoritmaları ile en düşük olan modeller ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Kesinlik oranı en yüksek olan modeller, rasgele orman, karar ağaçları ve k en yakın komşu algoritmaları ile en düşük olan modeller ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Duyarlılık oranı en yüksek olan modeller, rasgele orman ve karar ağaçları, yapay sinir ağları ve destek vektör makinaları algoritmaları ile en düşük olan modeller ise naive bayes ve lojistik regresyon algoritmaları ile oluşturulmuştur. İyi performansı rasgele orman, karar ağaçları, k en yakın komşu algoritmaları göstermiş olup performans düzeyi yüksek seviyededir. Kötü performans gösteren algoritmalar ise lojistik regresyon, naive bayes ve destek vektör makinaları, yapay sinir ağları algoritması olup orta seviyede performans göstermişlerdir. Rasgele orman algoritması %78,8 doğruluk oranı ile en yüksek performansı gösterirken, naive bayes algoritması ise %58,4 doğruluk oranı ile en düşük performansı göstermiştir. Genel olarak karar ağaçları ve rasgele orman algoritması diğer algoritmalardan yüksek performans ile belirgin olarak ayrılmaktadır. Lojistik regresyon ve naive bayes gibi istatistiksel algoritmalar diğer makine

öğrenmesi algoritmalarına göre belirgin olarak daha düşük bir performans ortaya koymuştur.

Eğitim ve test verisi oranlarındaki değişim model performanslarında %0,3-%3,7 aralığında bir değişime neden olmuştur. Bu değişim aralığı oldukça kısıtlı olup eğitim ve test verisi oranlarındaki değişimin model performansında belirgin bir artış ve azalışa neden olmadığı sonucuna varılmıştır.

İndirgenmemiş Yığın Veri Seti İçin 5-Kat Çapraz Doğrulamaya Göre Modellerin Performans Değerleri

İndirgenmemiş yığın veri setinde 14.146 satır veri ve 9 öznitelik modele alınmıştır. Yığın Veri Seti için 5-kat çapraz doğrulama yöntemi ile geliştirilen modellerin her bir kat için doğruluk performansı ve tüm katların ortalama doğruluk performans değerleri aşağıda yer verilmiştir.

Tablo 3.27 İndirgenmemiş Yığın Veri Seti İçin 5-Kat Çapraz Doğrulama Modellerin Doğruluk Performans Değerleri

	k-1	k-2	k-3	k-4	k-5	k-O
K En Yakın Komşu	0,712	0,741	0,736	0,727	0,713	0,726
Lojistik Regresyon	0,578	0,589	0,595	0,604	0,607	0,595
Naive Bayes	0,579	0,583	0,580	0,585	0,582	0,582
Karar Ağaçları	0,780	0,797	0,772	0,789	0,772	0,782
Rasgele Orman	0,782	0,791	0,779	0,795	0,782	0,786
Destek Vektör Makinaları	0,567	0,597	0,566	0,594	0,579	0,581
Yapay Sinir Ağları	0,681	0,671	0,661	0,677	0,692	0,676

Ortalama doğruluk oranı en yüksek modeller rasgele orman ve karar ağaçları algoritması ile oluşturulmuştur. Yüksek doğruluk oranında bu iki algoritmayı k en yakın komşu ve yapay sinir ağları algoritması takip etmektedir.

Ortalama doğruluk oranı en düşük modeller ise destek vektör makinaları, naive bayes ve lojistik regresyon algoritmaları ile oluşturulmuştur. En yüksek doğruluk performansı %79,7 oran ile 2-kat doğru çaprazlama ile karar ağaçları algoritmasında gerçekleşmiştir. En düşük doğruluk performansı ise %56,6 oran ile 3-kat doğru çaprazlama ile destek vektörleri algoritmasında gerçekleşmiştir.

İndirgenmemiş Yığın Veri Seti İçin 10-Kat Çapraz Doğrulamaya Göre Modellerin Performans Değerleri

İndirgenmemiş Yığın veri setinde 14.146 satır veri ve 9 öznitelik modele alınmıştır. Yığın Veri Seti için 10-kat çapraz doğrulama yöntemi ile geliştirilen modellerin her bir kat için doğruluk performansı ve tüm katların ortalama doğruluk performans değerleri aşağıda yer verilmiştir.

Tablo 3.28 İndirgenmemiş Yığın Veri Seti 10-Kat Çapraz Doğrulama Modellerin Doğruluk Performans Değerleri

	k-1	k-2	k-3	k-4	k-5	k-6	k-7	k-8	k-9	k-10	k-O
K En Yakın Komşu	0,731	0,714	0,736	0,738	0,729	0,738	0,736	0,732	0,699	0,719	0,727
Lojistik Regresyon	0,574	0,584	0,580	0,595	0,612	0,577	0,616	0,596	0,607	0,601	0,594
Naive Bayes	0,588	0,570	0,584	0,577	0,596	0,565	0,589	0,581	0,577	0,586	0,581
Karar Ağaçları	0,779	0,777	0,791	0,802	0,765	0,772	0,796	0,778	0,784	0,782	0,782
Rasgele Orman	0,778	0,782	0,789	0,800	0,770	0,781	0,801	0,779	0,789	0,788	0,786
Destek Vektör Makinaları	0,560	0,572	0,577	0,601	0,567	0,575	0,609	0,580	0,583	0,575	0,580
Yapay Sinir Ağları	0,676	0,651	0,669	0,673	0,657	0,678	0,706	0,644	0,691	0,661	0,671

Ortalama doğruluk oranı en yüksek modeller rasgele orman ve karar ağaçları algoritması ile oluşturulmuştur. Yüksek doğruluk oranında bu iki algoritmayı k en yakın komşu ve yapay sinir ağları algoritması takip etmektedir. Ortalama doğruluk oranı en düşük modeller ise destek vektör makinaları, naive bayes ve lojistik regresyon algoritmaları ile oluşturulmuştur. En yüksek doğruluk performansı %80,2 oran ile 4-kat doğru çaprazlama ile karar ağaçları algoritmasında gerçekleşmiştir. En düşük doğruluk performansı ise %56 oran ile 1-kat doğru çaprazlama ile destek vektörleri algoritmasında gerçekleşmiştir.

Çapraz doğrulama sayısının 5-kat ve 10-kat olmasına göre ortalama doğrulama oranları birlikte incelendiğinde %1'in altında çok düşük farkların olduğu tespit edilmiştir. Bu durum çapraz doğrulamadaki kat farklılığının model performansında belirgin bir artış ve azalışa neden olmadığı sonucunu ortaya koymaktadır.

İndirgenmemiş Yığın Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

İndirgenmemiş yığın veri seti hold out geçerleme yöntemine göre %70 eğitim-%30 test ve %80 eğitim-%20 test şeklinde ve k-kat çapraz doğrulama yöntemine göre de 5-kat ve 10-kat çapraz doğrulamaya tabi tutulmuş ve parametreleri aynı olan makine öğrenmesi algoritmaları ile modeller geliştirilerek modellerin performans değerleri doğruluk, kesinlik, duyarlılık ve f-ölçütü performans göstergeleri üzerinden tespit edilmiştir. Modellerin sadece doğruluk/ortalama doğruluk ölçütü üzerinden performans karşılaştırmaları yapılmıştır.

Tablo 3.29 İndirgenmemiş Yığın Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,724	0,731	0,726	0,727
Lojistik Regresyon	0,606	0,607	0,595	0,594
Naive Bayes	0,583	0,584	0,582	0,581
Karar Ağaçları	0,777	0,781	0,782	0,782
Rasgele Orman	0,784	0,788	0,786	0,786
Destek Vektör Mak.	0,602	0,593	0,581	0,580
Yapay Sinir Ağları	0,676	0,686	0,676	0,671

Karşılaştırmalı performans değerlerine göre en yüksek doğruluk oranı %78,8 oran ile rasgele orman algoritması ve %80 eğitim-%20 test model geçerleme yönteminde gerçekleşmiştir. İkinci sırada en yüksek doğruluk performansı %78,6 oran ile yine rasgele orman algoritmasında 5-kat ve 10-kat çapraz doğrulama yönteminde gerçekleşmiştir. En yüksek doğruluk oranında rasgele orman algoritmasını tüm geçerleme/doğrulama yöntemlerinde karar ağaçları algoritması ve k en yakın komşu algoritması takip etmektedir. En düşük doğruluk performansı %58 oran ile destek vektör makinaları algoritmasında 10-kat çapraz doğrulama yönteminde gerçekleşmiştir. İkinci sırada en düşük doğruluk performansı %58,1 oran ile yine destek vektör makinaları algoritmasında 5-kat çapraz doğrulama yönteminde ve naive bayes algoritmasında 10-kat çapraz doğrulama yönteminde gerçekleşmiştir. Destek vektör makinaları, naive bayes ve lojistik regresyon algoritmaları tüm geçerleme/doğrulama yöntemlerinde en düşük doğruluk performansı göstermiştir. Yapay sinir ağları algoritması ise tüm geçerleme/doğrulama yöntemlerinde orta

düzeyde doğruluk performansı göstererek en yüksek ve en düşük doğruluk performansı gösteren algoritmaların ortasında yer almaktadırlar.

Makine öğrenmesi algoritmalarının farklı tür ve şekildeki doğruluk performansı incelendiğinde doğruluk performansının belirleyici olarak model geçerleme/doğrulama yöntemlerinin etkisinin az olduğu esas belirleyicinin algoritmalar olduğu göze çarpmaktadır. Başka bir ifadeyle makine öğrenmesi algoritması sabit iken model geçerleme yöntemleri ve tiplerinde meydana gelen değişimler modelin doğruluk performansını çok az etkilemektedir.

İndirgenmiş Yığın Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Performans Değerleri

Buraya kadar indirgenmemiş yığın veri seti için modeller oluşturulmuş ve performans hesapları yapılmıştır. Bu bölümde ise daha önce istatistiksel metotlar gerçekleştirilen öznitelik seçme ve çıkarma işlemine göre indirgenmiş yığın veri seti için modeller oluşturuldu, performans hesabı ve değerlendirmesi yapılmıştır. İndirgenmiş yığın veri setinde indirgenmemiş yığın veri setine göre sadece cinsiyet özniteliği bulunmayıp diğer öznitelikler bulunmaktadır. İndirgenmiş yığın veri setinde 14.146 satır veri ve 8 öznitelik modele alınmıştır. İndirgenmiş yığın Veri Seti için Hold Out Model geçerleme yöntemi ve %70 eğitim-%30 test ve %80 eğitim-%20 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.30 İndirgenmiş Yığın Veri Seti İçin Hold Out Model Geçerleme Yöntemine Göre Modellerin Performans Değerleri ve Karşılaştırmaları

	%70 Eğitim-%30 Test				**%80 Eğitim-%20 Test**			
	Doğruluk (Accuracy)	**Kesinlik (Precision)**	**Duyarlılık (Recall)**	**F1 Skor**	**Doğruluk (Accuracy)**	**Kesinlik (Precision)**	**Duyarlılık (Recall)**	**F1 Skor**
K En Yakın Komşu	0,685	0,673	0,701	0,687	0,722	0,688	0,803	0,741
Lojistik Regresyon	0,607	0,596	0,628	0,612	0,602	0,596	0,616	0,606
Naive Bayes	0,574	0,648	0,293	0,403	0,574	0,651	0,303	0,414
Karar Ağaçları	0,750	0,713	0,824	0,765	0,753	0,722	0,816	0,766
Rasgele Orman	0,756	0,723	0,82	0,768	0,755	0,726	0,814	0,767
Destek Vektör Mak.	0,599	0,569	0,767	0,653	0,589	0,563	0,759	0,647
Yapay Sinir Ağları	0,657	0,628	0,744	0,681	0,675	0,666	0,691	0,678

İndirgenmiş yığın veri setine ilişkin hold out model geçerleme yöntemine göre hazırlanan modelde en yüksek performans değerleri rasgele orman, karar ağaçları algoritmaları göstermiştir. En düşük performans değerlerini ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları göstermiştir. K en yakın komşu ve yapay sinir ağları diğer algoritmalara göre ortalama bir performans göstermiştir.

Eğitim ve test verisi oranlarındaki değişim algoritmaların performans göstergelerinin çok az bir değişime uğramasına neden olmuştur. Dolayısıyla indirgenmiş veri setinde de model performansını model geçerleme yöntemi belirlemeyip algoritmalar belirlemiştir.

Tablo 3.31 İndirgenmiş Yığın Veri Seti İçin K-Kat Çapraz Doğrulama Yöntemine Göre Modellerin Doğruluk Performans Değerleri ve Karşılaştırmaları

	5-Kat Çapraz Doğrulama	**10-Kat Çapraz Doğrulama**
	Ortalama Doğruluk	Ortalama Doğruluk
K En Yakın Komşu	0,700	0,699
Lojistik Regresyon	0,594	0,595
Naive Bayes	0,568	0,568
Karar Ağaçları	0,757	0,755
Rasgele Orman	0,760	0,759
Destek Vektör Mak.	0,579	0,578
Yapay Sinir Ağları	0,670	0,672

İndirgenmiş yığın veri setine ilişkin k-kat çapraz doğrulama yöntemine göre hazırlanan modelde en yüksek doğruluk performansı değerleri rasgele orman, karar ağaçları algoritmaları göstermiştir.

En düşük doğruluk performans değerlerini ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları göstermiştir. K en yakın komşu ve yapay sinir ağları diğer algoritmalara göre ortalama bir performans göstermiştir.

Çapraz doğrulama kat sayısındaki farklılık algoritmaların performans göstergelerinin çok az bir değişime uğramasına neden olmuştur. Dolayısıyla indirgenmiş veri setinde de model doğruluk performansını algoritmalar belirlemiştir.

Tablo 3.32 İndirgenmiş Yığın Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,685	0,722	0,700	0,699
Lojistik Regresyon	0,607	0,602	0,594	0,595
Naive Bayes	0,574	0,574	0,568	0,568
Karar Ağaçları	0,750	0,753	0,757	0,755
Rasgele Orman	0,756	0,755	0,760	0,759
Destek Vektör Mak.	0,599	0,589	0,579	0,578
Yapay Sinir Ağları	0,657	0,675	0,670	0,672

İndirgenmiş yığın veri seti için karşılaştırmalı performans değerlerine göre en yüksek doğruluk oranı %76 oran ile rasgele orman algoritmasında 5-kat çapraz doğrulama yönteminde gerçekleşmiştir. İkinci sırada en yüksek doğruluk performansı %75,9 oran ile yine rasgele orman algoritmasında 10-kat çapraz doğrulama yönteminde gerçekleşmiştir. En yüksek doğruluk performansında rasgele orman algoritmasını tüm geçerleme/doğrulama yöntemlerinde karar ağaçları algoritması ve k en yakın komşu algoritması takip etmektedir. En düşük doğruluk performansı %56,8 oran ile naive bayes algoritmasında 10-kat çapraz doğrulama yönteminde gerçekleşmiştir. İkinci, üçüncü ve dördüncü sırada en düşük doğruluk performansı yine naive bayes algoritmasında diğer model geçerleme/doğrulama yöntemlerinde gerçekleşmiştir. Naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları tüm geçerleme/doğrulama yöntemlerinde en düşük doğruluk performansı göstermiştir. Yapay sinir ağları algoritması ile k en yakın komşu algoritması ise tüm geçerleme/doğrulama yöntemlerinde orta düzeyde doğruluk performansı göstererek en yüksek ve en düşük doğruluk performansı gösteren algoritmaların ortasında yer almaktadırlar.

Makine öğrenmesi algoritmalarının farklı tür ve şekildeki doğruluk performansı incelendiğinde doğruluk performansının belirleyicisi olarak model geçerleme/doğrulama yöntemlerinin etkisinin az olduğu esas belirleyicinin algoritmalar olduğu göze çarpmaktadır. Diğer bir ifadeyle model algoritması sabit iken model geçerleme/doğrulama yöntemlerindeki değişimler model performanslarını yok denecek kadar az etkilemektedir.

İndirgenmiş Yığın Veri Seti ve İndirgenmemiş Yığın Veri Seti İle Oluşturulan Modeller Arasındaki Doğruluk Performansı Karşılaştırması

Tablo 3.33 İndirgenmiş Yığın Veri Seti ve İndirgenmemiş Yığın Veri Seti İle Oluşturulan Modeller Arasındaki Doğruluk Performansı Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,039	0,009	0,026	0,028
Lojistik Regresyon	-0,001	0,005	0,001	-0,001
Naive Bayes	0,009	0,01	0,014	0,013
Karar Ağaçları	0,027	0,028	0,025	0,027
Rasgele Orman	0,028	0,033	0,026	0,027
Destek Vektör Mak.	0,003	0,004	0,002	0,002
Yapay Sinir Ağları	0,019	0,011	0,006	-0,001

İndirgenmemiş ve indirgenmiş yığın veri setlerine göre oluşturulan modellerdeki doğruluk oranları arasında farklı algoritma ve model geçerleme/doğrulama yöntemlerine fark etmeksizin çok düşük bir fark olduğu anlaşılmaktadır. İki veri setine göre en yüksek doğruluk performansı farklılığı k en yakın komşu, karar ağaçları ve rasgele orman algoritmalarında olmuştur. Bu farklılık tüm model geçerleme/doğrulama yöntemleri için ortalama %3 civarındadır. Aynı zamanda bu algoritmalar en yüksek doğruluk performansı gösteren algoritmalardır. İki veri setine göre en düşük doğruluk performansı farklılığı destek vektör makinaları, lojistik regresyon ve naive bayes algoritmalarında olmuştur. Bu farklılık tüm model geçerleme/doğrulama yöntemleri için ortalama %1 civarındadır. Aynı zamanda bu algoritmalar en düşük doğruluk performansı gösteren algoritmalardır. İndirgenmiş veri seti indirgenmemiş veri setinden sadece cinsiyet özniteliği yönünden farklılık arz etmektedir. İndirgenmiş model hem daha düşük doğruluk performansı göstermiş hem de bir işsizin iş arama davranışında olmazsa olmaz özniteliği olan cinsiyetini model dışı bırakmıştır. Bu durum işsizlerin risk sınıflaması ve iş profillemesi için kabul edilemezdir. Dolayısıyla indirgenmemiş model indirgenmiş modele göre hem doğruluk performansı hem de öznitelik zenginliği yönünden üstünlük arz etmektedir.

Öznitelik çıkarma işlemi yüksek doğruluk performansı gösteren algoritmalarda daha yüksek oranda doğruluk performansı düşüşü gerçekleştirirken, düşük performans gösteren algoritmalarda daha düşük oranda doğruluk performansı düşüşüne neden olmaktadır. Öznitelik çıkarma işlemi ile gerçekleştirilen boyut azaltmada çıkarılan

öznitelik sayısı arttıkça doğruluk oranında azalma beklenmektedir. Dolayısıyla boyut azaltma işleminin her veri setine uygulanmayacağı değerlendirilmektedir. Özellikle tez çalışmasında olduğu gibi öznitelik sayısının az ve yetersiz olduğu veri setinde öznitelik çıkarma işlemi doğruluk performansını düşürdüğü gibi veriye dayalı bilgi kaybına da neden olmaktadır.

Örneklem Veri Seti Kapsamında Gerçekleştirilen Makine Öğrenmesi İşlemleri

Örneklem veri seti yığın veri setinden %20 oranda ve istatistiksel metotlarla elde edilmiş bir veri setidir. Bu konu başlığı altında hem indirgenmemiş hem de indirgenmiş örneklem veri seti, hold out model geçerleme yöntemine(%70 eğitim-%30 test ve %80 eğitim-%20 test) ve k çapraz doğrulama yöntemine(5-kat çapraz doğrulama ve 10 kat çapraz doğrulama) göre denetimli öğrenme algoritmaları ile modellenmiş ve geliştirilen modeller veri seti, algoritma ve model geçerlemeye/doğrulama türlerine göre performans değerleri karşılaştırılmıştır. Modelleme gerçekleştirilirken yığın veri setine uygulanan işlemlerin aynısı uygulanarak yığın ve örneklem veri seti arasındaki performans değerleri karşılaştırılmıştır.

İndirgenmemiş Örneklem Veri Seti İçin Hold Out-%70 Eğitim-%30 Test Model Geçerlemeye Göre Modellerin Performans Değerleri

İndirgenmemiş örneklem veri setinde 2.830 satır veri ve 9 öznitelik modele alınmıştır. Veri setindeki verilerin 1.981'i eğitim ve 849'u test verisi olarak kullanılmıştır. Örneklem Veri Seti için Hold Out Model geçerleme yöntemi ve %70 eğitim ve %30 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.34 İndirgenmemiş Örneklem Veri Seti için Hold Out-%70-%30 Modellerin Performans Değerleri

Sınıflama Algoritmaları	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skor
K En Yakın Komşu	0,665	0,653	0,726	0,687
Lojistik Regresyon	0,605	0,608	0,623	0,615
Naive Bayes	0,562	0,663	0,274	0,388
Karar Ağaçları	0,658	0,656	0,686	0,670
Rasgele Orman	0,693	0,669	0,779	0,720
Destek Vektör Makinaları	0,600	0,579	0,767	0,660
Yapay Sinir Ağları	0,642	0,633	0,698	0,664

Doğruluk oranı en yüksek olan modeller, rasgele orman, k en yakın komşu ve karar ağaçları algoritmaları ile en düşük olan modeller ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Kesinlik oranı en yüksek olan modeller rasgele orman, karar ağaçları, naive bayes ve k en yakın komşu algoritmaları ile en düşük olan modeller destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Duyarlılık oranı en yüksek olan modeller, rasgele orman, destek vektör makinaları ve karar ağaçları algoritmaları ile en düşük olan modeller ise naive bayes ve lojistik regresyon algoritmaları ile oluşturulmuştur. Rasgele orman, karar ağaçları, k en yakın komşu yükseğe yakın orta seviyede performans göstermişlerdir. Lojistik regresyon ve naive bayes, algoritmalar orta seviyede ancak diğer algoritmalara göre daha düşük performans göstermişlerdir. Rasgele orman algoritması %69,3 oran ile en iyi doğruluk performansı gösterirken, naive bayes algoritması ise %56,2 oranı ile en kötü doğruluk performansı göstermiştir.

Naive bayes, lojistik regresyon ve destek vektör makinaları algoritması düşük performans ile belirgin olarak ayrılmaktadır. Diğer algoritmaların performans değerleri birbirinden yüksek farklılıklar arz etmeyerek yakın değerlerdedir. Ancak rasgele orman algoritması iyi performans ile göze çarpmaktadır.

İndirgenmemiş Örneklem Veri Seti İçin Hold Out-%80 Eğitim-%20 Test Model Geçerlemeye Göre Modellerin Performans Değerleri

İndirgenmemiş örnekleme veri setinde 2.830 satır veri ve 9 öznitelik modele alınmıştır. Veri setindeki verilerin 2.264'i eğitim ve 566'ı test verisi olarak kullanılmıştır. Örneklem Veri Seti için Hold Out Model geçerleme yöntemi ve %80 eğitim ve %20 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.35 İndirgenmemiş Örneklem Veri Seti İçin Hold Out-%80-%20 Modellerin Performans Değerleri

Sınıflama Algoritmaları	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skor
K En Yakın Komşu	0,647	0,649	0,693	0,670
Lojistik Regresyon	0,599	0,617	0,594	0,605
Naive Bayes	0,602	0,715	0,386	0,501
Karar Ağaçları	0,724	0,728	0,747	0,737
Rasgele Orman	0,726	0,709	0,799	0,751
Destek Vektör Makinaları	0,624	0,609	0,765	0,678
Yapay Sinir Ağları	0,652	0,689	0,597	0,640

Doğruluk oranı en yüksek olan modeller, rasgele orman ve karar ağaçları algoritmaları ile en düşük olan modeller ise lojistik regresyon ve naive bayes algoritmaları ile oluşturulmuştur. Kesinlik oranı en yüksek olan modeller karar ağaçları, naive bayes ve rasgele orman, algoritmaları ile en düşük olan modeller destek vektör makinaları ve lojistik regresyon algoritmaları ile oluşturulmuştur. Duyarlılık oranı en yüksek olan modeller, rasgele orman, destek vektör makinaları ve karar ağaçları algoritmaları ile en düşük olan modeller ise naive bayes, lojistik regresyon ve yapay sinir ağları algoritmaları ile oluşturulmuştur. Rasgele orman, karar ağaçları, k en yakın komşu algoritmaları tabloda yer alan tüm performans göstergelerinde yüksek seviyede performans göstermişlerdir. Rasgele orman algoritması %72,6 oran ile en iyi doğruluk performansı gösterirken, lojistik regresyon algoritması ise %59,9 oranı ile en kötü doğruluk performansı göstermiştir. Rasgele orman ve karar ağaçları iyi performans ile naive bayes ve lojistik regresyon algoritması kötü performans ile belirgin olarak diğer algoritmalardan belirgin ayrılmaktadır. Eğitim ve test verisi oranlarındaki değişim model doğruluk performanslarında %0,6-%4,1 aralığında bir değişime neden olmuştur. Bu değişim aralığı oldukça kısıtlı olup eğitim ve test verisi oranlarındaki değişimin model doğruluk performansında belirgin bir artış ve azalışa neden olmadığı sonucuna varılmıştır.

İndirgenmemiş Örneklem Veri Seti 5-Kat Çapraz Doğrulamaya Göre Modellerin Performans Değerleri

İndirgenmemiş örnekleme veri setinde 2.830 satır veri ve 9 öznitelik modele alınmıştır. Örneklem veri Seti için 5-kat çapraz doğrulama yöntemi ile geliştirilen modellerin her bir kat için doğruluk performansı ve tüm katların ortalama doğruluk performans değerleri aşağıda yer verilmiştir.

Tablo 3.36 İndirgenmemiş Örneklem Veri Seti İçin 5-Kat Çapraz Doğrulama Modellerin Doğruluk Performans Değerleri

	k-1	k-2	k-3	k-4	k-5	k-Ort.
K En Yakın Komşu	0,611	0,617	0,673	0,611	0,654	0,633
Lojistik Regresyon	0,541	0,571	0,610	0,594	0,601	0,583
Naive Bayes	0,597	0,553	0,557	0,527	0,602	0,567
Karar Ağaçları	0,634	0,684	0,677	0,682	0,666	0,669
Rasgele Orman	0,629	0,684	0,686	0,687	0,696	0,676
Destek Vektör Makinaları	0,512	0,583	0,624	0,604	0,553	0,575
Yapay Sinir Ağları	0,564	0,604	0,663	0,643	0,618	0,618

Ortalama doğruluk oranı en yüksek modeller rasgele orman ve karar ağaçları algoritması ile oluşturulmuştur. Yüksek doğruluk oranında bu iki algoritmayı k en yakın komşu algoritması takip etmektedir. Ortalama doğruluk oranı en düşük modeller ise naive bayes, lojistik regresyon ve destek vektör makinaları algoritmaları ile oluşturulmuştur. En yüksek doğruluk performansı %69,6 oran ile 5-kat doğru çaprazlama ile rasgele orman algoritmasında gerçekleşmiştir. En düşük doğruluk performansı ise %51,2 oran ile 1-kat doğru çaprazlama ile destek vektörleri algoritmasında gerçekleşmiştir.

İndirgenmemiş Örneklem Veri Seti 10-Kat Çapraz Doğrulamaya Göre Modellerin Performans Değerleri

İndirgenmemiş örnekleme veri setinde 2.830 satır veri ve 9 öznitelik modele alınmıştır. Örneklem veri Seti için 10-kat çapraz doğrulama yöntemi ile geliştirilen modellerin her bir kat için doğruluk performansı ve tüm katların ortalama doğruluk performans değerleri aşağıda yer verilmiştir.

Tablo 3.37 Örneklem Veri Seti(İndirgenmemiş) 10-Kat Çapraz Doğrulama Modellerin Doğruluk Performans Değerleri

	k-1	k-2	k-3	k-4	k-5	k-6	k-7	k-8	k-9	k-10	k-O.
K En Yakın Komşu	0,530	0,696	0,633	0,650	0,721	0,629	0,608	0,636	0,654	0,640	0,640
Lojistik Regresyon	0,527	0,537	0,576	0,590	0,608	0,604	0,530	0,654	0,601	0,597	0,582
Naive Bayes	0,565	0,604	0,530	0,597	0,587	0,519	0,527	0,558	0,622	0,576	0,569
Karar Ağaçları	0,590	0,693	0,668	0,682	0,714	0,671	0,686	0,664	0,661	0,689	0,672
Rasgele Orman	0,587	0,700	0,678	0,703	0,731	0,671	0,700	0,661	0,693	0,693	0,682
Destek Vektör Makinaları	0,505	0,519	0,590	0,604	0,576	0,633	0,569	0,657	0,544	0,558	0,576
Yapay Sinir Ağları	0,569	0,618	0,590	0,661	0,671	0,608	0,565	0,657	0,618	0,625	0,618

Ortalama doğruluk oranı en yüksek modeller rasgele orman ve karar ağaçları algoritması ile oluşturulmuştur. Yüksek doğruluk oranında bu iki algoritmayı k ne yakın komşu algoritması takip etmektedir. Ortalama doğruluk oranı en düşük modeller ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları ile

oluşturulmuştur. En yüksek doğruluk performansı %73,1 oran ile 5-kat doğru çaprazlama ile rasgele orman algoritmasında gerçekleşmiştir. En düşük doğruluk performansı ise %50,5 oran ile 1-kat doğru çaprazlama ile destek vektörleri algoritmasında gerçekleşmiştir.

Çapraz doğrulama sayısının 5-kat ve 10-kat olmasına göre ortalama doğrulama oranları birlikte incelendiğinde %1'in altında çok düşük farkların olduğu tespit edilmiştir. Bu durum çapraz doğrulamadaki kat farklılığının model performansında belirgin bir artış ve azalışa neden olmadığı sonucunu ortaya koymaktadır.

İndirgenmemiş Örneklem Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

İndirgenmemiş örneklem veri seti hold out geçerleme yöntemine göre %70 eğitim-%30 test ve %80 eğitim-%20 test şeklinde ve k-kat çapraz doğrulama yöntemine göre de 5-kat ve 10-kat çapraz doğrulamaya tabi tutulmuş ve parametreleri aynı olan makine öğrenmesi algoritmaları ile modeller geliştirilerek modellerin performans değerleri doğruluk, kesinlik, duyarlılık ve f-ölçütü performans göstergeleri üzerinden tespit edilmiştir. Modellerin sadece doğruluk/ortalama doğruluk ölçütü üzerinden performans karşılaştırmaları yapılmıştır.

Tablo 3.38 İndirgenmemiş Örneklem Veri Seti Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,665	0,647	0,633	0,640
Lojistik Regresyon	0,605	0,599	0,583	0,582
Naive Bayes	0,562	0,602	0,567	0,569
Karar Ağaçları	0,658	0,724	0,669	0,672
Rasgele Orman	0,693	0,726	0,676	0,682
Destek Vektör Mak.	0,600	0,624	0,575	0,576
Yapay Sinir Ağları	0,642	0,652	0,618	0,618

Karşılaştırmalı performans değerlerine göre en yüksek doğruluk oranı %72,6 oran ile rasgele orman algoritması ve %80 eğitim-%20 test model geçerleme yönteminde gerçekleşmiştir. İkinci sırada en yüksek doğruluk performansı %72,4 oran ile karar ağaçları algoritması ve yine %80 eğitim-%20 test model geçerleme yönteminde

gerçekleşmiştir. En yüksek doğruluk oranında rasgele orman algoritmasını tüm geçerleme/doğrulama yöntemlerinde karar ağaçları algoritması ve k en yakın komşu algoritması takip etmektedir. En düşük doğruluk performansı %56,2 oran ile naive bayes algoritmasında %70 eğitim-%30 test model geçerleme yönteminde gerçekleşmiştir. İkinci sırada en düşük doğruluk performansı %56,2 oran ile yine naive bayes algoritmasında 5-kat çapraz doğrulama yönteminde gerçekleşmiştir. Destek vektör makinaları, naive bayes ve lojistik regresyon algoritmaları tüm geçerleme/doğrulama yöntemlerinde en düşük doğruluk performansı göstermiştir. Yapay sinir ağları algoritması ise tüm geçerleme/doğrulama yöntemlerinde orta düzeyde doğruluk performansı göstererek en yüksek ve en düşük doğruluk performansı gösteren algoritmaların ortasında yer almaktadırlar.

Makine öğrenmesi algoritmalarının farklı tür ve şekildeki doğruluk performansı incelendiğinde doğruluk performansının belirleyici olarak model geçerleme/doğrulama yöntemlerinin etkisinin az olduğu esas belirleyicinin algoritmalar olduğu göze çarpmaktadır. Başka bir ifadeyle makine öğrenmesi algoritması sabit iken model geçerleme yöntemleri ve tiplerinde meydana gelen değişimler modelin doğruluk performansını çok az etkilemektedir.

İndirgenmiş Örneklem Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Oluşturulan Modellerin Performans Değerleri

Buraya kadar indirgenmemiş örneklem veri seti için farklı makine öğrenmesi algoritmaları ve modele geçerleme yöntemlerine göre modeller oluşturulmuş ve performans hesapları yapılmıştır. Bu bölümde ise daha önce istatistiksel metotlar ile gerçekleştirilen öznitelik seçme ve çıkarma işlemine göre indirgenmiş örneklem veri seti için modeller oluşturuldu, performans hesabı ve değerlendirmesi yapılmıştır.

İndirgenmiş örneklem veri setinde indirgenmemiş örneklem veri setine göre sadece cinsiyet ve ikamet özniteliği bulunmayıp diğer öznitelikler bulunmaktadır. İndirgenmiş örneklem veri setinde 2.830 satır veri ve 7 öznitelik modele alınmıştır. İndirgenmiş örneklem veri Seti için Hold Out Model geçerleme yöntemi ve %70 eğitim-%30 test ve %80 eğitim-%20 test verisi için geliştirilen modellerin performans değerleri aşağıda yer verilmiştir.

Tablo 3.39 İndirgenmiş Örneklem Veri Seti İçin Hold Out Model Geçerleme Yöntemine Göre Modellerin Performans Değerleri ve Karşılaştırmaları

	%70 Eğitim-%30 Test				%80 Eğitim-%20 Test			
	Doğruluk (Accurac	Kesinlik (Precisio	Duyarlılık (Recall)	F1 Skor	Doğruluk (Accurac	Kesinlik (Precisio	Duyarlılık (Recall)	F1 Skor
K En Yakın Komşu	0,637	0,628	0,695	0,660	0,627	0,635	0,659	0,647
Lojistik Regresyon	0,601	0,608	0,598	0,603	0,606	0,622	0,608	0,615
Naive Bayes	0,549	0,662	0,223	0,334	0,581	0,741	0,294	0,421
Karar Ağaçları	0,640	0,633	0,686	0,658	0,675	0,682	0,696	0,689
Rasgele Orman	0,661	0,644	0,740	0,688	0,700	0,682	0,785	0,730
Destek Vektör Mak.	0,595	0,571	0,802	0,667	0,606	0,591	0,775	0,671
Yapay Sinir Ağları	0,617	0,620	0,633	0,626	0,636	0,649	0,645	,647

İndirgenmiş örneklem veri setine ilişkin hold out model geçerleme yöntemine göre hazırlanan modelde en iyi performans değerleri rasgele orman, karar ağaçları ve k en yakın komşu algoritmaları göstermiştir. En kötü performans değerlerini ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları göstermiştir. Yapay sinir ağları iyi ve kötü performans gösteren algoritmalara göre ortalama bir performans göstermiştir.

İndirgenmiş örneklem veri setinde eğitim ve test verisi oranlarındaki değişim algoritmaların performans göstergelerinin çok düşük bir değişime uğramasına neden olmuştur. Ancak bu çok düşük değişim yüksek performans gösteren algoritmalarda daha yüksek düşük performans gösteren algoritmalarda daha düşük şekilde gerçekleşmiştir. Başka bir ifadeyle eğitim ve test verisi oranlarındaki değişim kaynaklı oluşan performans artış/azalışı algoritma performansı ile doğrusal bir ilişki içindedir.

Tablo 3.40 İndirgenmiş Örneklem Veri Seti İçin K-Kat Çapraz Doğrulama Yöntemine Göre Modellerin Performans Değerleri ve Karşılaştırmaları

	5-Kat Çapraz Doğrulama	10-Kat Çapraz Doğrulama
	Ortalama Doğruluk	Ortalama Doğruluk
K En Yakın Komşu	0,612	0,616
Lojistik Regresyon	0,584	0,584
Naive Bayes	0,548	0,553
Karar Ağaçları	0,650	0,649
Rasgele Orman	0,657	0,653
Destek Vektör Mak.	0,580	0,573
Yapay Sinir Ağları	0,608	0,614

İndirgenmiş örneklem veri setine ilişkin k-kat çapraz doğrulama yöntemine göre hazırlanan modelde en iyi performans değerleri rasgele orman, karar ağaçları algoritmaları göstermiştir. En kötü performans değerlerini ise naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları göstermiştir. K en yakın komşu ve yapay sinir ağları iyi ve kötü performans gösteren algoritmalara göre ortalama bir performans göstermiştir.

Çapraz doğrulama kat sayısındaki farklılık algoritmaların performans göstergelerinin çok az bir değişime uğramasına neden olmuştur. Dolayısıyla indirgenmiş veri setinde de model performansını algoritmalar belirlemiştir.

Tablo 3.41 İndirgenmiş Örneklem Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Modellerin Doğruluk Ölçütü Üzerinden Performans Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,637	0,627	0,612	0,616
Lojistik Regresyon	0,601	0,606	0,584	0,584
Naive Bayes	0,549	0,581	0,548	0,553
Karar Ağaçları	0,640	0,675	0,650	0,649
Rasgele Orman	0,661	0,700	0,657	0,653
Destek Vektör Mak.	0,595	0,606	0,580	0,573
Yapay Sinir Ağları	0,617	0,636	0,608	0,614

İndirgenmiş örneklem veri seti için karşılaştırmalı performans değerlerine göre en yüksek doğruluk oranı %70 oran ile rasgele orman algoritmasında %80 eğitim-%20 test model geçerleme yönteminde gerçekleşmiştir. İkinci sırada en yüksek doğruluk performansı %66,1 oran ile yine rasgele orman algoritmasında %70 eğitim-%30 test model geçerleme yönteminde gerçekleşmiştir. En yüksek doğruluk oranında rasgele orman algoritmasını tüm geçerleme/doğrulama yöntemlerinde karar ağaçları algoritması ve k en yakın komşu algoritması takip etmektedir. En düşük doğruluk performansı %54,8 oran ile naive bayes algoritmasında 5-kat çapraz doğrulama yönteminde gerçekleşmiştir. İkinci, üçüncü ve dördüncü sırada en düşük doğruluk performansı yine naive bayes algoritmasında diğer model geçerleme/doğrulama yöntemlerinde gerçekleşmiştir. Naive bayes, destek vektör makinaları ve lojistik regresyon algoritmaları tüm geçerleme/doğrulama yöntemlerinde en düşük doğruluk performansı göstermiştir. Yapay sinir ağları algoritması ile k en yakın komşu

algoritması ise tüm geçerleme/doğrulama yöntemlerinde orta düzeyde doğruluk performansı göstererek en yüksek ve en düşük doğruluk performansı gösteren algoritmaların ortasında yer almaktadırlar.

Makine öğrenmesi algoritmalarının farklı tür ve şekildeki doğruluk performansı incelendiğinde doğruluk performansının belirleyici olarak model geçerleme/doğrulama yöntemlerinin etkisinin az olduğu esas belirleyicinin algoritmalar olduğu göze çarpmaktadır. Diğer bir ifadeyle model algoritması sabit iken model geçerleme/doğrulama yöntemlerindeki değişimler model performanslarını yok denecek kadar az etkilemektedir. Bu çıkarımın aynısı daha yığın veri seti içinde geçerlidir.

İndirgenmiş Örneklem Veri Seti ve İndirgenmemiş Örneklem Veri Seti İle Oluşturulan Modeller Arasındaki Doğruluk Performansı Karşılaştırması

Tablo 3.42 İndirgenmiş Örneklem Veri Seti ve İndirgenmemiş Örneklem Veri Seti İle Oluşturulan Modeller Arasındaki Doğruluk Performansı Karşılaştırması

	Hold Out Geçerleme		**K-Kat Çapraz Doğrulama**	
	%70 Eğitim %30 Test	**%80 Eğitim %20 Test**	**5-Kat Çapraz Doğrulama**	**10-Kat Çapraz Doğrulama**
K En Yakın Komşu	0,028	0,02	0,021	0,024
Lojistik Regresyon	0,004	-0,007	-0,001	-0,002
Naive Bayes	0,013	0,021	0,019	0,016
Karar Ağaçları	0,018	0,049	0,019	0,023
Rasgele Orman	0,032	0,026	0,019	0,029
Destek Vektör Mak.	0,005	0,018	-0,005	0,003
Yapay Sinir Ağları	0,025	0,016	0,01	0,004

İndirgenmemiş ve indirgenmiş örneklem veri setlerine göre oluşturulan modellerdeki doğruluk oranları arasında farklı algoritma ve model geçerleme/doğrulama yöntemlerine fark etmeksizin çok düşük bir fark olduğu anlaşılmaktadır. İndirgenmiş ve indirgenmemiş iki örneklem veri setine göre en yüksek doğruluk performansı farklılığı rasgele orman, k en yakın komşu ve karar ağaçları algoritmalarında olmuştur. Bu farklılık tüm model geçerleme/doğrulama yöntemleri için ortalama %2,5 civarındadır. Aynı zamanda bu algoritmalar en yüksek doğruluk performansı gösteren algoritmalardır.

İndirgenmiş ve indirgenmemiş iki örneklem veri setine göre en düşük doğruluk performansı farklılığı destek vektör makinaları, lojistik regresyon ve naive bayes algoritmalarında olmuştur. Bu farklılık tüm model geçerleme/doğrulama yöntemleri için ortalama %1 civarındadır. Aynı zamanda bu algoritmalar en düşük doğruluk performansı gösteren algoritmalardır.

İndirgenmiş veri seti indirgenmemiş veri setinden sadece cinsiyet ve ikamet özniteliği yönünden farklılık arz etmektedir. İndirgenmiş model hem daha düşük doğruluk performansı göstermiş hem de bir işsizin iş arama davranışında olmazsa olmaz özniteliği olan cinsiyet ve ikamet model dışı bırakmıştır. Bu durum işsizlerin risk sınıflaması ve iş profillemesi için kabul edilemezdir. Dolayısıyla indirgenmemiş model indirgenmiş modele göre hem doğruluk performansı hem de öznitelik zenginliği yönünden üstünlük arz etmektedir.

Öznitelik çıkarma işlemi yüksek doğruluk performansı gösteren algoritmalarda daha yüksek oranda doğruluk performansı düşüşü gerçekleştirirken, düşük performans gösteren algoritmalarda daha düşük oranda doğruluk performansı düşüşüne neden olmaktadır. Öznitelik çıkarma işlemi ile gerçekleştirilen boyut azaltmada çıkarılan öznitelik sayısı arttıkça doğruluk oranında azalma beklenmektedir. Dolayısıyla boyut azaltma işleminin her veri setine uygulanmayacağı değerlendirilmektedir. Özellikle tez çalışmasında olduğu gibi öznitelik sayısının az ve yetersiz olduğu veri setinde öznitelik çıkarma işlemi doğruluk performansını düşürdüğü gibi veriye dayalı bilgi kaybına da neden olmaktadır.

Yukarıda yer alan değerlendirmelerin tamamı neredeyse yığın veri seti içinde daha önce yapılmıştır. Bu durum yığın veri seti ile örneklem veri seti sonuçları arasında çok yüksek benzerlikler olduğunu ortaya koymaktadır. Ancak her iki örneklem veri seti ile yapılan tüm modellemelerin bir tabloda gösterilmesi bu benzerlikleri daha net ortaya koyacaktır.

İndirgenmiş ve İndirgenmemiş Yığın ve Örneklem Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Oluşturulan Modellerin Doğruluk Performans Değerleri

İndirgenmiş ve indirgenmemiş yığın ve örneklem veri seti için farklı geçerleme/doğrulama tipi ve makine öğrenmesi algoritmasına göre 112 adet model oluşturulmuş ve oluşturan modellerin doğruluk performanslarına yer verilmiştir.

Tablo 3.43 İndirgenmiş ve İndirgenmemiş Yığın ve Örneklem Veri Seti İçin Geçerleme/Doğrulama Tipi ve Şekline Göre Oluşturulan Modellerin Doğruluk Performans Değerleri

			K En Yakın Komşu	Lojistik Regresyon	Naive Bayes	Karar Ağaçları	Rasgele Orman	Destek Vektör	Yapay Sinir Ağları	Ortalama
Yığın	**İndirgenmemiş**	**%70 Eğitim-%30 Test**	0,724	0,606	0,583	**0,777**	**0,784**	0,602	0,676	0,679
		%80 Eğitim-%20 Test	**0,731**	**0,607**	**0,584**	**0,781**	**0,788**	**0,593**	**0,686**	**0,681**
		5-Kat Çapraz Doğrulama	0,726	0,595	0,582	**0,782**	**0,786**	0,581	0,676	0,675
		10-Kat Çapraz Doğrulama	0,727	0,594	0,581	**0,782**	**0,786**	0,580	0,671	0,674
	İndirgenmiş	**%70 Eğitim-%30 Test**	0,685	0,607	0,574	0,750	**0,756**	0,599	0,657	0,661
		%80 Eğitim-%20 Test	0,722	0,602	0,574	0,753	**0,755**	0,589	0,675	0,667
		5-Kat Çapraz Doğrulama	0,700	0,594	0,568	0,757	**0,760**	0,579	0,67	0,661
		10-Kat Çapraz Doğrulama	0,699	0,595	0,568	0,755	**0,759**	0,578	0,672	0,661
Örneklem	**İndirgenmemiş**	**%70 Eğitim-%30 Test**	0,665	0,605	0,562	0,658	**0,693**	0,600	0,642	0,632
		%80 Eğitim-%20 Test	0,647	0,599	0,602	0,724	**0,726**	0,624	0,652	0,653
		5-Kat Çapraz Doğrulama	0,633	0,583	0,567	0,669	**0,676**	0,575	0,618	0,617
		10-Kat Çapraz Doğrulama	0,64	0,582	0,569	0,672	**0,682**	0,576	0,618	0,620
	İndirgenmiş	**%70 Eğitim-%30 Test**	0,637	0,601	0,549	0,640	**0,661**	0,595	0,617	0,614
		%80 Eğitim-%20 Test	0,627	0,606	0,581	0,675	**0,700**	0,606	0,636	0,633
		5-Kat Çapraz Doğrulama	0,612	0,584	0,548	0,650	**0,657**	0,58	0,608	0,606
		10-Kat Çapraz Doğrulama	0,616	0,584	0,553	0,649	**0,653**	0,573	0,614	0,606
		Ortalama	0,674	0,597	0,572	0,717	**0,726**	0,589	0,649	0,646

Farklı tip ve şekildeki model geçerleme/doğrulama ve makine öğrenmesi algoritmaları ile oluşturulan modellerin neredeyse tamamında yığın veri seti ile oluşturulan modellerin doğruluk performans oranı örneklem veri setinden daha fazladır. Yığın veri seti modelleri ile ortalama yaklaşık %67 doğruluk performansı sergilenirken,

örneklem veri seti ile oluşturulan modeller ile ortalama yaklaşık %62 doğruluk performansı sergilenmiştir. Yığın veri seti ile oluşturulan modeller örneklem veri setine göre oluşturulan modellere göre ortalama yaklaşık %5 oranında daha fazla doğruluk performansı göstermektedir. Esasında bu performans farklılığı çok yüksek olmayıp örneklem veri seti için iyi bir model performansı olarak kabul edilebilir. Ayrıca bu sonuç, yığın veri seti ile modellemenin daha maliyetli olduğu durumlarda örneklem veri setinin tercih edilebilirliğini göstermektedir. Bununla birlikte örneklem veri setinin yığın veri setinden çok daha farklı olmayan doğruluk performansı göstermesi örneklem veri setinin yığını iyi temsil ettiği çıkarımı yapılabilir. Hâlihazırda tez çalışmasında belirlenen örneklem veri seti hem örneklem sayısı hem de sistematikliği yönünden yığından iyi bir şekilde seçilmiştir. Oluşturulacak bir makine öğrenmesi modellemesinde yığın veri seti mi yoksa örneklem veri setinin mi kullanılacağına somut araştırmanın konusu, amacı ve maliyetine göre araştırmacı belirleyecektir. Tez çalışmasında işsizlerin İŞKUR'a başvuru sonrası bir yıl içerisinde işsiz kalma riski tespit edilmekte olup bu risk tespitinin modellenmesinde hem yığın hem de örneklem verisi kullanılabilir olup yığın veri seti daha iyi bir modelleme fırsatı sunmaktadır.

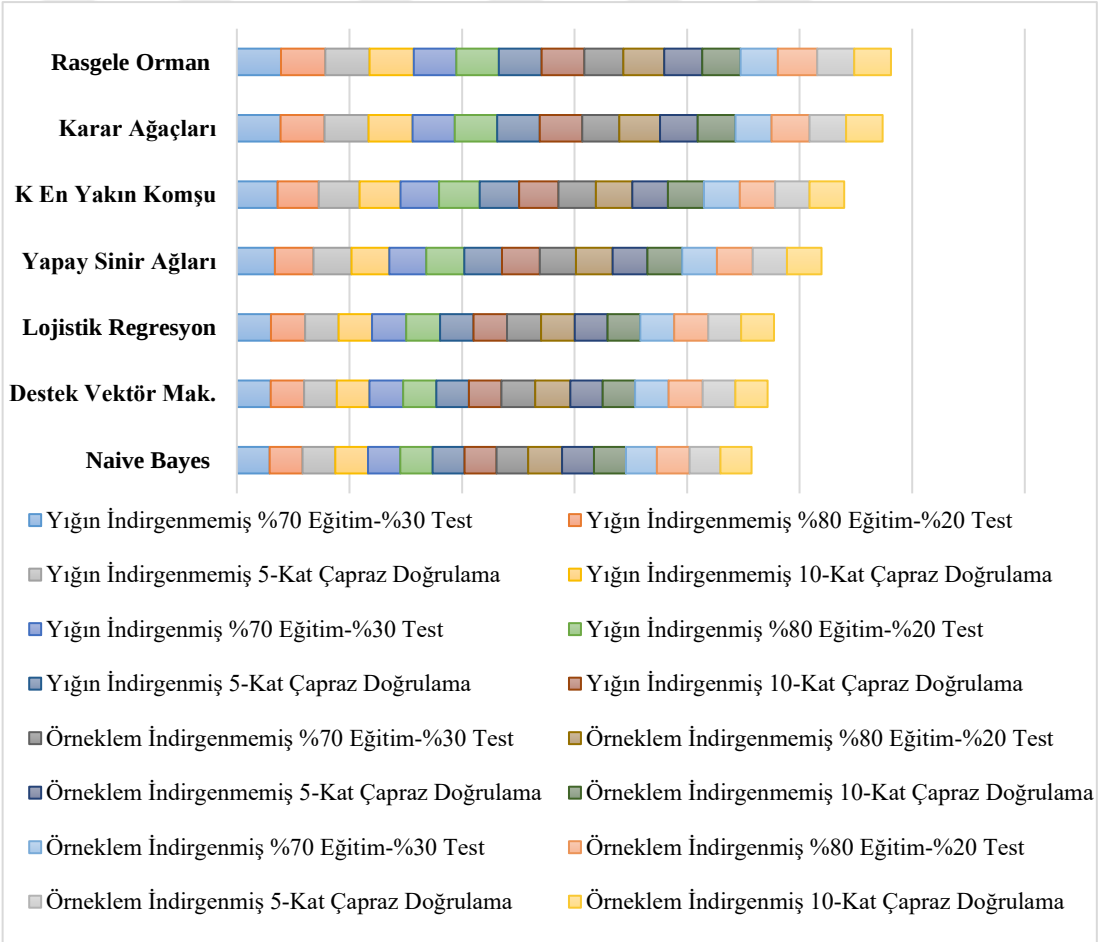
Farklı tip ve şekildeki model geçerleme/doğrulama ve makine öğrenmesi algoritmaları ile yığın ve örneklem veri setine göre oluşturulan modellerin neredeyse tamamında indirgenmemiş veri seti ile oluşturulan modellerin doğruluk performans oranı indirgenmiş veri setinden daha fazladır. İndirgenmemiş yığın veri seti ortalama yaklaşık %68 doğruluk performansı gösterirken indirgenmiş yığın veri set ise ortalama yaklaşık %66 doğruluk performansı göstermiştir. İndirgenmemiş örneklem veri seti ortalama yaklaşık %63 doğruluk performansı gösterirken indirgenmiş örneklem veri set ise ortalama yaklaşık %61 doğruluk performansı göstermiştir. İndirgenmiş veri seti ile oluşturulan modeller indirgenmemiş veri setine göre oluşturulan modellere göre ortalama yaklaşık %2 oranında daha fazla doğruluk performansı göstermektedir. İndirgenmemiş veri seti ile indirgenmiş veri seti arasındaki bu çok düşük orandaki doğruluk performans farklılığı indirme işleminde çıkarılan öznitelik sayısının azlığından kaynaklanmaktadır. Dolayısıyla çıkarılan öznitelik sayısı arttıkça doğruluk performans oranındaki farkın artacağı değerlendirilmektedir. Burada hem indirgenmiş hem de indirgenmemiş veri seti ile tez çalışması modellenebilir. Ancak indirme işlemi gerçekleştirilirken çıkarılan cinsiyet ve ikamet özniteliğinim işsiz kalma risk

değerlendirilmesi için olmazsa olmaz nitelik arz etmesi ve indirgenmemiş veri seti ile daha yüksek doğruluk performansı elde edilmesi tez çalışması için indirgenmemiş veri setini daha fazla tercih edilebilir kılmaktadır. Bu noktada indirgenmemiş yığın veri seti birlikteliğinin daha iyi bir modelleme fırsatı sunmaktadır.

Farklı tipteki makine öğrenmesi algoritmaları ile indirgenmiş ve indirgenmemiş yığın ve örneklem veri setine göre oluşturulan modellerin neredeyse tamamında test ve eğitim verisi şeklinde holdout model geçerleme yöntemi ile oluşturulan modellerin doğruluk performans k-kat çapraz doğrulama yöntemi ile oluşturulan modellerin doğruluk performansından daha fazladır. Ancak bu fark yüksek bir farklılık arz etmemekte olup yığın veri seti ile oluşturulan modellerde ortalama yaklaşık olarak %0,5 ve daha az iken örneklem veri seti ile oluşturulan modellerde ise ortalama yaklaşık %2,4 ve daha az şeklindedir. Başka bir ifadeyle modellerin doğruluk performansında model geçerleme yöntemi ile model çapraz doğrulama yöntemi arasında belirgin bir fark bulunmamaktadır. Bu noktada her iki yönteminde kullanılabilir. Ancak önemle belirtmek gerekir ki k-kat çapraz doğrulama yöntemi veri setinin yapısı ile doğrudan ilintili olup tabakalı ve sıralı hale getirilen veri setinde oldukça düşük performans göstermektedir. Tez çalışmasında yığından örneklem çekme işlemi için yığın veri seti tabakalanmış ve sıralanmış olması durumunda gerçekleştirilen k-kat çapraz doğruluma yöntemi ile oluşturulan modeller düşük doğruluk performansı göstermiş ve bu durumun aşılması için veri seti random edilerek sistematik yapı ortadan kaldırılmış ve k-kat çapraz doğrulama ile tekrar oluşturulan modeller daha yüksek doğruluk performansı göstermiştir. Test-eğitim verisi ile oluşturulan modeller k-kat çapraz doğrulama ile oluşturulan modellere göre az da olsa daha yüksek doğruluk performansı göstermesi ve test-eğitim verisi ile oluşturulan modellerin k-kat çapraz doğrulamaya göre oluşturulan modellere göre veri setinin yapısından daha az etkilenmesi tez çalışması için test-eğitim verisi ile model oluşturmayı daha tercih edilebilir kılmaktadır. Bu bağlamda indirgenmemiş yığın veri setinin test-eğitim verisi olarak ayrılarak gerçekleştirilen makine öğrenmesi yöntemi ile daha iyi bir modelleme fırsatı sunmaktadır.

Farklı tipteki makine öğrenmesi algoritmaları ile indirgenmiş ve indirgenmemiş yığın ve örneklem veri setine göre oluşturulan modellerin neredeyse tamamında %80 eğitim-%20 test verisi şeklinde oluşturulan modellerin doğruluk performans %70 eğitim-%30 test verisi şeklinde oluşturulan modellerin doğruluk performansından

daha fazladır. Ancak bu fark yüksek bir farklılık arz etmemekte olup yığın veri seti ile oluşturulan modellerde ortalama yaklaşık olarak %0,6 ve daha az iken örneklem veri seti ile oluşturulan modellerde ise ortalama yaklaşık %2,1 ve daha az şeklindedir. Başka bir ifadeyle modellerin doğruluk performansında %80 eğitim-%20 test verisi şeklinde ve %70 eğitim-%30 test verisi şeklinde yöntemi arasında belirgin bir fark bulunmamaktadır. Bu noktada her iki yönteminde kullanılabilir. %80 eğitim-%20 test verisi ile oluşturulan modeller %70 eğitim-%30 test ile oluşturulan modellere göre az da olsa daha yüksek doğruluk performansı göstermesi nedeniyle tez çalışması için %80 eğitim-%20 test ile model oluşturmayı daha tercih edilebilir ve elverişli kılmaktadır. Bu bağlamda indirgenmemiş yığın veri setinin %80 eğitim-%20 test olarak ayrılarak gerçekleştirilen makine öğrenmesi yöntemi ile daha iyi bir modelleme fırsatı sunmaktadır.



Şekil 3.3 Makine Öğrenmesi Modellerinin Doğruluk Performans Karşılaştırması

Oluşturulan modellerin tamamında rasgele orman algoritması ile oluşturulan modellerin doğruluk performans diğer algoritmalar ile oluşturulan modellerin

doğruluk performansından daha fazladır. Rasgele orman algoritması ise oluşturulan modellerin ortalama yaklaşık doğruluk performansı %72,6'dır. Bu doğruluk performansını %71,7 ile karar ağaçları, %67,4 ile k en yakın komşu, %64,9 ile yapay sinir ağları, %59,7 ile lojistik regresyon, %58,9 ile destek vektör makinaları ve %57,2 ile naive bayes algoritmaları takip etmektedir. Rasgele orman ve karar ağaçları algoritmalarının diğer algoritmalara göre daha yüksek doğruluk performansı göstermesi ve algoritmalar arasında doğruluk performansı yönünden belirgin bir fark olmaması her iki algoritmanın kullanılmasını elverişli kılmaktadır. Bu bağlamda indirgenmemiş yığın veri setinin %80 eğitim-%20 test olarak ayrılarak rasgele orman ve karar ağaçları algoritmaları ile gerçekleştirilen makine öğrenmesi yöntemleri ile daha iyi bir modelleme fırsatı sunmaktadır.

Oluşturulan tablo ve grafikte yer yer alan tüm bilgiler birlikte değerlendirildiğinde modellerin doğruluk performanslarında asıl belirleyicilerin makine öğrenmesi algoritmaları ve veri seti türünün olduğu, model geçerleme ve doğrulama yöntemlerindeki değişimin modelin doğruluk performansında yok denecek kadar az seviyede bir etkisi olduğu, veri setinin öznitelik çıkartılması ile oluşan indirgenmiş veri seti ile oluşturulan modelin doğruluk performansının çıkarılan öznitelik sayısı ile doğrusal bir ilişki gösterdiği tespit edilmiştir. Ayrıca en yüksek doğruluk performansı gösteren indirgenmemiş yığın veri seti için rasgele orman ve karar ağaçları algoritmaları için temel parametrelerde değişiklik yapılarak doğruluk performansı iyileştirmesi yapılarak yeni modeller oluşturulabilir.

3.11 Makine Öğrenmesi Modellerine İlişkin Performans İyileştirmesi

Veri seti, yığın/örneklem, indirgenmemiş/indirgenmiş, model geçerleme/model doğrulama, %80 Eğitim-%20 Test/ %70 Eğitim-%30 Test, 5-kat/10-kat çapraz doğrulama ve farklı tipteki makine algoritmaları ile modellenmiş ve en yüksek doğruluk performansı indirgenmemiş yığın veri seti için rasgele orman ve karar ağaçları ve bu algoritmalara en yakın olarak k en yakın komşu algoritmasında gerçekleşmiştir. Hem model geçerleme hem de model doğrulama yöntemleri doğruluk performansı üzerinde çok düşük farklılık arz etmektedir. Esasında yapılan tüm bu işlemler ile performans iyileştirme büyük ölçekte gerçekleştirilmiş ve makine öğrenme algoritmalarının ve veri seti türünün doğruluk performansındaki etkileri tespit edilmiştir.

Algoritmaların doğruluk performanslarının daha da arttırılması için yapılması gereken bir diğer işlemde algoritma parametrelerinin değiştirilmesidir. Burada hâlihazırda en yüksek doğruluk performans gösteren rasgele orman, karar ağaçları ve k en yakın komşu algoritmasında makine öğrenmesi algoritmalarının parametrelerinde değişim yaparak daha yüksek bir doğruluk performansı yakalanması amaçlanmıştır. En yüksek doğruluk performansı indirgenmemiş yığın veri seti için %80 Eğitim- %20 Test model geçerleme işlemi ile gerçekleştirildiği için parametreler üzerinden model iyileştirme işlemi bu veri seti ve model geçerleme ile yapılmıştır.

Tablo 3.44 Karar Ağaçları ve Rasgele Orman Algoritmasına İlişkin İyileştirilmiş Performans Değerleri

	KARAR AĞAÇLARI		**RASGELE ORMAN**	
	Bölünme Yöntemi (Criterion) Gini	**Bölünme Yöntemi (Criterion) Entropy**	**Bölünme Yöntemi (Criterion) Gini**	**Bölünme Yöntemi (Criterion) Entropy**
Doğruluk (Accuracy)	0,781	0,782	0,789	0,788
Kesinlik (Precision)	0,745	0,747	0,753	0,752
Duyarlılık(Recall)	0,848	0,848	0,855	0,855
F-Ölçütü	0,793	0,794	0,801	0,800

Hem karar ağaçlarında hem de rasgele orman algoritmasında performans iyileştirme sadece bölünme yöntemi(criterion) üzerinden gerçekleştirilmiştir. Diğer parametreler[29] sabit tutulmuştur. Gini ve entropy bölünme yöntemine göre hem karar ağaçlarında hem rasgele orman algoritmasında bölünme yöntemlerindeki farklılık yok denecek kadar düşük bir oranda(yaklaşık binde bir) değişime neden olmuştur. Dolayısıyla ana parametre olarak sayılan bölünme yöntemindeki farklılıklar karar ağacı ve rasgele orman algoritmalarının başta doğruluk olmak üzere, kesinlik ve duyarlılık performans değerleri üzerinde belirgin bir iyileştirme sağlamamaktadır.

[29] **Karar ağacında sabit tutulan parametreler;**
splitter='best',max_depth=None,min_samples_split=2,min_samples_leaf=1,min_weight_fraction_leaf=0.0,max_features=None,random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0)

Rasgele orman algoritmasında sabit tutulan parametreler;
n_estimators=10,max_depth=None,min_samples_split=2,min_samples_leaf=1,min_weight_fraction_leaf=0.0,max_features='sqrt',max_leaf_nodes=None,min_impurity_decrease=0.0,bootstrap=True,oob_score=False,n_jobs=None,random_state=None,verbose=0,warm_start=False,class_weight=None,ccp_alpha=0.0,max_samples=None)

Tablo 3.45 K-NN Algoritmasına İlişkin İyileştirilmiş Performans Değerleri

	1-NN	2-NN	3-NN	4-NN	5-NN	6-NN	7-NN	8-NN	9-NN	10-NN
Doğruluk (Accuracy)	0,731	0,693	0,737	0,733	0,740	0,736	0,738	0,746	0,742	0,746
Kesinlik (Precision)	0,728	0,761	0,711	0,731	0,694	0,713	0,687	0,706	0,693	0,704
Duyarlılık (Recall)	0,731	0,556	0,791	0,728	0,850	0,782	0,865	0,834	0,863	0,843
F-Ölçütü	0,729	0,642	0,749	0,729	0,764	0,746	0,766	0,765	0,769	0,767

K-En yakın komşu algoritması uzaklık ölçüleri, komşu sayısı ve komşu ağırıklandırma olmak üzere bir takım ana parametrelere sahiptir. Ana parametrelerde meydana gelen değişimlerin algoritma performansı üzerine etkileri bulunmaktadır. Tez çalışmasında minkowski uzaklık ölçüsünde sadece komşu sayısında meydana gelen değişimlerin algoritma performansı üzerindeki etkileri incelenmiştir. Komşu sayısının 1-NN olduğu durumda %73,1 olan doğruluk performansı, 10-NN olduğu durumda %74,6 olmuştur. Dolayısıyla yaklaşık %1,5 oranında bir doğruluk performans artışı gerçekleşmiştir. Bu bilgi komşu sayısındaki artışın doğruluk performansında belirgin bir artışa neden olmadığı sonucunu ortaya koymuştur. Ancak önemle belirtmek gerekir ki duyarlılık ölçütünde yaklaşık %11,3 oranda belirgin bir performans artışı gözlemlenmiştir. Duyarlılık artışı ile İŞKUR'a başvuru sonrası bir yıl süre içerisinde gerçekte işe yerleşmiş olanların doğru tahmin edilme oranının da artış meydana gelmiştir.

Her üç algoritma için parametre değişimleri üzerinden meydana gelen performans değişimleri birlikte değerlendirildiğinde, algoritmaların parametre değişimleri algoritma performanslarında belirgin bir artış veya azalışa neden olmamaktadır. Ancak %1'lik bir doğru sınıflama artışının çok önem arz ettiği araştırmalarda parametre değişimleri kaynaklı oluşan küçük performans artışları göz ardı edilmemesi gerekmektedir. Tez çalışmasında İŞKUR'a başvuran işsizlerin başvuru sonrası bir yıl içerisinde işsiz kalma riskleri tespit edilmeye çalışıldığı için düşük performans artışları değer olarak kabul edilmektedir. Bu noktada rasgele orman algoritması gini bölünme yöntemine göre en yüksek doğruluk performansı göstermesi nedeniyle tez çalışmasında tercih edilebilir kılmaktadır.

Makine öğrenmesine ilişkin tüm performans iyileştirme teknikleri birlikte değerlendirildiğinde indirgenmemiş yığın veri setinin %80 eğitim-%20 test olarak

ayrılarak rasgele orman algoritması ile gini bölünme yöntemine gerçekleştirilen daha yüksek bir model performansı elde edilmiş nihai model olarak seçilmiştir.

Nihai modele ulaştıktan sonra yapılması gereken bir veri üzerinden nihai model ile sınıflandırma tahmin işlemi gerçekleştirilerek çalışma probleminin çözümü için gerekli olan bilgiye ulaşılmıştır. Nihai model veri setinde yer alan ilk 10 işsiz için sınanmış ve doğru ve yanlış sınıflandırmaları gösterilmiştir.

Tablo 3.46 İndirgenmemiş Yığın Veri Seti İçin %80 Eğitim-%20 Test Model Geçerleme Yöntemine Göre Rasgele Orman Algoritması İle Gini Bölünme Yöntemine Göre Oluşturulan Makine Öğrenmesi Modelinin Veri Setinde Yer Alan İlk 10 İşsiz İçin Sınanması

Sıra	Cinsiyet	Yaş	Sosyal Durum	Başvuru Türü	Meslek Grup	Medeni Durum	İşsizlik Maaşı Alma Durumu	Öğrenim	İkamet		İŞE YERLEŞME DURUMU	İŞE YERLEŞME TAHMİNİ	DOĞRULUK SINIFLANDIRMASI
1	2	3	2	2	4	1	0	5	2		0	1	Yanlış Sınıflandırma
2	1	1	2	2	6	1	0	3	2		1	1	Doğru Sınıflandırma
3	2	4	2	2	5	4	0	5	2		0	0	Doğru Sınıflandırma
4	2	2	2	2	6	1	0	3	2		0	0	Doğru Sınıflandırma
5	1	1	2	1	3	1	0	3	2		0	0	Doğru Sınıflandırma
6	2	5	2	1	3	4	0	2	2		0	0	Doğru Sınıflandırma
7	2	6	2	1	3	4	0	2	2		1	0	Yanlış Sınıflandırma
8	1	3	2	2	3	1	0	3	1		0	1	Yanlış Sınıflandırma
9	1	7	2	2	2	4	0	2	2		0	0	Doğru Sınıflandırma
10	1	4	2	2	5	4	0	3	1		1	1	Doğru Sınıflandırma

3.12 İstatistiksel Testler İle Çalışma Varsayımlarının Test Edilmesi

Çalışmanın iki ana varsayımdan oluştuğunu birinci bölümde belirtilmiştir. Bu varsayımlardan birincisi; İŞKUR'a başvuran işsizlerin, başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmeme durumu ile işsizin kişisel, demografik ve işgücüsel bilgileri arasında istatistiksel anlamlı bir ilişkinin olup olmadığı, ikincisi; İŞKUR'a başvuran işsizlerin, başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmeme doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir farkın olup olmadığı;

Birinci varsayıma ilişkin hipotez aşağıdaki şekilde oluşturulmuştur.

H_{10}=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki yoktur.

H_{11}=İşsizin bilgileri(cinsiyet, yaş vb.) ile işe yerleşmesi arasında bir ilişki vardır.

Tablo 3.47 Ki-kare Analizi Sonuçlarına Göre İşe Yerleşme ile İşsizlerin Özellikleri Arasındaki İlişkiler

ÖZNİTELİKLER	Test istatistiği	S. d	P Değeri	İlişki Varlığı	İlişki Gücü	İlişki Derecesi
İkamet	26,15	1	<0.001	Var	0,043	Çok Zayıf
Sosyal Durum	75,47	1	<0.001	Var	0,0728	Çok Zayıf
Cinsiyet	121,91	1	<0.001	Var	0,0924	Çok Zayıf
Yaş1	439,93	6	<0.001	Var	0,1737	Çok Zayıf
Medeni Durum1	388,64	3	<0.001	Var	0,1635	Çok Zayıf
Sosyal Yardım Alma Durumu	5,94	1	0.015	Var	0,0205	Çok Zayıf
Öğrenim	81,21	4	<0.001	Var	0,0756	Çok Zayıf
Meslek	378,04	6	<0.001	Var	0,1613	Çok Zayıf
Başvuru Türü	68,32	1	<0.001	Var	0,0693	Çok Zayıf
İşsizlik ödeneği Alma Durumu	27,25	1	<0.001	Var	0,0439	Çok Zayıf

Ki-kare analiz sonuçlarına göre işsizin işe yerleşmesi durumu ile işsizin tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel olarak anlamlı bir ilişki bulunmaktadır[30]. Kontenjans katsayısı ile ilişki gücü[31] incelendiğinde işsizin işe yerleşmesi ile tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. Ancak zayıf ilişkiye sahip bu özniteliklerin bir araya getirilmesi daha güçlü bir yapı/model oluşturabilmektedir. Analiz sonuçları İŞKUR'a başvuran işsizin, başvuru sonrası bir yıl içerisinde işe yerleşmesinde temel düzeyde kişisel, demografik ve iş arama bilgilerinin çok zayıfta olsa bir etkisi olduğunu göstermektedir.

Hâlihazırda İŞKUR iş arayan kayıt bilgileri de söz konusu temel bilgilerden oluşmaktadır. Bu durum İŞKUR'un temel işsiz bilgilerinden işsizlerin zaman bağlı olarak işsiz kalma risklerinin tahmin edilmesine ve iş profillemesine fırsat sunmaktadır. Burada önemli olan husus ise risk değerlendirilmesinin iyi yapılmasıdır. Risk değerlendirmesinin iyi yapılması ise açıklama oranı ve doğru sınıflama oranı yüksek performanslı modellemeler ile ancak mümkündür. Bu performanslı modellerin

[30] İşsizin sosyal yardım alma durumu özniteliğinin ilişki derecesinin ve test istatistiğinin düşük olması ve bu özniteliği ait veri sayısının az olması nedeniyle bu özniteliğe mesafeli yaklaşılmıştır.

[31] **Pearson'un Kontenjans Katsayısı (Contingency Coefficient):** Kontenjans katsayısı, katsayısının IxJ boyutlu tablolardaki iki değişken arasındaki ilişkinin büyüklüğünü ölçen biçimidir [48].

oluşturulması için hâlihazırda klasik çok değişkenli istatistiksel yöntemler ile daha modern makine öğrenmesi algoritmaları kullanılmaktadır. Ancak bu yöntemlerin hangisi/hangilerinin kullanılacağı modellerin performans göstergeleri ve birbirlerine olan üstünlüklerinin tespiti ile mümkündür. Tez çalışmasının ikinci varsayımı ile yöntemlerin model performansları test edilmiştir.

İkinci varsayıma ilişkin hipotez aşağıdaki şekilde oluşturulmuştur.

H_{20}=İşsizin işe yerleşmesinin doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir fark yoktur.
H_{21}=İşsizin işe yerleşmesinin doğru sınıflandırılmasında çok değişkenli istatistiksel yöntemler ile yapay zekâ makine öğrenmesi yöntemleri arasında istatistiksel anlamlı bir fark vardır.
Her ne kadar konuyla ilgili kaynaklarda lojistik regresyon modellemesi ve naive bayes algoritması bir makine öğrenmesi algoritması olarak gösterilmiş olsa da bu algoritmalar esasında istatistik kökenli olup bir takım olasılık ve istatistik hesaplar ile oluşturulmuşlardır. Ancak bu çalışmada sadece lojistik regresyon algoritması ile diğer makine öğrenmesi algoritmalarının doğruluk sınıflandırması yönünden aralarında istatistiksel bir fark olup olmadığı test edilmiştir.

Çok değişkenli istatistiksel modelleme yöntemi olan lojistik regresyon algoritması ile diğer makine öğrenmesi algoritması algoritmalarının doğru ve yanlış sınıflandırma sayıları ve ki-kare analizi sonuçlarına yer verilmiştir. Söz konusu tablo indirgenmemiş örneklem veri setinin %80 Eğitim-%20 Test model eğitimi ile algoritmaların doğruluk performanslarına göre oluşturulmuştur.

Tablo 3.48 Lojistik Regresyon ve Makine Öğrenmesi Algoritmalarının Sınıflandırma Tablosu ve Ki-kara Analizi Sonuçları

	Doğru Sınıflama	**Yanlış Sınıflama**	**Test İstatistiği**	**S.d**	**P Değeri**
Lojistik Regresyon	1719	1111			
Rasgele Orman	2234	596	224,7	1	<0.001
Naive Bayes	1653	1177	2,9	1	0,089
Karar Ağaçları	2209	621	201,9	1	<0.001
Destek Vektör Makinaları	1677	1153	1,1	1	0,292
Yapay Sinir Ağları	1940	890	38,8	1	<0.001
K En yakın Komşu	2068	763	98,4	1	<0.001

Lojistik regresyon algoritması ile makine öğrenmesi algoritmalarının doğru sınıflandırma sayıları arasında matematiksel farklar olduğu tespit edilmiştir. Oluşan bu matematiksel farkın istatistiksel olarak anlamlı olup olmadığı ki-kare analiz ile test edilmiştir. Ki-kare analizi sonuçlarına göre p değeri 0,05'ten küçük olan rasgele orman, karar ağaçları, yapay sinir ağları ve k en yakın komşu algoritmaları ile lojistik regresyon algoritması arasında doğruluk sınıflandırması yönünde %95 güven düzeyinde istatistiksel olarak anlamlı fark bulunmakta olup H_{20} hipotezi reddedilir. P değeri 0,05'ten büyük olan naive bayes ve destek vektör makinaları ile lojistik regresyon algoritması arasında doğruluk sınıflandırması yönünde %95 güven düzeyinde istatistiksel olarak anlamlı fark bulunmamakta olup H_0 hipotezi reddedilmez.

Doğruluk sınıflandırması yönünden Lojistik regresyon algoritması ile arasında belirgin bir fark olmayan naive bayes ve destek vektör makinaları algoritmaları yaklaşık %60 doğruluk performansı sergilerken, doğruluk sınıflandırması yönünden arasında belirgin fark olan k en yakın komşu ve yapay sinir ağları algoritması yaklaşık %70, rasgele orman ve karar ağaçları algoritması ise yaklaşık %80 doğruluk performansı sergilemektedir. Bu sonuçlara göre makine öğrenmesi algoritmalarının çoğunluğunun çok değişkenli lojistik regresyon analizine göre daha yüksek bir doğruluk performansı gösterdiği tespit edilmiştir.

3.13 İşe Yerleştirmede Etkin Faktörlere İlişkin Değerlendirme

Çalışmanın birinci bölümünde iş arayanın bir işe yerleşerek istihdam edilmesinde birtakım ana faktörler ve bu ana faktörlerin içerisinde bulunan birçok alt faktör belirleyici ve etkileyici konumda olduğu ve istihdamın gerçekleşmesinde iş arayan, açık iş ve işgücü piyasası şeklinde üç ana faktör olduğu belirtilmiştir. Tez çalışmasında İŞKUR'a başvuran işsizlerin başvuru sonrası bir yıl içerisinde işe yerleşmesi/yerleşmemesi işsize ait kişisel, demografik ve iş arama bilgileri üzerinden bir tespit ve modelleme işlemi gerçekleştirilmiştir. Başka bir ifadeyle işsizlerin işe yerleşmesinde iş arayan ana faktörü ve ana faktörü belirleyen bir takım alt faktörler üzerinden değerlendirme yapılmıştır. Bu bağlam da iş arayanın ikamet, cinsiyet, sosyal durum, yaş, öğrenim, meslek, medeni durum, başvuru türü, işsizlik ödeneği alma durumu gibi temel İŞKUR iş arayan kayıt bilgileri üzerinden bir sınıflama ve modelleme işlemi yapılmış olup işsizin işe olan ihtiyacı, iş arama ve çalışma isteği, iş

arama donanımı ve iş arama davranışı üzerinden bir sınıflama ve modelleme yapılmamıştır.

İşsizin temel İŞKUR iş arayan kayıt işlemleri üzerinden sınıflama ve modelleme işlemi öncelikle klasik çok değişkenli lojistik regresyon analizi ile gerçekleştirilmiş ve yaklaşık %59 oranında bir doğru sınıflama oranı ve yaklaşık %5 civarında bir açıklama oranı elde dilmiştir. Başka bir ifadeyle lojistik regresyon modellemesi bize orta düzeyde bir doğru sınıflama oranı ve çok düşük bir açıklama oranı sunmuştur. Bu durum işsizin işe yerleşmesini yaklaşık %95 oranında açıklayan başkaca ana ve alt faktörlerin olduğunu göstermektedir. Bu ana faktörlerin açık iş ve işgücü piyasası ana faktörleri olduğu alt faktörlerin ise iş arayan ana faktörü altındaki işe olan ihtiyacı, iş arama ve çalışma isteği, iş arama donanımı ve iş arama davranışı alt faktörleri olduğu anlaşılmaktadır.

Lojistik regresyon modellemesinin orta düzeyde bir doğru sınıflama göstermesi başkaca sınıflama tekniklerinin kullanımını da zorunlu kılmıştır. Bu noktada yapay zekâ makine öğrenmesi teknikleri kullanılarak doğru sınıflama oranı yaklaşık %80 oranına kadar yükseltilmiştir. Her ne kadar makine öğrenmesi teknikleri ile işe yerleşmede doğru sınıflama oranı yükseltilmiş olsa da açıklama oranında bir değişim olmamıştır. Bu noktada çok düşük bir seviyede de olsa model açıklama oranı bizlere büyük bir bilgi sunmaktadır. Diğer bir ifadeyle modelin ideal açıklama oranına ulaşması için modelde olması gereken ana ve alt faktörlere yüksek işaretler göndermektedir. Bununla birlikte %5 açıklama oranını meydana getiren ikamet, cinsiyet, sosyal durum, yaş, öğrenim, meslek, medeni durum, başvuru türü, işsizlik ödeneği alma durumu gibi temel İŞKUR iş arayan kayıt bilgilerinden hangisinin modele en yüksek katkı yaptığını bilmek işe yerleşmede etkin alt faktörlerin bilinmesine ve iş arayan danışmanlık içeriğinin belirlenmesine büyük katkı yapacaktır.

Model içerisinde yer alan özniteliklerin modele ne ölçüde katkı sağladıklarını bulmak için özniteliklerin modeldeki standartlaştırılmış regresyon katsayılarına ve ODDS[32] değerlerine bakmak gerekmektedir. Model öncelikle ana ve sonrasında öznitelik çıkartma işlemi ile amaçlanan model şeklinde oluşturulmuştur. Dolayısıyla her iki model için standart regresyon katsayıları ve ODDS üzerinden özniteliklerin modele katkısı irdelenmiştir.

[32] ODDS; Bir olayın olma olasılığının olmama olasılığına oranı.

Tablo 3.49 Lojistik Regresyon Modelinde Yer Alan Özniteliklerin Standardize Edilmiş Regresyon Katsayıları

	Parametreler	**Standardize Edilmiş Regresyon Katsayıları**	**ODDS**
1.Ana Model	Const(Sabit Değer)	-0,007	
	Cinsiyet	-0,02	Referans
	Sosyal Durum	-0,209	10,5
	Yaş1	-0,296	14,8
	Medeni Durum2	-0,146	7,3
	Öğrenim	0,233	11,7
	Meslek	0,187	9,4
	Başvuru Türü	-0,257	12,9
	İkamet	0,083	4,2
	İşsizlik Ödeneği Alma Durumu	-0,18	9,0
2.Amaçlanan Model	Const(Sabit Değer)	-0,007	
	Sosyal Durum	-0,212	2,6
	Yaş1	-0,298	3,7
	Medeni Durum2	-0,15	1,9
	Öğrenim	0,237	2,9
	Meslek	0,189	2,3
	Başvuru Türü	-0,261	3,2
	İkamet	0,081	Referans
	İşsizlik Ödeneği Alma Durumu	-0,182	2,2

Özniteliklerin standardize edilmiş regresyon katsayılarına göre oluşturulmuş standart ana modele göre modele pozitif katkı 0,233 oranla öğrenim, 0,187 meslek, 0,083 ile ikamet öznitelikleri tarafından olmuştur. Modele negatif katkı ise -0,296 oranla yaş1, -0,257 oranla başvuru türü, -0,209 oranla sosyal durum, -0,18 oranla işsizlik ödeneği alma durumu, -0,146 oranla medeni durum2 ve -0,02 oranla cinsiyet öznitelikleri tarafından olmuştur. Burada pozitif katkıları işsizin işe yerleşme ihtimalini arttıran negatif katkılar ise işe yerleşme ihtimalini azaltan özniteliklerdir. Diğer bir ifadeyle işsizin öğrenim ve meslek bilgisi üzerinden sadece değerlendirildiğinde işe yerleşme ihtimali artarken, işsizin yaşı, başvuru türü, sosyal durumu ve tecrübesi(işsizlik ödeneği alam durumu) hesaba katıldığında artan işe yerleşme ihtimali düşmektedir. Burada işsizin işe yerleşme ihtimalini en az etkileyen öznitelik cinsiyettir. Hâlihazırda cinsiyet özniteliği modelden çıkartılmıştır. Cinsiyet özniteliğine göre yaş değişkeni yaklaşık 15 kat, başvuru türü 13 kat, sosyal durum 11 kat, meslek 9 kat, işsizlik ödeneği alma durumu(tecrübe) 9 kat, medeni durum 7 kat ve ikamet 4 kat daha fazla

işsizin işe yerleşmesinde etki etmektedir. Amaçlanan modelle cinsiyet özniteliği çıkartılıp ikamet özniteliği referans kabul edilip ODDS oranları hesaplandığında modele katkı yapan öznitelikler arasındaki fark korunarak cinsiyet özniteliği kaynaklı yüksek katkı değerleri ortadan kalkmıştır.

ODDS değerlerine göre işsizin işe yerleşme ihtimalinin tahmin edilmesinde işsizin yaşı, başvuru türü, öğrenimi, sosyal durumu, mesleği, medeni durumu, işsizlik ödeneği alma durumu, ikameti ve cinsiyeti sıralı bir şekilde en çoktan en aza şeklinde katkı sağlamaktadır. Bu bilgi işsizin işe yerleşmesinde etkin belirleyicileri de ortay koymaktadır. Bu belirleyiciler başta işsizin yaşı, sonra başvuru türü(çalışırken işsiz kalan/ilk kez iş arayan, öğrenimi, iş arama kişisel durumu(engelli/normal) ve mesleğidir. Esasında bu bilgiler işsizin iş arama temel bilgileridir. İşsizin medeni durumu, ikameti ve cinsiyeti ise işsizin iş arama dışındaki kişisel bilgileridir. Yine işsizin işsizlik ödeneği alma durumuna ait bilgi ise yine doğrudan iş arama dışı bir bilgidir. Standart regresyon katsayıları ve ODDS değerleri üzerinden tespit edilen bu bilgiler, iş arayan danışmanlık hizmetinde danışmanın nitelikli bir danışmanlık sunumu için işsizin hangi özniteliklerine daha fazla önem vermesi gerekliliğini de ortaya koymaktadır. Burada dikkat edilmesi gereken en önemli husus ise işsizin işe yerleşmesinde model dışında yer alan başkaca etkin ana ve alt faktörlerin olduğu ve bu faktörlerin oldukça yüksek bir açıklama oranına sahip olduğudur.

3.14 İşe Yerleşme Olasılıklarına Göre Risk Gruplarının Belirlenmesi

İşsizlerin risk gruplarının belirlenmesinde iki yöntem kullanılabilir. Bunlardan biri sınıflama algoritmalarına göre işsizin işe yerleşmesi(y=1) veya işe yerleşmemesi(y=0) şekline göre "işsiz kalma riski taşıyanlar/işsiz kalma riski taşımayanlar" şeklinde iki bir sınıflamadır. Bu sınıflama oldukça önemli ve temel bir sınıflandırma olmasına karşı keskin sınırlar içermesi nedeniyle bazı dezavantajlar içermekte olup işsizin danışmanlık ihtiyacının seviyelerinin belirlenmesine sınırlı bir katkı sağlayacaktır. Risk gruplarının belirlenmesinde bir diğer yöntem ise işsizin işsiz kalma olasılıklarına göre "Çok Yüksek Riskli, Yüksek Riskli, Riskli, Düşük Riskli ve Çok Düşük Riskli" türünde bir likert ölçeği sınıflamasına tabi tutulmasıdır. Likert ölçeği şeklinde beşli bir işsiz kalma risk sınıflandırması ile işsizin risk grupları ve risk gruplarına göre işsizin iş arayan danışmanlığına olan ihtiyacı daha net ortaya koyulmuş olacaktır.

Çok değişkenli lojistik regresyon modellemesi her ne kadar orta düzeyde bir doğru sınıflama oranı verse de işsizin İŞKUR'a başvuru süresi sonrası bir yıl içerisinde işsiz kalma ihtimaline ilişkin önemli bir olasılık değeri türeterek işsizin likert ölçeği şeklinde işsiz kalma risk grubunun belirlenmesinde önemli yer tutmaktadır. Ancak doğru sınıflama oranın istenilen seviyede olmaması bir sorun olarak çözüm beklemektedir. Bu sorunun çözümünde iki yöntem bulunmaktadır. Bunlardan biri yüksek bir doğruluk oranı imkânı sunan rasgele orman algoritmaları ile lojistik regresyon algoritmasının birlikte kullanılması, diğeri rasgele orman algoritmasının regresyon tipinin kullanılması ve en son olarak lojistik regresyon algoritmasının doğru sınıflama oranın arttırılmasıdır. Doğru sınıflama oranın arttırılması için ise modele yeni öznitelikler eklenmesi gerekmektedir. Bu noktada rasgele orman algoritmasının regresyon şeklinin uygulanması hem yüksek bir doğru sınıflandırma ve açıklama oranı hem de işsizin işe yerleşme/işsiz kalma olasılarının elde edilmesine ve işsiz kalma risk sınıflandırması ile işsizin risk gruplarının belirlenmesine imkân sunmaktadır.

Hem çok değişkenli lojistik regresyon hem de rasgele orman algoritması 0,50'nin altındaki değerleri sıfır(işe yerleşmedi) ve 0,50'nin üzerindeki değerleri ise bir(işe yerleşti) şeklinde bir sınıflama yapmaktadır. Ancak bu sınıflama yöntemi çok keskin olup örneğin 0,49 değerinde işe yerleşme olasılığına sahip işsizleri doğrudan işe yerleşmedi olarak atamaktadır. Yine 0,60 değerinde bir işe yerleşme olasılığına sahip işsizleri de doğrudan işe yerleşmedi olarak atamaktadır. Bu durum ise doğru sınıflandırma oranını doğrudan düşürmektedir.

Burada özellikle 0,40-0,60 aralığında işe yerleşme olasılığına sahip işsizlerin işe yerleşme/yerleşmeme durumu kritik bir önem taşımaktadır. Başka bir ifadeyle 0,40-0,60 aralığında işe yerleşme olasılığına sahip işsizler işe yerleşedebilir yerleşemeyedebilr. Bu çıkarımı bu aralıkta işe yerleşen ve yerleşmeyen işsiz oranları ve ODDS oranları desteklemektedir. Dolayısıyla bu aralık için kesin ifadelerle işe yerleşti/yerleşmedi şeklinde bir sınıflamadan kaçınılarak "riskli" düzeyi kullanılması ve diğer risk gruplarının bu temel üzerine bina edilmesi risk gruplarının belirlenmesi doğru sınıflama oranı kaynaklı sorunu büyük ölçüde çözecektir.

Tablo 3.50 Rasgele Orman Regresyon Sonuçları

Göstergeler	Oran
Tahmin Edilen İşe Yerleşme İle Gerçek İşe Yerleşme Arasındaki Korelasyon(r)	0,69
Tahmin Edilen İşe Yerleşmenin Gerçek İşe Yerleşmeyi Açıklama Oranı($\boldsymbol{R^2}$)	0,48
Doğru Sınıflama Oranı	0,80
Ortalama Kare Hatası(Mean Squared Error(MES))	0,13
Ortalama Mutlak Hata(Mean Absolute Error(MAE))	0,26

Rasgele orman regresyon sonuçlarına göre tahmin edilen işe yerleşme olasılıkları ile gerçek işe yerleşme olasılıkları arasında 0,69 oranında yüksek bir doğrusal ilişki bulunmaktadır. Oluşturulan rasgele orman regresyon modelinin açıklama oranı 0,48 ile oldukça iyi durumdadır. Ayrıca hem mutlak hata hem de ortalama hata karesi düşük seviyededir. Bununla birlikte yaklaşık 0,80 oranında yüksek seviyede bir doğru sınıflama oranı elde edilmiştir Rasgele orman algoritması regresyon modeli ile hem işe yerleşme/işsiz kalma olasılıkları hesaplanmış, hem yüksek bir açıklama oranı elde edilmiş ve hem de yüksek bir doğru sınıflama performansı sağlanmıştır. Bu haliyle rasgele orman algoritması tez çalışması için en uygun ve en iyi model olarak değerlendirilebilir.

Bu tespitlerden sonra sırada işe yerleşme/işsiz kalma olasılıklarına göre işsizlerin beşli likert ölçeğinde işsiz kalma risk guruplarının belirlenmesi işlemi gelmektedir. İşsizlerin risk gruplarının belirlenmesinde öncelikle doğru bir işsiz kalma sınıflamasının sağlanması gerekmektedir. Bu sınıflandırmanın yapılmasında daha önce belirtildiği üzere işsizlerin işe yerleşme/yerleşmeme belirsizliğinin en yüksek olduğu 0,40-0,60 oransal aralığı temel alınarak bu aralıkta işe yerleşme ihtimaline sahip işsizler "riskli" olarak sınıflandırılmıştır. Bu temel üzerine diğer risk aralıkları ve sınıfları oluşturulmuştur.

Tablo 3.51 İşsiz Kalma Risk Sınıflaması

Risk Puan	İşe Yerleşme Olasılığı	Risk Sınıflaması	Açıklama
5	0,00-0,20	Çok Yüksek Riskli	Çok yüksek olasılıkla işe yerleşmeme
4	0,20-0,40	Yüksek Riskli	İşe yerleşmeme olasılığı daha yüksek
3	0,40-0,60	Riskli	İşe yerleşme/yerleşmeme belirsizliği
2	0,60-0,80	Düşük Riskli	İşe yerleşme olasılığı daha yüksek
1	0,80-1,00	Çok Düşük Riskli	Çok yüksek olasılıkla işe yerleşme

İşsiz kalma risk sınıflaması beşli likert ölçeği şeklinde oluşturulmuş olup sınıflar arasında eşit oranlı bir artış/azalış bulunmaktadır. Bu sınıflandırma ile her bir işsiz için bir işsiz kalma risk puanı ve risk sınıfı belirlenmiştir. Risk sınıflandırmasına göre İŞKUR'a ilgili dönemde kayıt yaptıran her bir işsiz için işe yerleşme olasılığı, risk grubu ve risk puanı belirlenmiştir.

Tablo 3.52 İşsizlerin Rasgele Orman Algoritması Regresyon Modellemesine Göre İşe Yerleşme Olasılığı, Risk Grubu ve Risk Sınıfı(Örnek Tablo)

Sıra	Cinsiyet	Sosyal Durum	Yaş	Öğrenim	Başvuru Durumu	Meslek Grup	Medeni Durum	İkamet	İşsizlik Maaşı Alma Durumu	İşe Yerleşme Olasılıkları	Risk Grubu	Risk Puanı
1	2	2	3	5	2	4	1	2	0	0,65	Düşük Riskli	2
2	2	2	4	5	2	5	4	2	0	0,00	Çok Y. Riskli	5
3	2	2	2	3	2	6	1	2	0	0,00	Çok Y. Riskli	5
4	1	2	1	3	1	3	1	2	0	0,39	Yüksek Riskli	4
5	2	2	5	2	1	3	4	2	0	0,27	Yüksek Riskli	4
6	1	2	3	3	2	3	1	1	0	0,51	Riskli	3
7	1	2	7	2	2	2	4	2	0	0,00	Çok Y. Riskli	5
8	1	2	1	3	1	3	1	2	0	0,39	Yüksek Riskli	4
9	2	2	7	2	2	3	4	2	0	0,00	Çok Y. Riskli	5
10	1	1	4	3	1	3	1	2	0	0,53	Riskli	3
11	1	2	2	3	1	6	1	2	0	0,44	Riskli	3
12	2	2	5	2	2	3	4	2	0	0,00	Çok Y. Riskli	5
13	2	2	5	2	1	3	2	2	0	0,00	Çok Y. Riskli	5
14	1	2	6	2	2	3	4	2	0	0,47	Riskli	3
15	2	2	5	2	2	5	4	1	0	0,00	Çok Y. Riskli	5
16	2	2	6	2	1	7	4	2	0	0,00	Çok Y. Riskli	5
17	2	2	4	2	1	3	4	2	0	0,00	Çok Y. Riskli	5
18	2	2	3	5	2	4	1	2	0	0,65	Düşük Riskli	2
19	2	2	3	4	2	3	4	2	0	0,00	Çok Y. Riskli	5

Oluşturulan örnek tabloda yer alan sonuçlar incelendiğinde işe yerleşme olasılığı sıfır olan işsizler dikkat çekmektedir. Bununla birlikte tabloda gösterilmemiş olmakla bazı işsizlerin işe yerleşme olasılığı bir olarak hesaplanmıştır. Hayatın olağan akışında bir işsizin bir işgücü piyasasında işe yerleşme ihtimalinin sıfır ve bir olması imkânsız bir durumdur. Böyle bir sonucun çıkması rasgele orman algoritmasının yapısından kaynaklandığı ve çözülmesi gereken bir sorun olduğu değerlendirilmektedir. Bu

sorunun çözümünde çok değişkenli doğrusal regresyon ile rasgele orman regresyona göre işe yerleşme ihtimali sıfır ve bir olan işsizlerin işe yerleşme ihtimalleri hesaplanmıştır. Hesaplamalara göre doğrusal regresyon modeli ile elde edilen işe yerleşme ihtimalinin sıfır ve bir gibi mutlak değerlerden uzak daha makul değerler ortaya koyduğu tespit edilmiştir. Ancak çoklu doğrusal regresyon sınıflamasının rasgele orman regresyona göre daha düşük doğru sınıflama performansı göstermesi çoklu doğrusal regresyonu doğrudan kullanımına engel teşkil etmektedir. Bu engelin aşılmasında rasgele orman regresyon ve çoklu doğrusal/lojistik regresyonun birlikte kullanılabilir. Ancak tez çalışmasında sadece rasgele orman regresyon algoritması sonuçlarına göre risk değerlendirmesi yapılmıştır.

Tablo 3.53 İşsizlerin Risk Gruplarına Göre Sınıflandırma Oranı

Risk Grubu	İşsiz Sayısı	Oran
Çok Yüksek Riskli	3.682	48,7%
Yüksek Riskli	871	11,5%
Riskli	1.245	16,5%
Düşük Riskli	1.280	16,9%
Çok Düşük Riskli	488	6,4%
	7.566	100,0%

İşsizlerin risk gruplarına göre risk sınıflandırma oranına bakıldığında Sivas ili işgücü piyasasında İŞKUR ilgili dönemde kayıt yaptırıp işsizlerin %48,7'si çok riskli grupta yer alarak çok yüksek bir olasılıkla başvuru sonrası bir yıl içerisinde işsiz kalmaktadır. Başka bir ifadeyle İŞKUR'a kayıt yaptıran her iki işsizden biri Sivas ilinde başvuru sonrası bir yıl içerisinde işe yerleşememektedir. Başvuru sonrası bir yıl içerisinde çok yüksek bir olasılıkla işe yerleşen işsizlerin oranı çok düşük olup %6,4'tür. İşe yerleşip yerleşmeyeceği riskini beraber taşıyan işsizlerin oranı %16,5'tir. Riskli, yüksek riskli ve çok yüksek riskli işsizler birlikte değerlendirildiğinde işsiz kalma riski taşıyan işsizlerin oranı %76,6'ya çıkmakta olup bu oran çok yüksek seviyededir. Diğer bir ifadeyle Sivas ili işgücü piyasasında İŞKUR'a ilgili dönemde kayıt yaptırarak iş arayan her dört işsizden ancak biri işsiz kalma riski taşımamaktadır. Esasında ilgili döneme ait bu sonuç Sivas ili işgücü piyasasın istihdam edilebilirliği hakkında önemli sonuçlar içermektedir. Ancak unutulmaması gereken husus bu sonuçların İŞKUR'a başvuru yapan işsizleri ait olduğu ve İŞKUR dışındaki iş arama kanalları ile iş arayan işsizlere ait olmadığıdır.

Tablo 3.54 İşsizlerin Genel Risk Göstergeleri

Genel Risk Göstergeleri	
İşsiz Kalma Risk Düzeyi	Yüksek Riskli
İşsiz Kalma Risk Oranı	0,76
İşsiz Kalama Risk Puanı(Ortalama)	3,79
İşe Yerleşme Oranı	0,06
İşe Yerleşme Olasılığı(Ortalama)	0,30

İşsizlere ait genel risk göstergelerine göre Sivas ili işgücü piyasasında İŞKUR aracılığıyla ilgili dönemde işsizlerin başvuru sonrası ortalama işe yerleşme olasılığı 0,30 olup düşük seviyededir. Düşük seviyedeki işe yerleşme olasılığı 3,79 risk puanı ile yüksek seviyede bir işsiz kalma riski şeklinde tezahür etmiştir. İşsizlerin genel olarak işsiz kalma risk oranı %76 olup bu oran yüksek düzeyde işsiz kalma riskine tekabül etmektedir. Bu bilgiler Sivas ilinde İŞKUR aracılıyla iş arayan işsizlerin işe yerleşme güçlüğünü ortaya koymaktadır. Hâlihazırda çok düşük seviyedeki %6 işe yerleşme oranı tespit edilen işe yerleşme güçlüğünü teyit etmektedir. Bu güçlüğün aşılmasında işsizlerin işsiz kalma risk tespitlerinin doğruluk performansı yüksek yöntemlerle yapılması gerekliliği bir kez daha ortaya çıkmaktadır.

4. TARTIŞMA VE SONUÇ

İşsizlik sebepleri, sonuçları ve etkileri itibariyle çok boyutlu ve karmaşık bir yapı arz etmektedir. İşsizlik sorunun çözülmesi ile birçok alanda fayda sağlayacaktır. İşsizlik sorunun çözümü ise istihdamdan geçmektedir. İstihdamın arttırılması ve yeni istihdam alanların oluşturulması ise makro ve mikro istihdam politikaları uygulanmaktadır. Makroekonomik istihdam politikalar ile makro göstergeler üzerinden işsizlik sorunu incelerekken mikro ekonomik istihdam politikaları ile mikro düzeyde araçlarla soruna çözüm aranır. Mikro ekonomik istihdam politikaları aktif ve pasif istihdam politikalarıdır. Aktif istihdam politikalarının ana gövdesini ise danışmanlık hizmetleri oluşturmaktadır.

İŞKUR iş arayan danışmanlığı ile işsizlerin iş arama süreçlerinin yönetimi ve yürütümü gerçekleştirilerek işsizler işe ve mesleğe yönlendirilmekte ve işsizlerin iş arama becerilerinin gelişimi ve işgücü piyasası hakkında bilgilendirme ana faaliyetleri yürütülmektir. İş arayan danışmanlığının etkin, verimli ve nitelikli olarak gerçekleştirilmesi bireysel eylem planı dâhilinde gerçekleştirilmektedir. Amaca uygun ve danışmanlık hizmetlerinin hedeflerinin gerçekleştirilmesine imkân tanıyan iyi bir eylem planın hazırlanması ise danışmanın işsizi kişisel, demografik ve iş arama özelliklerine göre işsiz kalma risklerinin meydana çıkarılması ile başka bir deyişle iş profilleme ile mümkündür. İş profilleme işlemi sadece danışman tarafından gerçekleştirildiğinde ise kimi zaman danışman tarafından kimi zamanda işsiz tarafından sübjektif nedenler(eksik tanıma/tanıtma vb.) dolayısıyla iyi bir şekilde gerçekleştirilmemektedir. Bu noktada iyi bir iş profilleme işleminin gerçekleştirilmesi iş arayanların iş bulma olasılıklarına göre işsiz kalma riskinin tespit edilmesi diğer bir ifadeyle istatistiksel profilleme işlemi gerçekleştirilmelidir.

İstatistiksel profilleme de hedef değişken, işsizin belirli bir zaman diliminde işe yerleşme/yerleşmemesidir. Bugün dünyada başta OECD ülkeleri olmak üzere birçok ülkede işsizlerin profillemesi işlemi istatistiksel yöntemlerle ağırlıkla lojistik regresyon analizi ile gerçekleştirilmektedir. Bilim ve teknikte gerçekleştirilen ilerlemeler profilleme işleminin yeni tekniklerle yapılmasına fırsat tanımaktadır. Özellikle yapay zekâ ve makine öğrenmesi alanındaki gelişmeler makine öğrenmesi algoritmalarını profilleme işleminde istatistiksel yöntemlere alternatif hale getirmektedir. Günümüzde bazı ülkeler rasgele orman makine öğrenmesi algoritması

ile işsiz/iş arayan profilleme işlemini gerçekleştirmektedir. Bu iki ana yöntemin profilleme işleminde birbirlerine olan üstünlükleri ve tercih sebeplerinin ortaya konulması ise bir soru olarak karşımızda durmaktadır.

Bu çalışmada 2022 yılı Eylül, Ekim ve Kasım aylarında Sivas İŞKUR'a iş için başvuran işsizlerin kişisel, demografik ve işgücüsel bazı bilgileri üzerinden başvuru sonrası bir yıl içerisinde işe yerleşme/yerleşmemeleri klasik istatistiksel yöntemler ve yapay zekâ makine öğrenmesi yöntemleri ile modellenmesi, sınıflandırılması, yöntemlerin karşılaştırılması, işsiz kalma risk sınıflarının ve işe yerleşmede etkin değişkenlerin belirlenmesi gerçekleştirilmiştir. Çalışmanın iki ana sorusu/problemi bulunmaktadır, Bunlardan birincisi; "İşsizlerin zamana göre işsiz kalma risklerinin tespitinde yapay zekâ makine öğrenmesi yöntemleri, istatistiksel yöntemlere göre daha mı etkili?" olduğu ikincisi ise "İşsizlerin işe yerleşmesinde etkin faktörlerin/değişkenlerin neler olduğu?" dur. Ayrıca çalışmanın ana soruları altında makine öğrenmesi algoritmaların, model eğitim yöntemleri ve veri setine göre model performansları da irdelenmiştir. İŞKUR verilerine göre Sivas ilinde 2023 yılı itibariyle yaklaşık 25 bin işsiz, %3 oranında açık iş bulunmakta olup il işsizliğinin önemsenecek bir yapıda olduğu ve çözümünde yeni yöntemler ve tekniklerin kullanılması gerekliliğini ortaya koymaktadır.

Çalışması kapsamında veri madenciliği ile veri setinin modellenmeye uygun hale getirilmesi, çok değişkenli istatistiksel yöntem ile modellenmenin ve sınıflandırılmanın sağlanması, makine öğrenmesi algoritması ile sınıflandırmaların sağlanması, teknik, yöntem ve algoritmaların kıyas edilmesi ve varsayımların test edilmesi, işe yerleşmede etkin ana ve alt faktörlerin tespiti ve işsiz kalma olasılığı ve bu olasılığa göre işsiz risk gruplarının belirlenmesi olmak üzere altı aşama gerçekleştirilmiştir.

İŞKUR veri seti toplam 13.457 satır iş arayanın kişisel, demografik ve işgücüsel bilgilerinden oluşmaktadır. Çalışmanın konusu işsizlerin zamana göre işsiz kalma risklerinin tespit edilmesi olduğu için daha iyi şartlarda iş arayan ve emekli olup iş arayanlar veri setinden çıkartılmış başvuru türü "Çalışırken İşsiz Kalan" ve "ilk kez İş Hayatına Atılan" olmak üzere iki kategoriye ayrılmıştır. Ham veri seti analiz ve modelleme yapmaya elverişli olmadığından veri ön işlemeye tabi tutulmuştur. Veri ön işleme kapsamında uyruk, ikamet, sosyal durum, medeni durum ve öğrenim

özniteliklerinde kategori sayısında değişim gerçekleştirilmiştir. Ayrıca her bir meslek için meslek ana grupları belirlenerek meslek modellemeye elverişli hale getirilmiştir. Veri dönüşümü ile tüm öznitelikler kategorik hale getirilmiştir. Özniteliklerin tamamı aynı ölçü biriminde olduğu için standartlaştırma işlemi yapılmamıştır. Başvuru sonrası bir yıl içerisinde işe yerleşmeyenlerin sayısı 12.054 ve işe yerleşenlerin sayısı ise 1.403'tür. Bu haliyle veri seti hedef değişken üzerinden dengelilik arz etmediğinden veri setindeki bu dengesizlik işe yerleşenlere ait satır özelliklerine göre arttırılma yapılarak giderilmiş ve işe yerleşenlerin sayısı 7.050 ve işe yerleşemeyenlerin sayısı 7.096 seviyesine getirilmiştir. Veri setine boyutlandırma işlemi, öznitelik seçme tekniği kullanılarak filter ve sarmal metot kullanılarak gerçekleştirilmiştir. Filter metotta istatistiksel korelasyon ve ki-kare analiz kullanılmış, sarmal metotta lojistik regresyona analizi geri arama tekniği kullanılarak boyutlandırma işlemi gerçekleştirilmiştir. Tüm bu işlemlerle veri seti iyi bir modellemeye hazır hale getirilmiştir.

İŞKUR'a kayıt yaptıran işsizlerin kayıt sonrası işe yerleşme(1)/yerleşmeme(0) durumu şeklinde sınıflandırılmıştır. Sınıflandırma işlemi çok değişkenli istatistiksel modelleme ve denetimli makine öğrenmesi algoritmaları kullanılarak yapılmıştır. Verilerin tamamının kategorik olması ve çalışmanın amacına uygun çok değişkenli istatistiksel analiz yöntemi olan çok değişkenli lojistik regresyon modellemesi kullanılmıştır. Yine amaca uygun olarak denetimli öğrenme algoritmalarından Rasgele Orman, K-En Yakın Komşu, Naive Bayes, Karar Ağaçları ve Destek Vektör Makineleri, Yapay Sinir Ağları algoritmaları sınıflama işlemi için kullanılmıştır.

Çok değişkenli lojistik regresyon modellemesi ile model parametreleri, model açıklama oranı ve doğru sınıflandırma performansı hesaplanarak değerlendirilmiştir. Binary lojistik regresyon modellemesine göre yordanan değişkenler 0 ile 1 arasında olasılıklı sonuçlar almaktadır. Sonuçların 0,50'den büyük olması durumunda işsizler işe yerleşmiş, 0,50'den küçük olması durumunda ise yerleşmemiş olarak kabul edilmektedir. Oluşturulan hem çok değişkenli lojistik regresyon modeli hem de model parametreleri %95 anlamlılık düzeyi ile istatistiksel olarak anlamlıdır. Modellemenin başlangıcından itibaren 2 öznitelik(cinsiyet, sosyal yardım alma durumu) çıkartılarak 8 öznitelikli(sosyal yardım, yaş, medeni durum, öğrenim, meslek grup, başvuru türü, ikamet ve işsizlik ödeneği alma durumu) modele ulaşılmıştır Model açıklama oranına(R^2) göre modelde yer alan öznitelikler hedef değişkeni olan işsizin işe

yerleşme durumunu %5,1 oranında açıklamaktadır. Sınıflandırma tablosuna(hata matrisi) göre doğru sınıflandırma oranı %59 olup çok değişkenli lojistik regresyon modellemesine göre orta seviyede sınıflandırma işlemi gerçekleştirilmiştir. Model açıklama oranın ve doğru sınıflandırma oranın düşüklüğü iki şeyin önemini ve gerekliliğini ortaya kılmaktadır. Bunlardan birincisi başvuru sonrası işsizin işe yerleşmesinde etkin ve açıklama oranını yükseltecek diğer ana ve alt faktörlerin belirlenmesi, ikincisi ise daha yüksek sınıflandırma oranına sahip sınıflandırma tekniklerinin uygulanma gerekliliğidir.

İŞKUR'a başvuran işsizlerin, başvuru sonrası bir yıl içerisinde işe yerleşme durumu ile işsizin kişisel, demografik ve işgücüsel bilgileri arasında istatistiksel anlamlı bir ilişkinin olup olmadığının tespiti için kategorik verilere uygun ilişki tespiti ki-kare analizi yapılmıştır. Ki-kare analizine göre işsizin işe yerleşmesi ile modelde yer alan tüm öznitelikleri arasında %95 güven düzeyinde istatistiksel anlamlı bir ilişki bulunmaktadır. Kontenjans katsayısına göre işsizin işe yerleşmesi ile tüm öznitelikleri arasındaki çok zayıf ilişki bulunmaktadır. Ancak zayıf ilişkiye sahip özniteliklerin bir araya getirilmesi daha güçlü bir model oluşturabilmektedir. İlişki gücüne göre en yüksek ilişki işsizin yaşı, medeni durumu ve meslek grubunda gözlenirken en düşük ilişki ise işsizin ikameti, sosyal yardım ve işsizlik ödeneği alma durumu arasında gözlemlenmektedir. Bu bilgiler İŞKUR temel işsiz kayıt bilgilerinin iş profillemede gerekliliğini ve önemini ortaya koymaktadır.

Lojistik regresyon analizi ve ki-kare analizi sonuçlarına göre ODDS Ratio oranları üzerinden özniteliklere göre risk grupları belirlenmiştir. İşsizin sosyal yardım alma özniteliği dışındaki tüm öznitelikleri ile başvuru sonrası bir yıl içerisinde işe yerleşmesi arasında istatistiksel anlamlı bir ilişki olduğu için söz konusu öznitelik dışındaki tüm öznitelikler için risk tespiti yapılmıştı ve risk tespitinde kategoriler "riskli" ve "daha riskli" olmak üzere iki gruba ayrılmıştır. Grupların tasnifi, işe yerleşmeme ODDS oranı düşük olan kategoriler "riskli", ODDS oranı yüksek olan kategoriler ise "daha riskli" olarak gerçekleştirilmiştir. Risk gruplandırmasına göre ilçe de oturanlar, normal statüde olanlar, kadınlar, orta ve daha ileri yaşta olanlar, evliler, ilkokul ve altı eğitime sahip olanlar, tesis ve makine operatörü ve montajcısı dışında mesleğe sahip olanlar, ilk kez iş hayatına atılanalar ve işsizlik ödeneği almayanlar İŞKUR'a başvuru sonrası bir yıl içerisinde işsiz kalma riski daha yüksektir.

Örneğin Sivas merkez de ikamet eden işsizler ilçede oturan işsizlere göre daha düşük bir işsiz kalma riskine sahiptir.

Çok değişkenli istatistiksel analizler ile gerçekleştirilen modelleme, sınıflama ve test işlemlerinden sonra yapay zekâ makine öğrenmesi algoritmaları ile aynı işlemler gerçekleştirilmiştir. Denetimli makine öğrenmesi algoritmaları ile oluşturulan modeller için hem Hold Out hem de K-Çapraz Doğrulama yöntemi kullanılmıştır. Çalışmada iki farklı veri seti bulunmaktadır. Bunlardan birincisi tüm işsizlerin yer aldığı yığın ve diğeri ise yığından %20 oranda çekilen örneklemdir. Öznitelik seçme ve çıkarma işlemine göre indirgenmiş veri seti ile indirgenmemiş veri seti için makine öğrenmesi algoritmaları uygulanmış ve her iki veri setinin model performansları kıyas edilmiştir. Modelin eğitimi dört farklı şekilde gerçekleştirilmiştir. Birincisinde veri setinin %70 eğitim, %30'u test, ikincisinde veri setinin %80'i eğitim ve %20'si test, üçüncüsünde K-5 çapraz doğrulama ve dördüncüsünde K-10 Çapraz doğrulamadır. Model performansları Doğruluk, Kesinlik, F-Ölçütü ve Duyarlılık gibi ölçütler üzerinden değerlendirilmiştir. Model algoritması, geçerleme yöntemi, veri seti, özellik seçimi, eğitim şekline göre model performans değerlendirmeye tabi tutulmuştur. Performans değerlerine göre en iyi model nihai model olarak seçilmiştir. Ayrıca en yüksek doğruluk performansı gösteren rasgele orman regresyon modellemesi ile zamana göre işsiz kalma risk tespiti yapılmış ve değerlendirilmiştir.

İndirgenmiş ve indirgenmemiş yığın ve örneklem veri seti için farklı geçerleme/doğrulama tipi ve makine öğrenmesi algoritmasına göre toplam 112 adet model oluşturulmuş ve oluşturan modeller için doğruluk performansları hesaplanmıştır. Farklı tip ve şekildeki model geçerleme/doğrulama ve makine öğrenmesi algoritmaları ile oluşturulan modellerin neredeyse tamamında en yüksek doğruluk performansı indirgenmemiş yığın veri setinde, holdout model geçerleme yöntemi ile %80 Eğitim-%20 Test model makine eğitimde rasgele orman algoritmasında gerçekleştirilmiştir. Makine öğrenmesi algoritmalarının doğruluk performansı incelendiğinde doğruluk performansında ana belirleyici faktörün model algoritmaları olduğu tespit edilmiştir. Model veri setindeki değişim(yığın/örneklem) ortalama yaklaşık %5, boyutlandırmadaki değişim(indirgenmemiş/indirgenmiş) model geçerlemedeki değişim(Hold Out/K Çaprazlama), model eğitimindeki değişim(%80 Eğitim/%20 Eğitim), model doğrulamada değişim ortalama yaklaşık %2'nin altında model doğruluk performansını etkilemektedir. Bununla birlikte model

parametreleri üzerinden gerçekleştirilen model iyileştirmesinde model performansını en fazla %1 oranında arttırmaktadır. Bu durum model doğruluk performansında model algoritması türünün ve veri setinin önemini ortaya koymaktadır.

Rasgele orman algoritması ile gerçekleştirilen tüm modellerde doğruluk performansı diğer algoritmalara göre daha yüksektir. Rasgele orman algoritması ise oluşturulan modellerin ortalama yaklaşık doğruluk performansı %72,6'dır. Bu doğruluk performansını %71,7 ile karar ağaçları, %67,4 ile k en yakın komşu, %64,9 ile yapay sinir ağları, %59,7 ile lojistik regresyon, %58,9 ile destek vektör makinaları ve %57,2 ile naive bayes algoritmaları takip etmektedir. Lojistik regresyon ve naive bayes gibi istatistik tabanlı algoritmalar ile oluşturulan modeller makine öğrenmesi tabanlı algoritmalardan daha düşük doğruluk performansı göstermiştir. Bunun istisnası ise destek vektörü modellemesidir. Veri seti türü düşükte olsa model doğruluk performansında bir etki yapmaktadır. Veri setinin boyutlandırılması, model eğitimi türü ve şekli model doğruluk performansında belirgin bir fark oluşturamamıştır. En yüksek doğruluk performansı, indirgenmemiş yığın veri seti için holdout model geçerleme yöntemine göre oluşturulan modeller için diğer performans göstergeleri incelendiğinde rasgele orman algoritması ile oluşturulan model doğruluk, kesinlik, duyarlılık ve F1 skor performansında diğer modellere çoğunlukla üstünlük göstermiştir. Karar ağaçları ve rasgele orman algoritması diğer algoritmalardan yüksek performans ile belirgin olarak ayrılmaktadır. Lojistik regresyon ve naive bayes gibi istatistiksel modeller makine öğrenmesi modellerine göre belirgin olarak daha düşük bir performans ortaya koymuştur.

Çalışmada sadece lojistik regresyon algoritması ile diğer makine öğrenmesi algoritmalarının doğruluk sınıflandırması yönünden aralarında istatistiksel olarak anlamlı bir fark olup olmadığı ki-kare analizi ile test edilmiştir. Ki-kare analizi sonuçlarına göre p değeri 0,05'ten küçük olan rasgele orman, karar ağaçları, yapay sinir ağları ve k en yakın komşu algoritmaları ile lojistik regresyon algoritması arasında doğruluk sınıflandırması yönünde %95 güven düzeyinde istatistiksel olarak anlamlı fark bulunmaktadır. P değeri 0,05'ten büyük olan naive bayes ve destek vektör makinaları ile lojistik regresyon algoritması arasında doğruluk sınıflandırması yönünde %95 güven düzeyinde istatistiksel olarak anlamlı fark bulunmamaktadır. Doğruluk sınıflandırması yönünden lojistik regresyon algoritması ile arasında istatistiksel fark olmayan naive bayes ve destek vektör makinaları algoritmaları

yaklaşık %60 doğruluk performansı sergilerken, doğruluk sınıflandırması yönünden arasında istatistiksel fark olan k en yakın komşu ve yapay sinir ağları algoritması yaklaşık %70, rasgele orman ve karar ağaçları algoritması ise yaklaşık %80 doğruluk performansı sergilemektedir. Bu sonuçlara göre makine öğrenmesi algoritmalarının çoğunluğu çok değişkenli lojistik regresyon analizine göre daha yüksek bir doğruluk performansı göstermiştir.

En yüksek doğruluk performansı indirgenmemiş yığın veri seti için %80 Eğitim- %20 Test model geçerleme işlemi ile gerçekleştirildiği için parametreler üzerinden model iyileştirme işlemi bu veri seti ve model geçerleme ile yapılmıştır. Hem karar ağaçlarında hem de rasgele orman algoritmasında performans iyileştirme diğer parametreler sabit tutularak sadece bölünme yöntemi(criterion) üzerinden gerçekleştirilmiştir. Gini ve entropy bölünme yöntemine göre hem karar ağaçlarında hem rasgele orman algoritmasında bölünme yöntemlerindeki farklılık yok denecek kadar düşük bir oranda(yaklaşık binde bir) değişime neden olmuştur. Dolayısıyla ana parametre olarak sayılan bölünme yöntemindeki farklılıklar karar ağacı ve rasgele orman algoritmalarının başta doğruluk olmak üzere, kesinlik ve duyarlılık performans değerleri üzerinde belirgin bir iyileştirme sağlamamaktadır. K-En yakın komşu algoritmasında performans iyileştirme diğer parametreler sabit iken komşu sayısında gerçekleştirilmiştir. Komşu sayısının 1-NN olduğu durumda %73,1 olan doğruluk performansı, 10-NN olduğu durumda %74,6 olmuştur. Dolayısıyla yaklaşık %1,5 oranında bir doğruluk performans artışı gerçekleşmiştir. Bu bilgi komşu sayısındaki artışın doğruluk performansında belirgin bir artışa neden olmadığı sonucunu ortaya koymuştur. Ancak önemle belirtmek gerekir ki duyarlılık ölçütünde yaklaşık %11,3 oranda belirgin bir performans artışı gözlemlenmiştir. Duyarlılık artışı ile İŞKUR'a başvuru sonrası bir yıl süre içerisinde gerçekte işe yerleşmiş olanların doğru tahmin edilme oranının da artış meydana gelmiştir. Her üç algoritma için parametre değişimleri üzerinden meydana gelen performans değişimleri birlikte değerlendirildiğinde, algoritmaların parametre değişimleri algoritma performanslarında belirgin bir artış veya azalışa neden olmamaktadır.

Makine öğrenmesi ilişkin tüm performans iyileştirme teknikleri birlikte değerlendirildiğinde indirgenmemiş yığın veri setinin %80 eğitim-%20 test olarak gini bölünme yöntemine gerçekleştirilen rasgele orman algoritması en yüksek performans

göstermiştir. Hâlihazırda OECD ülkelerinden Belçika ve Yeni Zelanda da iş profilleme ve risk tespiti bu algoritmaya göre yapılmaktadır.

İşsizlerin risk gruplarının belirlenmesinde 0,40-0,60 aralığında kritik işe yerleşme olasılığına sahip işsizlerin için kesin ifadelerle işe yerleşti/yerleşmedi şeklinde bir sınıflamadan kaçınılarak "riskli" düzeyi kullanılması ve diğer risk sınıflarının bu temel üzerine bina edilmesi sağlanarak diğer risk sınıfları eşit aralıklı şekilde "Çok Yüksek Riskli, Yüksek Riskli, Riskli, Düşük Riskli ve Çok Düşük Riskli" türünde bir likert ölçeği risk sınıflandırması yapılmıştır. Sınıflama için işsizlerin zamana göre işsiz kalma olasılıklarının yüksek sınıflandırma performansı sergileyen rasgele orman algoritması regresyon modellemesi üzerinden yapılmıştır. Rasgele orman regresyon sonuçlarına göre tahmin edilen işe yerleşme olasılıkları ile gerçek işe yerleşme olasılıkları arasında 0,69 oranında yüksek korelasyon bulunmaktadır. Oluşturulan rasgele orman regresyon modelinin açıklama oranı($\boldsymbol{R^2}$) 0,48 ile oldukça iyi durumdadır. Ayrıca hem mutlak hata(0,26) hem de ortalama hata karesi(0,13) düşük seviyededir. Bununla birlikte 0,80 oranında yüksek seviyede bir doğru sınıflama oranı elde edilmiştir Rasgele orman algoritması regresyon modeli ile hem işe yerleşme/işsiz kalma olasılıkları hesaplanmış, hem yüksek bir açıklama oranı elde edilmiş ve hem de yüksek bir doğru sınıflama performansı sağlanmıştır. Bu tespitlerden sonra işe yerleşme/işsiz kalma olasılıklarına göre işsiz kalma risk sınıfları oluşturulmuştur.

İşsiz kalma risk sınıflaması eşit aralıklı olup her işsiz için bir işsiz kalma risk sınıfı belirlenmiştir. Risk sınıflarına göre ilgili dönemde İŞKUR'a başvuran 7.566 işsiz tasnif edilmiştir. Risk sınıflandırma oranına göre Sivas'ta İŞKUR ilgili dönemde kayıt yaptıran işsizlerin %48,7'si çok riskli grupta yer alarak çok yüksek olasılıkla başvuru sonrası bir yıl içerisinde işsiz kalmaktadır. Başvuru sonrası bir yıl içerisinde çok yüksek olasılıkla işe yerleşen işsizlerin oranı çok düşük olup %6,4'tür. İşe yerleşip yerleşmeyeceği riskini beraber taşıyan işsizlerin oranı %16,5'tir. Riskli, yüksek riskli ve çok yüksek riskli işsizler birlikte değerlendirildiğinde işsiz kalma riski taşıyan işsizlerin oranı %76,6 olup yüksek seviyededir. Diğer bir ifadeyle Sivas ilinde her dört işsizden üçü işsiz kalma riski taşımaktadır. İşsizlerin başvuru sonrası ortalama işe yerleşme olasılığı 0,30 olup düşük seviyededir. Ayrıca çok düşük seviyedeki(%6) işe yerleşme oranı ildeki yüksek düzeyde işsiz kalma riskini teyit etmektedir.

Çalışma sonuçları hep birlikte değerlendirildiğine aşağıda yer alan net sonuçlara ulaşılmıştır.

- ✓ Profilleme işleminde yapay zekâ ve makine öğrenmesi modellerinin çok değişkenli istatistiksel modellere göre daha yüksek performans sergilediği,
- ✓ Model doğruluk performansında ana belirleyici faktörün model algoritmalarının olduğu,
- ✓ Yığın veri seti ile yapılan modellemelerinin ortalama yaklaşık %5 daha fazla doğruluk performansı sağladığı,
- ✓ Karar ağaçları ve rasgele orman gibi ağaç tipi algoritmaların diğer makine öğrenmesi algoritmalarına göre belirgin doğruluk performansı farkı gösterdiği,
- ✓ İyi bir modelleme işleme için iyi bir veri ön işleme/veri tasarımı süreci işletilmesinin gerekliliği,
- ✓ Rasgele orman regresyon ile hem yüksek doğruluk hem de açıklama oranı elde edildiği,
- ✓ Veri setinde yapılan boyutlandırma işleminin model doğruluk performansını çok düşük seviyede düşürmesiyle birlikte modelde olması zorunlu olan cinsiyet gibi değişkenleri model dışı bırakması nedeniyle boyut indirgeme işlemine gidilmemesi,
- ✓ İşsizlik sorunun çözümünde işsiz kalma risk tespitine dayalı profillemenin önemi ve gerekliliği,
- ✓ İŞKUR temel işsiz kayıt bilgilerinin neredeyse tamamının işsizin İŞKUR'a başvuru sonrası bir yıl içerisinde işe yerleşmesi ile ilişkili olduğu ve bu ilişkinin çok zayıf olmasına rağmen doğru teknik ve yöntemin seçilmesiyle yüksek performans sağlayan bir modelleme sağlanacağı,
- ✓ Açıklama ve doğruluk oranı iyi bir modelleme işlemi için işsizlerin başta iş arama davranışı olmak üzere başkaca özniteliklerinin modellemeye dâhil edilmesi,
- ✓ Risk sınıflandırılmasının belirlenmesinde 0,40-0,60 aralığında işe yerleşme/işsiz kalma riskinin taşıyanların temel kabul edilerek diğer risk gruplarının bu gruba göre dizayn edilmesi,
- ✓ İlgili dönemde İŞKUR'a başvuran işsizlerin başvuru sonrası bir yıl içerisinde %76,6 oranla yüksek seviyede işsiz kalma riski taşıdığı ve bu profillemenin zorunluluğu,

- ✓ İŞKUR tarafından gerçekleştirilecek işsiz kalma riski tabanlı profilleme işleminde çok değişkenli istatistiksel modelleme ve makine öğrenmesi modelleme işlemlerinin birlikte yapılmasının gerekliliği tespit edilmiş ve değerlendirilmiştir.

KAYNAKLAR

[1] **Takcı, H.** (2020). *Teori ve Uygulamada Veri Madenciliği*. Nobel Akademik Yayıncılık No: 3218, 211s, Ankara.

[2] **Türkiye Cumhuriyeti Cumhurbaşkanlığı Dijital Dönüşüm Ofisi.** (2021). *"Ulusal Yapay Zekâ Stratejisi 2021-2025*", 95s.

[3] **Ersöz, F., Çınar, Y.** (2021). Veri Madenciliği ve Makine Öğrenimi Yaklaşımlarının Karşılaştırılması: Tekstil Sektöründe Bir Uygulama. *Avrupa Bilim ve Teknoloji Dergisi*, 21(2021), 397-414

[4] **Emeç, H., Üçdoğruk Birecikli, Ş.,Acar Balaylar, N.** (2021). Türkiye'de Genç İşsizliğinin Profili: Panel Logit Model Tahmini. *Journal of Management & Labor/Yönetim ve Çalışma Dergisi,* 5(2), 137-154

[5] **Işık, M.** (2017). Profil Temelli Danışmanlık Kapsamında Bireysel Eylem Planlarının Oluşturulması: Dünya Uygulamaları Ve Türkiye Simülasyonu. *İŞKUR Uzmanlık Tezi*, 195s, Ankara.

[6] **Desiere, Sam., Langenbucher, K., Struyven, L. Y.** (2019). Statistical Profiling in Public Employment Services: An international comparison. *OECD Social, Employment and Migration Working Papers* No. 224, 28 s.

[7] **Aktaş, Y.** (2019). Aktif İstihdam Politikası Araçlarından İŞKUR İşbaşı Eğitim Programlarının Mikro ekonomik Etki Değerlendirmesi Ve Lojistik Regresyon Modellemesi Sivas İli Araştırması. *İŞKUR İl Analizi*, 810s, Sivas.

[8] Url-1 <(https://sozluk.gov.tr/>, alındığı tarih: 08.07.2024.

[9] **İçen, G., Günay, S.** (2014). Uzman Sistemler ve İstatistik. *İstatistikçiler Dergisi: İstatistik &Aktüerya*, 7(2014), 35-37

[10] Url-2<(https://uib.org.tr/tr/kbfile/yapay-zekâ-ve-yeni-teknolojiler>, alındığı tarih: 02.05.2024.

[11] **Öztemel, E.** (2020). Yapay Zekâ ve İnsanlığın Geleceği. Bilişim Teknolojileri ve İletişim: Birey ve Toplum Güvenliği, 95-112

[12] **Eldem, A., Eldem, H., Palalı, A.** (2017). Görüntü İşleme Teknikleri İle Yüz Algılama Sistemi Geliştirme. *BEU Fen Bilimleri Dergisi*, 6(2), 44-48

[13] Url-3<https://tr.wikipedia.org/wiki/Uzman_sistemler>, alındığı tarih: 02.05.2024.

[14] **Uğuz, S.** (2021). *Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Bir Yapay Zekâ Ekolü*. Nobel Akademik Yayıncılık No: 2565, 297s, Ankara.

[15] **Tüzüntürk, S.** (2009). Veri Madenciliği ve İstatistik. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Dergisi*, XXIX(1), 65-90

[16] Url-4 <https://tr.wikipedia.org/wiki/Süreç>, alındığı tarih: 06.05.2024.

[17] **Alaybeg, F. (**2019). Veri Madenciliği Giriş, Yöntemleri ve Metodolojileri. https://furkanalaybeg.medium.com/veri-madenciliği-ve-yöntemleri-d0e2fd238e44. Erişim Tarihi: 08.05.2024.

[18] **Günbatur, E.** (2019). Türkiye'de otomobil sigortası sahtekârlıklarının makine öğrenmesi yöntemleri ile tespit edilmesi. *Hacettepe Üniversitesi. Endüstri Mühendisliği Ana Bilim Dalı(Yüksek Lisans Tezi),* 97s, Ankara.

[19] **Kapanoğlu, M., Abdalla, S., Gültekin, Ö., Er, F., Sönmez, H.** (2019). *İşletme Analitiği*. Anadolu Üniversitesi Yayını No: 3409, 214s, Eskişehir.

[20] Url-5 <https://tr.wikipedia.org/wiki/ogrenme >, alındığı tarih: 06.05.2024.

[21] Url-6 <https://ilge.com.tr/>, alındığı tarih: 08.05.2024.

[22] Url-7<https://bulutistan.com/blog/makine-ogrenmesi-karar-agaci-decision-tree-nedir/>, alındığı tarih: 15.05.2024.

[23] **Torun, H.** (2022). Karar Ağacı (Decision Tree).https://hakan.io/karar-agaci-decision-tree/. Erişim Tarihi: 15.05.2024.

[24] **Atcılı, A.** (2022). Karar Ağaçları Algoritmasi.https://medium.com/machine-learning-türkiye/karar-agaclari-algoritmasi-b823c23997d0. Erişim Tarihi: 15.05.2024.

[25] **Demir, E.** (2021). Üretim planlama ve kontrol süreçlerinde veri madenciliğinin yeri ve çok kriterli karar alma yaklaşımlarıyla çözüm önerileri: bir işletme uygulaması. *Marmara Üniversitesi. Ekonometri Ana Bilim Dalı(Doktora Tezi*), 259s, İstanbul.

[26] **Breiman, L.** (2001), Random Forests, Machine Learning, 45 (1), 5–32.

[27] **Miraç, Ö.** (2022). Phthon İle Sınıflandırma Analizleri(Rastgele Orman Algortiması).https://miracozturk.com/python-ile-siniflandirma-analizleri-rastgele-orman-random-forest-algoritmasi/. Erişim Tarihi: 15.05.2024.

[28] Url-8<https://www.geeksforgeeks.org/random-forest-regression-in-python/>, alındığı tarih: 16.05.2024.

[29] **Metlek, S., Kayaalp, K**. (2020). Makine Öğrenmesinde, Teoriden Örnek Matlab Uygulamalarına Kadar Destek Vektör Makineleri. *Iksad Publicationsı,* 93s, Ankara

[30] **Uğurluoğlu, K**. (2021). Naive Bayes Sınıflandırması- Machine Learning.

https://kaanugurluoglu123.medium.com/naive-bayes-sınıflandırması-machine-learning-5d086326cc60. Erişim Tarihi: 17.05.2024.

[31] **Cengiz, A.** Makine Öğrenmesi Ders Notları.

https://avys.omu.edu.tr/storage/app/public/macengiz/133359/5.HAFTA.pdf. Erişim Tarihi: 16.05.2024.

[32] Url-9<https://www.geeksforgeeks.org/support-vector-machine-algorithm/?ref=lbp >, alındığı tarih: 16.05.2024.

[33] Url-10<https://www.geeksforgeeks.org/k-nearest-neighbours/?ref=lbp>, alındığı tarih: 16.05.2024.

[34] **Miraç, Ö.** (2021). Python İle Sınıflandırma Analizleri(En Yakın Komşu Algoritması(KNN)). https://miracozturk.com/python-ile-siniflandirma-analizleri-knn-k-nearest-neighbours-k-en-yakin-komsu-algoritmasi/. Erişim Tarihi: 16.05.2024.

[35] **Kabalcı, E.** (2014). Yapay Sinir Ağları (Artificial Neural Networks) Ders Notları.

https://ekblc.wordpress.com/wp-content/uploads/2013/09/ysa.pdf. Erişim Tarihi: 16.05.2024.

[36] **Keskenler, K., Keskenler, E.F**. (2017). Geçmişten Günümüze Yapay Sinir Ağları ve Tarihçesi. *Takvim-i Vekayi* 5(2), 8-18

[37] Url-11<https://tr.wikipedia.org/wiki/Sinir_hücresi>, alındığı tarih: 17.05.2024.

[38] **Öztürk, M.F., Ergin Şahin, M.** (2018). Yapay Sinir Ağları ve Yapay Zekâya Genel Bakış. *Takvim-i Vekayi* 6(2), 25-36

[39] **Çınar, U.K.** (2018). Yapay Sinir Ağları ve R Programıyla Uygulama. https://www.veribilimiokulu.com/yapay-sinir-aglari/. Erişim Tarihi: 17.05.2024.

[40] **Miraç, Ö.** (2021). Python İle Sınıflandırma Analizleri(Yapay Sinir Ağları). https://miracozturk.com/python-ile-siniflandirma-analizleri-yapay-sinir-aglari-ysa/. Erişim Tarihi: 17.05.2024.

[41] **Miraç, Ö.** (2021). Python Çapraz Doğrulama Teknikleri. https://miracozturk.com/capraz-dogrulama-teknikleri-cross-validation/.Erişim Tarihi: 17.05.2024.

[42] **Şahin, S.** (2021). Makine Öğrenmesi Yöntemleri ile Ortaokul Öğrenci Başarılarının Tespiti Ve Bir Uygulama. *İstanbul Üniversitesi. Enformatik Ana Bilim Dalı(Yüksek Lisans Tezi)*, 139s, İstanbul

[43] Url-11<https://en.wikipedia.org/wiki/Statistical_learning_theory>, alındığı tarih: 21.05.2024.

[44] **Ağırdan, D.** (2020). Göbek Bağı Bir: İstatistiksel Öğrenme vs Makine Öğrenmesi.https://medium.com/deep-learning-turkiye/göbek-bağı1-bir-istatistiksel-öğrenenme-vs-makine-öğrenmesi. Erişim Tarihi: 21.05.2024.

[45] **Gamgam, H., Altunkaynak, B.** (2015). *SPSS Uygulamalı Regresyon Analizi*. Seçkin Yayıncılık No:441, 226s, Ankara.

[46] **Alpar, R.** (2017). Uygulamalı *Çok Değişkenli İstatistiksel Yöntemler*. Detay Yayıncılık No:429, 820s, Ankara

[47] **Akça, M.F.** (2021). Lojistik Regresyon Nedir? Nasıl Çalışır? https://mfakca.medium.com/lojistik-regresyon-nedir-nasıl-çalışır-4e1d2951c5c1. Erişim Tarihi: 02.07.2024.

[48] **Karagöz, Y.** (2010). İlişki Katsayıları İle Öğrenci Başarısını Etkileyen Faktörlerin Belirlenmesi. *Elektronik Sosyal Bilimler Dergisi*, 9(32), 425-446

[49] **Esin, A., Ekni, M., Gamgam, H**. (1997). *Sağlık Bilimlerinde İstatistik*. Gazi Üniversitesi Yayını No:171, 508s, Ankara.