ÇANKIRI KARATEKİN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

LİSE GİRİŞ SINAVINDA ÖĞRENCİ BAŞARILARININ MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE ANALİZİ

Mehmet ŞENLİGİL

ELEKTRONİK-BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

ÇANKIRI
2023

Her hakkı saklıdır

TEZ ONAYI

Mehmet ŞENLİGİL tarafından hazırlanan **"Lise Giriş Sınavında Öğrenci Başarılarının Makine Öğrenmesi Teknikleri ile Analizi"** adlı tez çalışması 25/07/2023 tarihinde aşağıdaki jüri tarafından oy birliği/oy çokluğu ile Çankırı Karatekin Üniversitesi Fen Bilimleri Enstitüsü Elektronik ve Bilgisayar Mühendisliği Anabilim Dalında **Yüksek Lisans Tezi** olarak kabul edilmiştir.

Danışman **:** Dr. Öğr. Üyesi Selim BUYRUKOĞLU

Jüri Üyeleri **:**

Başkan **:** Dr. Öğr. Üyesi Selim BUYRUKOĞLU
Bilgisayar Mühendisliği Anabilim Dalı
Çankırı Karatekin Üniversitesi

Üye **:** Dr. Öğr. Üyesi. Fuat TÜRK
Bilgisayar Mühendisliği Anabilim Dalı
Kırıkkale Üniversitesi

Üye **:** Dr. Öğr. Üyesi Mustafa KARHAN
Bilgisayar Mühendisliği Anabilim Dalı
Çankırı Karatekin Üniversitesi

Yukarıdaki sonucu onaylarım

Prof. Dr. Hamit ALYAR

Enstitü Müdürü

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Çankırı Karatekin Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğine göre hazırlamış olduğum "**Lise Giriş Sınavında Öğrenci Başarılarının Makine Öğrenmesi Teknikleri İle Analizi**" konulu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, tezin Çankırı Karatekin Üniversitesi Fen Bilimleri Enstitüsü'nden başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve bu çalışmanın Çankırı Karatekin Üniversitesi tarafından kullanılan "Bilimsel İntihal Tespit Programı"yla tarandığını, "intihal içermediğini" beyan ederim. Çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması halinde ortaya çıkacak tüm ahlaki ve hukuki sonuçlara razı olduğumu bildiririm. Çankırı Karatekin Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca gereğinin yapılmasını arz ederim (25.07.2023).

Mehmet ŞENLİGİL

ÖZET

Yüksek Lisans Tezi

LİSE GİRİŞ SINAVINDA ÖĞRENCİ BAŞARILARININ MAKİNE ÖĞRENMESİ TEKNİKLERİ İLE ANALİZİ

Mehmet ŞENLİGİL

Çankırı Karatekin Üniversitesi
Fen Bilimleri Enstitüsü
Elektronik-Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Selim BUYRUKOĞLU

Son yıllarda Eğitsel Veri Madenciliği çalışmalarında Makine Öğrenmesi algoritmaları oldukça yaygın biçimde kullanılmaya başlanmıştır. Bu çalışmada öğrencilerin ders başarı notları ile bireysel ve çevresel özelliklerini içeren veri seti kullanarak LGS sınavındaki başarı tahminlemesi yapılmıştır. Çalışma aynı zamanda tek tabanlı ve topluluk (Ensemble) algoritmalarının performanslarını karşılaştırılmak suretiyle en iyi modeli de önermektedir. Oluşturulan modelleri kıyaslamada R-Squared değeri referans kabul edilirken MSE, RMSE, MAE ve MAPE metriklerine de bakılmıştır. Çalışmada aynı zamanda Feature Selection metodlarının da model performansına etkileri araştırılmıştır. Elde edilen sonuçlarda Feature Selection uygulanmadan tüm veri seti kullanılarak Stacking mimaride Super Learner model ile en iyi performans elde edilmiştir. Modelin R-squared değeri 0,79 olurken MSE 161,32 olarak hesaplanmıştır. Çalışmanın sonuçları çalışılan modelin gelişime açık bir model olduğunu ortaya koymaktadır. Öte yandan öğrencilerin öğrenim ve kişisel özelliklerini içeren veriler kullanılarak merkezi sınavlardaki başarı düzeylerinin modellenebileceğini de göstermektedir. Kullanılan bu teknik ilerleyen dönemlerde öğrencilerin güçlü ve zayıf yönlerinin belirlenmesinde ve öğrencilerin başarısının arttırılmasında önemli bir yere sahip olabilir.

2023, 72 sayfa

ANAHTAR KELİMELER: Eğitsel veri madenciliği, Makine öğrenmesi, Öznitelik seçimi, Topluluk öğrenmesi

ABSTRACT

Master of Science Thesis

STUDENT ACHIEVEMENTS IN HIGH SCHOOL ENTRANCE EXAM ANALYSIS WITH MACHINE LEARNING TECHNIQUES

Mehmet ŞENLİGİL

Çankırı Karatekin University
Graduate School of Natural and Applied Sciences
Department of Electronics and Computer Engineering

Advisor: Asst. Dr. Selim BUYRUKOĞLU

In recent years, Machine Learning algorithms have become widely used in Educational Data Mining studies. This study aims to predict the success in the High School Transition Exam (LGS) using a dataset that includes students' academic performance scores as well as individual and environmental characteristics. The study also proposes the best model by comparing the performances of both single-base and ensemble algorithms. In the comparison of the models, the R-Squared value is considered as a reference, while metrics such as MSE, RMSE, MAE, and MAPE are also examined. Additionally, the study investigates the impact of Feature Selection methods on the model's performance. The results show that the best performance is achieved by utilizing the Super Learner model in a Stacking architecture, using the entire dataset without applying Feature Selection. The model's R-squared value is found to be 0.79, and the MSE is calculated as 161.32. This study demonstrates that the model is open to further improvements based on the obtained results. Moreover, it indicates that the success levels in central exams can be modeled by utilizing data that includes students' educational and personal characteristics. This technique, used in the study, can play an important role in identifying students' strengths and weaknesses and improving their academic performance in the future.

2023, 72 pages

Keywords: Educational data mining, Machine learning, Feature selection, Ensemble learning

ÖNSÖZ VE TEŞEKKÜR

“Lise Giriş Sınavında Öğrenci Başarılarının Makine Öğrenmesi Teknikleri ile Analizi” isimli bu tez çalışması 2021-2023 yılları arasında hazırlanarak Çankırı Karatekin Üniversitesi Fen Bilimleri Enstitüsü’ne “Yüksek Lisans” tezi olarak sunulmuştur.Yüksek lisans eğitimi ve tez süreci boyunca gerek yaklaşımı, bilgisi ve birikimiyle hep yanımda olan gerekse farklı fikir ve bakış açılarıyla ufkumu genişleten ve sorunları çözme konusunda desteğini esirgemeyen saygı değer hocam ve tez danışmanım Sn. Dr. Öğr. Üyesi Selim BUYRUKOĞLU’na teşekkürlerimi sunarım. Ayrıca yüksek lisansa başlamada ön ayak olan ve her daim yanınızda olmasını isteyeceğiniz bir yol arkadaşı olan Servet ERDOĞAN’a, yorucu ve yoğun geçen bu süreçte bana gösterdikleri sevgi, destek, anlayış ve fedakârlıklarından dolayı sevgili eşim Sümeyra ŞENLİGİL’e ve çocuklarım Barış Kaan ŞENLİGİL ile Rüzgâr ŞENLİGİL’e de özel teşekkürlerimi sunarım.

Mehmet ŞENLİGİL

Çankırı, Temmuz 2023

İÇİNDEKİLER

ÖZET ..... **i**
ABSTRACT ..... **ii**
ÖNSÖZ VE TEŞEKKÜR ..... **iii**
İÇİNDEKİLER ..... **iv**
KISALTMALAR DİZİNİ ..... **vi**
ŞEKİLLER DİZİNİ ..... **vii**
ÇİZELGELER DİZİNİ ..... **viii**
1. GİRİŞ **1**
1.1 Öğrenci Başarısının Önemi **1**
1.2 Amaç ve Hedefler **3**
1.3 Araştırma Soruları **4**
1.4 Çalışmanın Literatüre Katkısı **4**
1.5 Tezin Akış Şeması **4**
2. LİTERATÜR ÖZETİ **6**
2.1 Öğrenci Başarısında Özellik Seçimi Önemi **8**
2.2 Öğrenci Başarısını Tahmin Etmede Yaygın Olarak Kullanılan Makine Öğrenmesi Algoritmaları **14**
2.3 Özellik Seçim Yöntemleri ve Makine Öğrenmesi Algoritmaları **29**
2.4 Özellik Seçim Yöntemleri **31**
2.4.1 Filter-based methods **32**
2.4.2 Wrapper (sarmalayıcı) based yöntemler **33**
2.4.3 Gömülü yöntemler **34**
2.4.4 Filter ve wrapper yöntemleri arasındaki temel farkları **35**
2.5 Makine Öğrenmesi Algoritmaları **35**
2.5.1 Tek tabanlı makine öğrenmesi **37**
2.5.2 Topluluk öğrenme algoritmaları **39**
3. MATERYAL VE METOT **43**
3.1 Veri Seti **43**
3.2 Veri Tanım ve Tipleri **45**
3.3 Veri Ön İşleme **46**

3.4 Öznitelik Seçimi 47
3.5 Veri Temizleme ve Veri Doğrulama 51
3.6 Optimum Modeli Belirleme 52
3.6.1 Tek tabanlı (single based) algoritma metrikleri 54
3.6.2 Topluluk öğrenme algoritma metrikleri 56
4. BULGULAR VE TARTIŞMA 58
4.1 Öznitelik Seçim Sonuçları 58
4.2 Tek Tabanlı Makine Öğrenme Modeli Sonuçları 59
4.3 Topluluk Tabanlı Makine Öğrenme Modeli Sonuçları 60
4.3.1 Bagging sonuçları 60
4.3.2 Boosting sonuçları 61
4.3.3 Stacking sonuçları 62
5. SONUÇLAR VE TARTIŞMA 64
KAYNAKLAR 66
ÖZGEÇMİŞ 72

KISALTMALAR DİZİNİ

AB	Adaboost (adaptive boosting)
XGBoost	Aşırı gradyan arttırma (extreme gradient boosting)
OPAK_ID	Benzersiz öğrenci ID (unique student ID)
BİLSEM	Bilim ve sanat merkezleri (science and art centers)
DNN	Derin sinir ağları (deep neural networks)
SVM	Destek vektör makine (support vector machine)
SVR	Destek vektör regresyon (support vector regression)
EVM	Eğitsel veri madenciliği (educational data mining)
DT	Karar ağacı (decision tree)
DTR	Karar ağacı regresyon (decision tree regression)
KNN	K-en yakın komşu (K-near neighbours)
CFS	Korelasyona dayalı özellik seçimi (correlation based Feature selection)
RMSE	Kök ortalama kare hata (root mean squared error)
LDA	Lineer diskriminant analizi (linear discriminant analysis)
LGS	Liselere geçiş sistemi (high school entrance system)
LR	Lojistik regresyon (logistic regression)
ML	Makine öğrenmesi (machine learning)
NB	Naive bayes regresyon (naive bayes regression)
MSE	Ortalama kare hata (mean squared error)
MAE	Ortalama mutlak hata (mean absolute error)
MAPE	Ortalama mutlak yüzde hatası (mean absolute percentage error)
OEGS	Ortaöğretime geçiş sistemi (secondary education transition system)
RFR	Rastgele orman regresyon (random forest regression)
ROSE	Rastgele yüksek örnekleme örnekleri (random over sampling examples)
SMOTE	Sentetik azınlık sınıf örnekleme (synthetic minority over-sampling technique)
SL	Süper öğrenici (super learner)
PCA	Temel bileşenler analizi (principal component analysis)
UCI	University of california Irvine
ANN	Yapay sinir ağları (artificial neural network)
SR	Yığınlama regresyon (stacking regression)

ŞEKİLLER DİZİNİ

Şekil 1.1 Tez akış şeması 5
Şekil 2.1 Veri madenciliği metodolojisinin tipik döngüsü (Şengür ve Karabatak 2018) 13
Şekil 2.2 Makine öğrenmesi ve eğitsel veri madenciliği (Pallathadka *et al.* 2021) 16
Şekil 2.3 Auto ML iş akış şeması (Aghalarova and Keser 2022) 18
Şekil 2.4 Önerilen yöntemin akış diyagramı (Aghalarova and Keser 2022) 19
Şekil 2.5 Destek vektör makinesi (SVM) (a), Rastgele orman (b), AdaBoost (c) tarafından verilen özellik önem düzeyleri ve bu üç modelden alınan puanların ortalama değerleri (d) (Albreiki *et al.* 2021) 20
Şekil 2.6 Kullanılan tekniklerin çalışmalara göre oranı (Rastrollo-Guerrero *et al.* 2020) 21
Şekil 2.7 Literatür taramasında ele alınan teknikler, hedefler ve algoritmalar hakkında temel istatistikler (Ghorbani and Ghousi 2020) 22
Şekil 2.8 Kullanılan makine öğrenimi yöntemlerinin grafik gösterimi (Balaji *et al.* 2021) 23
Şekil 2.9 Kayıt şekline göre öğrenci başarı dağılımları (Aydoğan ve Karcı 2019) 26
Şekil 2.10 Veri madenciliği modelleri (Düzgün 2017) 31
Şekil 2.11 Sınıflandırma algoritmaları (Balaji *et al.* 2021) 36
Şekil 2.12 Bagging ve Boosting model arasındaki fark (Chauhan 2021) 41
Şekil 3.1 Modelin akış diyagramı 43
Şekil 3.2 Veri başlıkları ve aldığı değerler 47
Şekil 3.3 Chi-square metodu ile yapılan öznitelik seçim sonuçları 48
Şekil 3.4 Veri seti özeti 52

ÇİZELGELER DİZİNİ

Çizelge 2.1 Bagging ve boosting model farkları 42
Çizelge 3.1 Veri Seti öznitelik listesi veri yapısı 44
Çizelge 3.2 Veri başlıkları ve içerdiği değerler skalası 45
Çizelge 3.3 Öznitelik seçim yöntemlerine göre alınan sonuçlar 50
Çizelge 3.4 Algoritmalara göre en iyi sonuç alınan hyper-parametre listesi 53
Çizelge 4.1 ANOVA ve chi-square teknikleri ile belirlenen öznitelik tablosu 59
Çizelge 4.2 Tek tabanlı makine öğrenmesi algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış) 59
Çizelge 4.3 Tek tabanlı makine öğrenmesi algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış) 60
Çizelge 4.4 Bagging algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış) 60
Çizelge 4.5 Bagging algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış) 61
Çizelge 4.6 Boosting algoritmalarının performans sonuçları (Chi-Square öznitelik seçimi uygulanmış) 62
Çizelge 4.7 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış) 62
Çizelge 4.8 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış) 63
Çizelge 4.9 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış) 63

1. GİRİŞ

1.1 Öğrenci Başarısının Önemi

İnsanlık tarihindeki büyük gelişmeler ve dönüşümler, insanın yaşadığı evreni keşfetme isteği ve bu bilgiyi hayata geçirmesiyle gerçekleşmiştir. Tarım devrimiyle başlayan dönüşüm, avcılık ve toplayıcılıkla geçimini sağlayan toplumların yerleşik hayata geçmesini sağlamıştır. Daha sonra 18. yüzyılın ikinci yarısından itibaren İngiltere'de başlayan Sanayi Devrimi, üretimde kullanılan insan gücünün yerini makine gücünün almasına olanak sağlamıştır. Bu dönüşümler toplumların sosyo-ekonomik yapılarında önemli değişikliklere yol açmıştır. Hızla artan nüfus, şehir hayatının başlamasına, sanat, mimari ve kültürün gelişmesine zemin oluşturmuştur (Özsoylu 2017). Büyük keşifler arasında yer alan ve 19. yüzyılın başında hayata geçirilen elektrik teknolojisi, 2.Sanayi Devrimi'nin gerçekleşmesine ve ülkeler arasında askeri alanda güç farklarının doğmasına neden olmuştur. 20. yüzyıl başlarında askeri üstünlüğün kullanımı ardı ardına iki büyük savaş yaşanmasına neden olmuştur. Özellikle 2.Dünya Savaşı'nın yarattığı büyük yıkımlar neticesinde sıcak savaşlar yerini bilgi ve teknoloji savaşlarına bırakmıştır. Bu sayede fen ve teknoloji alanındaki gelişmeler hız kazanmış, Teknoloji Devrimi diye adlandıran Elektronik Çağ yani 3.Sanayi Devrimi gerçekleşmiştir. Elektronik teknolojisi sayesinde insan eliyle yönetilen makinelerin yerini bilgisayarlar ve otomasyon sistemleri almıştır. 2000'li yılların başında elektronik teknolojisi de silikon teknolojisinin keşfi ile yerini mikro elektronik cihazlara bırakmış. Böylece Endüstri 4.0 diye adlandırılan ve yapay zekâ, robotik sistemler ve akıllı cihazları içeren Dijital Çağ başlamıştır (Kamber ve Bolatan 2019).

Tüm bu gelişmeler toplumlarda bilgi, beceri, kültür ve tarih mirasını sonraki nesillere aktarma, teknolojik gelişmelerin gerisinde kalmama ve üretme gerekliliğini doğurmuştur. Bunu gerçekleştirmek adına eğitim ve öğretim tüm toplumlar için her çağda büyük önem arz etmiştir. Eğitim alanında da ilerlemeler her geçen gün artan bir hızla devam etmektedir. Nitelikli iş gücü ve insan kaynağı yaratılması toplumsal gelişim bakımından olmazsa olmaz bir gereksinimdir. Nitelikli bireyler yetiştirmenin yolu da eğitimden

geçmektedir (Bağcı 2018). Nitekim iyi bir kariyer planlaması yapmak isteyen öğrenciler için okul ve alan seçimi önemli bir unsura dönüşmüştür. Ülkemizde uluslararası düzeyde öğrenci yetiştiren okul sayısı ile öğrenci sayısı arasındaki fark sınavla seçme sistemini zorunlu kılmaktadır (Çetin ve Bülbül 2017). Toplumda da öğrencileri genel beceri ve yetenekleri ile değerlendirmek yerine sınav başarı puanlarına göre sıralamanın daha doğru bir uygulama olduğu düşüncesi (Baykal 2014) ve merkezi sınav sistemine duyulan güven olgusu kademeler arası geçişte sınav sistemlerinin gerekliliğini ortaya koymaktadır (Çetin ve Ünsal 2019). Bu da öğrencilerin küçük yaşlardan itibaren hazırlandıkları bir sınav maratonu ve bunun sonucunda öğrencilerde oluşan psikolojik baskı ve ciddi kaygılar demektir. (Kayapinar 2006) Öğrenciler üzerindeki kaygıyı azaltmanın yolu öğrencilerin kişisel beceri ve yeteneklerini ön plana çıkaracak bir eğitim sisteminin kurgulanmasından ve öngörülebilirliği olan kariyer planlama tekniklerinin kullanımından geçmektedir. Bu alanda yürütülen yapay zekâ çalışmaları gelecekte çocukların kariyer planlamasında büyük öneme sahip olacaktır. Aynı zamanda bu çalışmaların merkezi sınavların öğrenciler üzerindeki olumsuz etkilerini azaltma ve başarıyı arttırma konusunda literatüre katkı sağlaması beklenmektedir.

Bilindiği üzere ülkemizde uygulanan 4+4+4 zorunlu eğitim sistemi ile ortaokulu bitiren her öğrencinin lise öğrenimine devam etme zorunluluğu getirilmiştir (Kılıç 2014). Öğrencilerin öğreniminde devam edeceği liselere yerleştirme işlemi iki şekilde gerçekleşmektedir. Bunlardan ilki sınavla öğrenci alan okullara kontenjanları dâhilinde LGS (Liselere Geçiş Sistemi) olarak adlandırılan ve merkezi olarak yapılan sınavdan alınan puan sıralamasına göre yerleştirmenin yapılmasıdır. İkinci yöntem ise sınava girmeyen ya da yerleştirilmesi yapılamayan öğrencileri, tercih ettikleri okul türlerinden (Anadolu Lisesi, Meslek Lisesi veya İmam Hatip Lisesi) adreslerine en yakın sınavsız öğrenci alan okullardan birine yerleştirilmesidir. Bu çalışmada kullanılan veri seti; 2021/2022 eğitim ve öğretim yılında Ağrı, Çankırı ve Edirne illerinde Millî Eğitim Bakanlığı tarafından LGS sınavına girmiş 8.sınıf öğrencilerinin sınav ve e-okulda tutulan çeşitli öğrenim bilgilerinden derlenmiştir. Bu alanda daha önce yapılmış tek örnek olan Ortaöğretime Geçiş Sistemi (OGES) yerleştirme puanlarının uzman sistemler ile tahmini (Uçar 2013) çalışmasında veri seti olarak veri tabanından rastgele seçilmiş 25.000 öğrencinin verisinin kullanıldığı belirtilmektedir.

Bu çalışma veri setindeki parametrelerin geçerli oranda temsil edilmesini garanti altına alması bakımından üç farklı bölgeden seçilen ildeki sınavına giren öğrencilerinin gerçek verileri ile yapılmıştır. Çalışmadaki amaç lise öğrenimlerine yön veren LGS sınavındaki yüzdelik başarı dilimini tahmin etmek ve öğrencinin kendini geliştirebilmesi için güçlü ve zayıf yanlarını ortaya çıkararak kariyer planlamasına katkı sunmaktır. Öte yandan bu alandaki veri madenciliği çalışmalarına katkıda bulunması da amaçlanmaktır.

Bu çerçeveden bakıldığında sınavlardaki başarının doğru şekilde ölçülmesi ve değerlendirilmesi kadar bilgi çağında yapay zekâ teknolojisinin yaygınlaştığı bir ortamda öngörülebilirlik oldukça önem kazanmıştır. Eğitimde niteliğin ve dolayısıyla başarının arttırılması bilgi toplumlarının gelişimi bakımından önemlidir. Ölçme ve değerlendirme teknikleri geliştikçe nicel ve skaler veriye dayalı daha objektif sonuçları üretmek de mümkün hale gelmiştir. Veri madenciliğinin ön plana çıktığı günümüz eğitim sisteminde yapay zekâ ve makine öğrenmesi algoritmaları yardımıyla bireylerin sadece sınav sonuçları ile değil genel yetenekleri, zekâ alanları ve ilgileri doğrultusunda da kariyer planlaması yapılabilmektedir. Bu alandaki çalışmalar gelecekte bizlere eğitim sistemi adına yeni bakış açıları sunacaktır.

1.2 Amaç ve Hedefler

Bu çalışmanın amacı, ülkemizde uygulanmakta olan Ortaöğretim Kurumlarına (Liselere) giriş sınavı için öğrencilerin sınav yüzdelik dilimlerinin makine öğrenmesi algoritmaları kullanılarak tahmin edilmesi, en iyi tahmini üreten model ile hyper-parametrelerin belirlenmesi ve sonuçların tartışılmasıdır. Bu çerçevede belirlenen çalışma hedefleri aşağıdaki gibidir;

- Veri setinin elde edilmesi,
- Veri setinde veri ön işleme adımlarının uygulanması,
- Veri setinde özellik seçim yöntemlerinin uygulanması,
- Veri setine uygun Tek (Single) Tabanlı Makine Öğrenmesi Algoritmaları ile Topluluk (Ensemble) Öğrenmesi Modellerinin belirlenmesi,

- Oluşturulan Modellerin en iyi performans veren Hyper-parametrelerinin tespit edilmesi ve performans metriklerinin elde edilmesi.

1.3 Araştırma Soruları

Bu çalışmanın temel amaç ve hedefleri doğrultusunda aşağıda belirtilen araştırma soruları hazırlanmıştır.

- LGS başarı sonuçlarının tespit edilmesinde tek tabanlı (single base) makine öğrenmesi algoritmaları, topluluk makine öğrenmesi algoritmalarından daha iyi performans sağlayabilir mi?
- LGS başarı sonuçlarının tespit edilmesinde topluluk makinme öğrenmesi algoritmalarından yığınlama (stacking ensemble) modeller, arttırma (boosting ensemble) modellere göre daha iyi performans sağlayabilir mi?
- LGS başarı sonuçlarının tespit edilmesinde topluluk makinme öğrenmesi algoritmalarından yığınlama (stacking ensemble) modeller, torbalama (bagging ensemble) modellere göre daha iyi performans sağlayabilir mi?

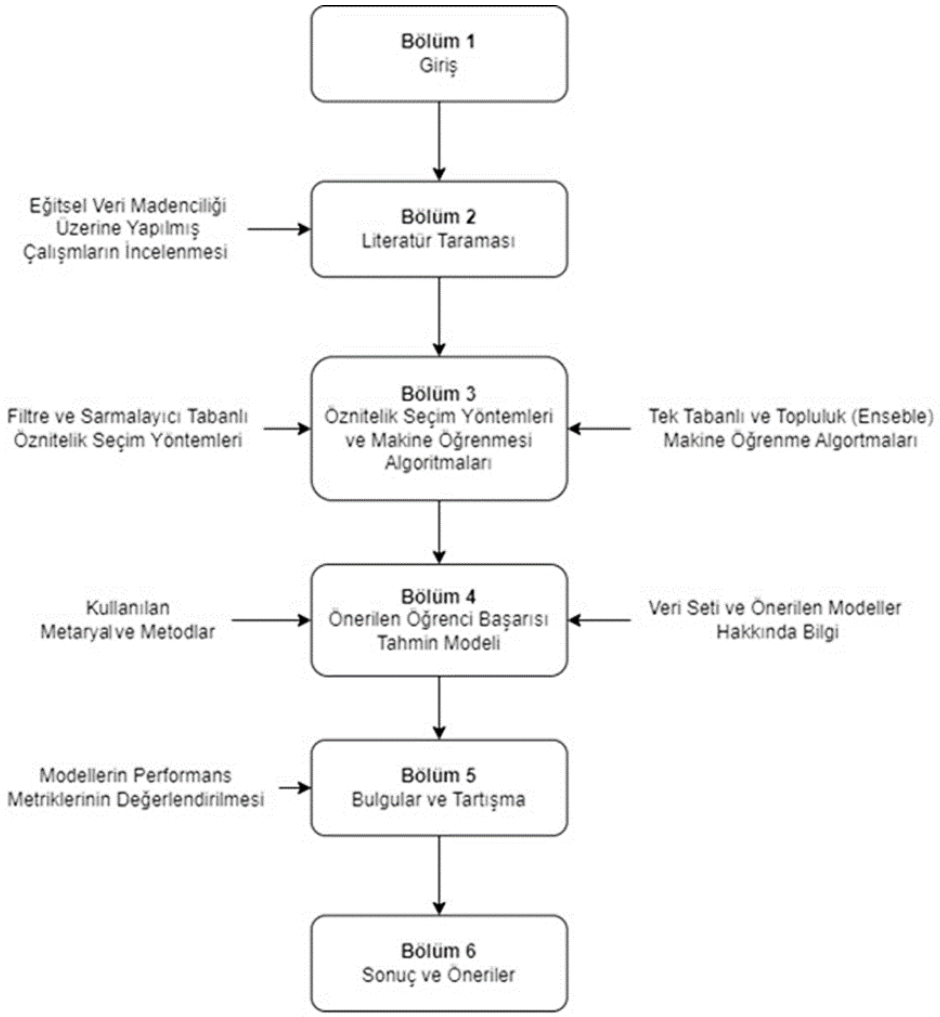
1.4 Çalışmanın Literatüre Katkısı

- Bu çalışmanın öğrencilerin eğitsel verilerine bakılarak merkezi sınavlarda girebilecekleri yüzdelik dilimin tahmin edilmesine, öğrencinin başarısına etki eden parametrelerin tespit edilmesine ve bu sayede öğrencilerin güçlü ve zayıf yönlerini ortaya koyacak modelin geliştirilmesine katkı sunması beklenmektedir.
- Tahmin için öngörülen farklı ML algoritmaları içinden en iyi performans gösteren modelin ve en uygun hyper-parametrelerin belirlenmesidir.

1.5 Tezin Akış Şeması

Tez çalışmamızın ilerleyişi Şekil 1.1'de sunulmuştur. Birinci bölümde, Eğitsel Veri Madenciliği hakkında genel bir bakış sunulmuştur. İkinci bölümde, literatürdeki özellik

seçim yöntemleri ve eğitsel veri madenciliği alanında yapılan çalışmalar incelenmiş, ayrıca makine öğrenimi algoritmalarının başarı tahmini üzerindeki etkileri araştırılmıştır. Bu alanda yapılmış çalışmalarda kullanılan teknikler detaylı bir şekilde incelenmiştir. Üçüncü bölümde, bu çalışmada kullanılan öznitelik seçme teknikleri ve makine öğrenimi algoritmaları detaylı bir şekilde açıklanmıştır. Dördüncü bölümde, önerilen yöntem hakkında bilgi verilmiştir. Veri seti ve veri ön işleme adımları, veri doğrulama ve değerlendirme metrikleri bu bölümde açıklanmıştır. Ayrıca, kullanılan makine öğrenimi algoritmalarının çalışma prensipleri de ele alınmıştır. Beşinci bölümde, önerilen makine öğrenimi algoritmalarının tahmin performansları karşılaştırılmıştır. Altıncı bölümde, çalışmanın sınırları, katkıları ve elde edilen sonuçlar ile gelecekte yapılması planlanan çalışmalara yönelik öneriler sunulmuştur.



Şekil 1.1 Tez akış şeması

2. LİTERATÜR ÖZETİ

Yapay zekâ, insansı zekâya sahip sistemlerin tasarlanması ve geliştirilmesiyle ilgilenen bir disiplindir. İnsanların karmaşık problemleri çözmek, öğrenmek ve kararlar vermek gibi yeteneklerini taklit etmeyi amaçlar. Makine öğrenmesi, istatistiksel ve algoritmik teknikleri kullanarak bilgisayarlara belirli bir görevi gerçekleme yeteneği kazandıran bir yöntem olarak açıklanabilir. Derin öğrenme ise yapay sinir ağlarının kullanıldığı bir makine öğrenmesi yaklaşımıdır. Makine öğrenmesi ve derin öğrenme yapay zekâ alanında önemli bir alt alan olarak ortaya çıkmıştır ve bilgisayar sistemlerinin verilerden öğrenme yeteneklerini geliştirmeyi hedefler. Yapay Zekâ kavramı ilk olarak 1950'lerde Alan Turing'in "Turing Testi" adlı makalesine dayanmaktadır. Turing, bir makinenin düşünme yeteneğine sahip olduğunu kabul edilebilmesi için, insan bir gözlemci tarafından insanla yapılan bir sohbetle ayırt edilememesi gerektiğini öne sürmüştür. Kavram olarak ise ilk kez 1956 Dortmund konferansında John McCarthy tarafından dile getirilmiştir (Arslan 2020). İlk etapta o dönemin sınırlı teknik yeterlikleri nedeniyle istenilen başarıyı elde edemese de ilerleyen dönemlerde özellikle hesaplaması uzun zaman olan matematiksel karmaşık problemlerin çözümünde oluşturulan algoritmik modeller sayesinde sonuçlar elde edilmeye başlanmıştır.

Makine Öğrenmesi, eldeki verileri kullanarak tahmin modellerini otomatik olarak oluşturur ve kararlar alır. Bunun için de regresyon ve sınıflandırma, tekniklerini kullanır. Regresyon, bir bağımlı değişkenin bağımsız değişkenler tarafından nasıl etkilendiğini anlamak için kullanılan temel bir istatiksel yöntemidir. Verilerin trendlerini analiz eder ve gelecekteki değerleri tahmin etmek için bu trendleri kullanır. Sınıflandırma ise verileri belirli sınıflara ayırma işlemidir. Makine Öğrenmesi, sınıflandırma modellerini eğiterek, girdi verilerini doğru sınıflara atayabilir.

Günümüzde yapay zekâ teknikleri öğrenme, akıllı tahminler yapma, karmaşık sorunları çözme, değişen koşullara uyum sağlama, farklı dillere ve deneyimlere uyum sağlama gibi bilgi yönetimi sürecinde doğrudan katkı sağlamaktadır. Eğitim alanında da yaygın olarak kullanılmaya başlanan yapay zekâ sayesinde öğrencilerin sınavlardaki performanslarını tahmin etmede ve başarı düzeylerini değerlendirmede artan bir öneme sahiptir.

Geleneksel yaklaşımda öğrencilerin başarı düzeyleri genel olarak sınav puanları ve ders içi performansları üzerinden belirlenmektedir. Bu yaklaşım değerlendirme adına tek boyutlu bir bakış açısı sunduğundan öğrencilerin gerçek durumlarını tespit etmede ve ihtiyaç duyulan çözümleri sunmada yetersiz kalabilmektedir. Modern yaklaşımda makine öğrenmesi ve yapay zekâ tekniklerinin kullanılması öğrencilerin başarılarını daha fazla parametre ile çok boyutlu analiz etme ve kişiselleştirilmiş çözümler sunma imkânı getirmektedir. Bu çalışmanın ilk ve önemli aşaması öznitelik belirleme ve veri ön işleme konusudur. Makine öğrenmesi ile yüksek performans gösteren bir model oluşturulabilmesi için veriyi oluşturan özniteliklerin doğru belirlenmesi ve veri ön işleme adımlarının dikkatle yapılması gerekmektedir. Öğrencilerin başarısını tahmin etmek için belirlenecek öznitelikler, öğrencinin profili (yaş, cinsiyet, sosyo-kültürel kimliği vb.) ile akademik verileri (sınav notları, devamsızlık sayısı vb.) gibi parametreler olabilir. Ayrıca modelin yüksek performans üretmesinde önemli bir yere sahip bir diğer unsur da veri ön işleme adımlarıdır. Bu adımlar eksik verilerin doldurulması, veri normalizasyonu, aykırı değerlerin tespiti ve ele alınması gibi işlemlerden oluşur. Makine öğrenmesi ile analiz çalışmasının ikinci aşama ise modelin oluşturulmasıdır. Elde edilmek istenen sonuca göre Sınıflandırma ya da Regresyon yöntemleri tercih edilir. Eğitimde başarının tahmin edilmesinde yaygın olarak kullanılan bir makine öğrenmesi yöntemi sınıflandırmadır. Sınıflandırma algoritmaları, öğrencilerin belirli bir kategoride (örneğin başarılı veya başarısız) sınıflandırılmasına yardımcı olur. Bu algoritmalar, eğitim verilerini kullanarak öğrencilerin sınıflandırma etiketlerini tahmin etmek için örüntüleri bulmaya çalışır. Sınıflandırma için yaygın olarak kullanılan algoritmalar arasında Naive Bayes (NB), Karar Ağaçları (RF), Destek Vektör Makineleri (SVM) ve K-En Yakın Komşu (KNN) gibi yöntemler bulunmaktadır. Bu algoritmalar, eğitim verilerinden öğrenilen modelleri kullanarak yeni öğrencilerin sınıflandırmasını yapabilir. Eğitimde başarının tahmin edilmesinde bir diğer kullanılan makine öğrenmesi yöntemi regresyondur. Regresyon, bir bağımlı değişkenin (örneğin öğrencinin not ortalaması) bağımsız değişkenlere (örneğin sınav puanları, devamsızlık sayısı, öğrenci profili vb.) bağlı olarak tahmin edilmesine odaklanır. Regresyon analizi, eğitim verilerinden öğrenilen bir modeli kullanarak, öğrencilerin başarı düzeylerini tahmin etmek için kullanılır. Lineer Regresyon, Lojistik Regresyon ve Doğrusal Olmayan Regresyon gibi regresyon yöntemleri bu alanda sıkça kullanılan tekniklerdir.

Özetle Yapay Zekâ, bilgisayarların belirli bir problemi çözmek için önceden belirlenmiş her adımı takip etmek yerine, algoritmalar aracılığıyla nasıl çözeceğini öğrenmelerini sağlamayı amaçlamaktadır. Bu, önceden belirlenmiş yolu izleyen programlardan ziyade, dinamik olarak ortaya çıkabilecek sorunlara daha önce öğrenilmiş çok sayıda içerik kullanarak çözüm sağlayan algoritmaların geliştirilmesi anlamına gelir (Gürsakal 2017). Yapay zekâ potansiyeli ve gidebileceği nokta hayal edilenin çok ötesindedir. Bu teknolojinin eğitime farklı açılardan güç kattığı ve öğretim sürecinde karşılaşılan sorunların aşımında daha büyük bir ivme yaratacağı kesindir (Arslan 2020).

2.1 Öğrenci Başarısında Özellik Seçimi Önemi

Günümüzde birçok kuruluşta her gün terabaytlarca veri üretilmektedir. Millî Eğitim Bakanlığı da büyük veri depolayan ve işleyen kamu kurumlarından biridir. Böylesi büyük hacimlerdeki veri tabanlarında depolanan veri başlıkları aynı zamanda özel veri madenciliği teknikleri ile ortaya çıkartılabilecek gizli ve tahmine dayanan stratejik bilgiler de içermektedir. Kurumlar, stratejik planlama süreçlerinde veri madenciliğinin önemini fark etmiş ve veri madenciliği teknikleri ile başarılı sonuçlar almaya başlamışlardır (Lee and Siau 2001).

Eğitimde başarının tahmin edilmesi için kullanılan makine öğrenmesi yöntemlerinin etkin olabilmesi için doğru özniteliklerin seçilmesi ve veri ön işleme adımlarının doğru bir şekilde uygulanması önemlidir (Tosunoğlu vd. 2021). Öznitelikler, öğrencilerin performanslarını ve karakteristiklerini temsil eden ölçülebilir değişkenlerdir. Veri ön işleme ise, veri setinin analiz için hazır hale getirilmesini sağlayan bir dizi işlemi içerir. Buna veri ön işleme çalışması denilmektedir. Bu çalışma birkaç işlem basamağından oluşmaktadır. Özniteliklerin doğru bir şekilde seçilmesi, başarı tahminleme modellerinin kalitesini etkileyen önemli bir faktördür. Öznitelik seçimi, gereksiz veya gürültülü özniteliklerin çıkarılması ve enformasyon içeren özniteliklerin belirlenmesiyle ilgilidir. İyi seçilmiş öznitelikler, modelin doğruluğunu artırabilir ve aşırı uyum (overfitting) sorununu önleyebilir.

Öznitelik seçim yöntemleri arasında bilgi kazancı, korelasyon analizi, öznitelik sıralaması ve öznitelik önemi değerlendirme gibi yöntemler bulunmaktadır. Bir diğer işlem veri temizlemedir. Veri setindeki hatalı, eksik veya tutarsız verilerin düzeltilmesi veya çıkarılması işlemidir. Bu adım, modelin doğru sonuçlar üretmesi için önemlidir. Veri temizleme aşamaları arasında eksik verilerin doldurulması, aykırı değerlerin tespiti, veri düzeltme ve veri standardizasyonu yer almaktadır. Ayrıca, öznitelik değerlerinin farklı ölçeklerde olması durumunda ortaya çıkabilecek yanlılığı gidermek için ölçekleme adı verilen farklı bir veri ön işleme adımı da uygulanır. Ölçekleme işlemi, öznitelik değerlerini belirli bir aralığa veya dağılıma dönüştürerek verilerin karşılaştırılabilir hale gelmesini sağlar. Ölçekleme yöntemleri arasında normalizasyon, standartlaştırma ve min-max ölçekleme gibi teknikler bulunmaktadır. Bir diğer işlem ise öznitelik çıkarımıdır. Bu işlem var olan özniteliklerden yeni özniteliklerin elde edilmesi işlemidir. Bu adım, veri setindeki bilgiyi artırmak veya daha iyi bir temsil sağlamak için kullanılır. Öznitelik çıkarımı yöntemleri arasında temel bileşen analizi (PCA), faktör analizi, birleştirme (aggregation) ve ayrıştırma (decomposition) gibi teknikler bulunmaktadır. Özniteliklerin doğru bir şekilde seçilmesi ve veri ön işleme adımlarının uygulanması, eğitimde başarının tahmin edilmesinde kullanılan makine öğrenmesi modellerinin kalitesini artırabilir. Doğru öznitelik seçimi, gereksiz bilgiyi azaltırken önemli bilgiyi korur. Veri temizleme, hatalı veya eksik verilerin analiz sonuçlarını etkilemesini engeller. Ölçekleme, farklı ölçeklerdeki verileri uyumlu hale getirirken öznitelik çıkarımı, daha iyi bir veri temsili sağlar. Bu adımların doğru bir şekilde uygulanması, başarı tahmini üreten modellerinin doğruluğunu ve etkinliğini artırabilir.

Bir modelin performansını belirlemenin en iyi yollarından biri de eşit koşullarda diğer modeller ile karşılaştırmaktır. Bunun için farklı makine öğrenmesi tekniklerini ve algoritmalarını oluşturup karşılaştırma yoluna gidilmiştir. Ancak iyi performans konusunda bilinen en yaygın görüşlerden birisi de daha iyi/üstün makinelere sahip olmanın performansa etkisidir. Benzer performanstaki makineleri kullanıyor olduğu halde fark yaratan faktörün özellik oluşturma ve özellik seçimi olduğu görülmektedir. Diğer bir ifadeyle, başarının anahtarı, uygun değişkenlere sahip olmak ve kullanılan tahmin modeli için en etkili değişkenleri seçebilmektir (Kaushik 2023).

Doğru bilgiye doğru zamanda sahip olmak, doğru kararları vermek bazen hayati önem taşımaktadır. Çoğu kuruluş için büyük bir endişe kaynağı olan veri toplama sorunu günümüzde neredeyse çözülmüştür. Gelecek yüz yılda, kuruluşlar veri toplama konusunda değil mevcut veriden bilgi üretmek konusunda yarışır hale geleceklerdir (Baker 1998). Eğitim alanındaki bir başka çalışmada veri madenciliği teknikleri ve makine öğrenmesi algoritmaları kullanımı öğrenciyi tanıma, öğrenme ortamını anlama ve eğitim verilerini keşfetme bakımından önemli rol oynadığından bahsedilmektedir. Son zamanlarda özellikle kamu kurumları dışındaki akademi kurumları, son derece rekabetçi bir ortamda faaliyet göstermektedir. Eğitim modeli belirleme, performans analizi, kaliteli eğitim hizmeti ve stratejik planlama konularında oldukça önemli aşama kaydetmişlerdir. Çalışma sonuçları incelendiğinde, Makine öğrenmesi teknikleri kullanılmak suretiyle öğrencilerin risk altında olup olmadığı ve öğrencilerin okulu bırakma olasılıkları tahmin edilmeye çalışılmıştır. Ayrıca çalışmada makine öğrenmesi yöntemleri ile yapılan analizler ile öğrencilerin performanslarını artırmak suretiyle kurumlarının da etkinliğinin artacağı belirtilmektedir (Albreiki *et al.* 2021). Bu nedenle geleceğin eğitim sistemi mimarisinde Eğitim Veri Madenciliği (EVM) büyük öneme sahip olacak ve bu alandaki çalışmalar yeni yaklaşımların ortaya çıkmasını sağlayacaktır.

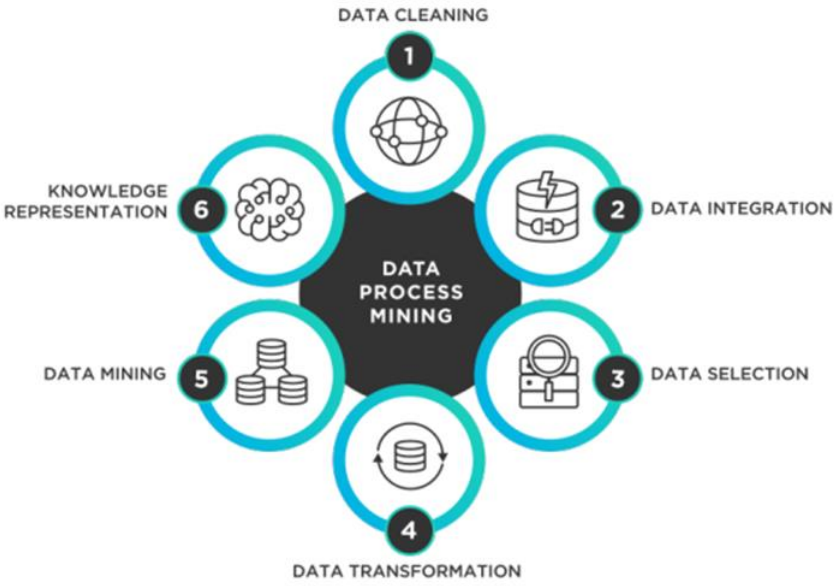
Obenshain (2004) tarafından veri madenciliği tekniklerinin sağlık verilerine uygulanması üzerine yapılmış çalışmada veri madenciliği, istatistik, veri tabanı teknolojileri, desen tanıma ve makine öğrenmesi çalışmalarının veriyi görselleştirme ve uzman sistemlerin ara yüzünü meydana getirme konusunda önemli role sahip olduğunu ifade etmektedir. Veri tabanları genellikle bir sorguyu veya işlevi içerir. Veri madenciliği ise eldeki verilerin birbiri ile ilişkisinden yeni veriler üretme işidir. Geleneksel istatiksel analiz yöntemleri genellikle belirli sayıda önceden tanımlanmış değişkene ihtiyaç duyulurken, veri madenciliği teknikleri yeni değişkenleri kendisi üretebilir ve çok fazla değişkeni kolayca işleyebilir. Bu sayede veri madenciliğinin bir süreç yönetim aracı olarak görülebileceğini söyler. Bunu da SEMMA (örneklem, keşfetme, değiştirme, modelleme ve değerlendirme) adı verilen bir yöntem ile ortaya koymaktadır. SEMMA yöntemi; Sample, Explore, Modify, Model ve Assess olarak adlandırılan beş aşamadan oluşur. Mimarinin yapısı; (1) örneklem- istatistiksel olarak temsil edici bir veri örneği çekmek, (2) keşfetmek- keşif, istatistiksel analiz ve görselleştirme tekniklerini uygulamak, (3)

değiştirmek (veya manipüle etmek)- en önemli değişkenleri seçmek ve dönüştürmek, (4) modellemek- tahmin sonuçları üretmek için model oluşturmak ve (5) değerlendirmek- modelin doğruluğunu test etmek şeklindedir. Sürecin her aşaması çıktı üretir, grafik ve tablo görüntüleri istenilen zamanda istenilen formatta ilgili kişilere sunulmak üzere oluşturulabilir (Obenshain 2004). Veriyi keşfetmek veri madenciliği bakımından çok önemli bir işlem basamağıdır. Doğru öznitelik seçimlerini belirlemek doğru tipte ve miktarda veri toplamak başarımı yükseltmede önemli yer tutmaktadır. Bunun için de çeşitli öznitelik seçim (feature selection) metotları kullanılmaktadır. Özellik seçimi, veri boyutunu azaltarak veri madencilik performansını artırmada kullanılan önemli bir önişleme adımıdır.

Özellikle, veri madenciliği, makine öğrenimi ve desen tanıma topluluklarında birçok özellik seçimi algoritması bulunmasına rağmen, hala aktif bir araştırma alanıdır. Son zamanlarda veri boyutlarındaki öznitelik ve değişken sayısının aşırı derecede (binlerce) artması nedeniyle, birçok öznitelik seçim algoritması etkililik ve verimlilik açısından ciddi zorlukla karşı karşıya kalmaktadır. Bu konuda yazılan bir makalede, bazı mevcut popüler öznitelik seçim algoritmalarını analiz etemkte ve bu algoritmaların güçlü yönlerine ve zorluklarına değinmektedir. Mevcut öznitelik seçim algoritmaları arasında bazıları yalnızca bağımlı değişkeni etkileyeni özniteliklerin seçimine odaklanırken gereksiz tekrarlanan değişkenlerin varlığına dikkat etmez. Tekrarlayan değişkenler nedeniyle gereksiz boyut artışı oluşur ve öğrenme performansını olumsuz etkiler. Bazı algoritmalar ise gürültülü verilerin varlığını dikkate almadan ilgili öznitelikleri seçerler. Gürültülü verinin varlığı, öğrenme performansını da etkiler ve hesaplama zamanını artırır. Bu tip çalışmalar, gereksiz ve gürültülü veri olmayan en iyi özellik ve alt küme seçimine sahip nitelikte bir çerçeve kullanımı gerekliliğine işaret etmektedir. Oluşturulan çerçeve, tüm veri türlerinde uygulanabilir ve boyut artışıyla ölçeklenebilir olmalıdır (Sutha and Tamilselvi 2015). Eğitim kurumları ve organizasyonlarında veri madenciliği kullanılarak çok sayıda çalışma yapılmış ve ciddi başarı elde edilmesi sebebiyle insanlar bu alana giderek daha fazla ilgi göstermeye başlamştır. İlgili çalışmada, veri setine yakın zamanda geliştirilmiş çokgen konik fonksiyonlar, sınıflandırma algoritması ve matematiksel analizler uygulanmıştır. Çalışma MATLAB ve WEKA'da uygulamaları ile yapılmıştır. Veri madenciliğinin en önemli çalışma alanlarından biri de büyük veri (Big Data) veri

madenciliğidir. Big Data veri madenciliği aynı zamanda Veri tabanında Bilgi Keşfi (VBK) olarak da adlandırılır. Bu alanda sıklıkla; denetimli ve denetimsiz sınıflandırma, karar ağaçları, sinir ağları, Naive Bayes, kümeleme gibi teknikler kullanılır (Irsoy *et al.* 2012). Örneğin, matematikte öğrenci başarısını öngörmek için mevcut bir sınıflandırma modeli küçük değişikliklerle önerilmiş ve kullanılan algoritmaya göre belirli bir doğrulukla, öğrencinin matematik dersini geçip geçemeyeceği söylenebilmektedir. Bu sonuçlara bakılarak eğitimciler ve öğrenciler ders başarısını artırmak adına süreci yönetebilecekleri veriyi elde edebilirler. Önerilen algoritma, farklı veri kümeleri için ikili sınıflandırma tekniği önermekte olup eğitimi alanındaki farklı konularda da kullanılabilir.

Bir başka çalışmada yine veri madenciliği teknikleri kullanılarak matematik sınıfındaki öğrencilerin başarıları belirlenmektedir. Başka bir deyişle, öğrenci başarıları ile öğrenci notları, demografik, sosyal bilgiler ve okul ile ilgili veriler gibi çeşitli öğrenci ilişkili veriler arasında herhangi bir bağlantı olup olmadığını belirlemek için veri madenciliği teknikleri kullanılmaktadır. Tahmin algoritmları olarak Karar Ağacı (DT), Ayırıcı Analiz (DA), Destek Vektör Makineleri (SVM), K-NN ve Topluluk Öğrenme Algoritmaları (Ensemle Learning Algorithms) kullanılmaktadır. Elde edilen sonuçlar umut vericidir ve öğrencilerin eğitim başarısını etkileyen ana değişkenleri belirlemektedir. Bu amaçla veri madenciliği tekniklerinden DT, DA, SVM ve ensemble learning yöntemleri genellikle daha iyi performans gösterirken, K-NN yöntemleri genellikle en kötü sonuçları vermiştir (Şengür ve Karabatak 2018). Şekil 2.1 ‘de Veri madenciliği metadolijisinin döngüsel yapısı gösterilmektedir.



Şekil 2.1 Veri madenciliği metodolojisinin tipik döngüsü (Şengür ve Karabatak 2018)

Abbasoğlu (2020) ise çalışmasında Eğitim Veri Madenciliğini, öğrenme ortamlarındaki benzersiz veri türlerini keşfetmek ve bu verilerle öğrenciler ile öğrenme ortamlarını daha iyi anlama üzerine çalışan yeni bir disiplin olarak tanımlamaktadır. Eğitsel veri madenciliğini bilgisayar, eğitim ve istatistik bilimlerininlerin bir kombinasyonu olarak ifade etmektedir. İlgili çalışmada; öğrencilerin demografik özelliklerinin ve sosyo-ekonomik durumlarının öğrencilerin genel başarı performansları üzerindeki etkilerini analiz etmektedir. Bu amaçla, 2019-2020 akademik yılının ikinci döneminde, Yalova ilindeki 5, 6, 7 ve 8. sınıflardan 1395 ortaokul öğrencisinin sosyo-demografik özelliklerine ilişkin veriler kullanılmış. E-Okul Yönetim Bilgi Sisteminden elde edilen veriler ile yapılan çalışmada sınıflandırma teknikleri ve algoritmaları ile yılsonu genel performans ortalamasının tahmin edilmesi amaçlanmıştır. 27 bağımsız değişkenin kullanıldığı çalışmada en yüksek korelasyon katsayısı veren sekiz algoritma seçilerek değerlendirme yapıldığı belirtilmektedir. Deneysel sonuçlara göre, genel başarı ortalaması tahmininde sınıflandırma yöntemleri başarılı sonuçlar verdiği ve genel başarı ortalaması tahmininde PCA kullanılarak yapılan analizde, Logistic sınıflandırma algoritması en iyi performansı gösterdiği belirtilmiştir.

2.2 Öğrenci Başarısını Tahmin Etmede Yaygın Olarak Kullanılan Makine Öğrenmesi Algoritmaları

Öğrenci başarısının tahmin edilmesi, eğitim alanında önemli bir araştırma konusudur. Yaygın olarak büyük veri (Big Data) içindeki gizli desenleri ve bağlantıları ortaya çıkarmak için kullanılan istatiksel çalışmaların bütününe veri madenciliği denilmektedir, Eğitsel veri madenciliği ise, eğitim alanındaki verileri inceleyen ve bu verilere dayanan uygulamaları kapsayan veri madenciliği alanıdır. Veri madenciliği, genelde finansal işlemler, bankacılık, sağlık, lojistik ve güvenlik gibi çeşitli alanlarda kullanılmaktadır. Bu alanda yürütülen veri madenciliği çalışmlarında müşterilerin tüketim alışkanlıklarına yönelik kümeleme, sınıflandırma ve birliktelik analizleriyle bir başka ifadeyle ilişki madenciliği yapılarak kişiye özel müşteri profili oluşturulmaktadır. Müşteri profilinin özelleştirilmesi kişiye özel teklif, süreç ve çözüm yönetimi konularında hızlı ve başarılı sonuçlar almak adına büyük önem taşımaktadır. Aynı yaklaşımın eğitim verileri üzerinde öğrenci profillerini çıkartmak suretiyle uygulanması böylece başarıyı arttıracak izleme ve değerlendirme yöntemlerinin sunulması veri madenciliği çalışma alanlarından biri haline gelmiştir. Bir çalışmada bu vurgu ile yola çıkılmış ve bir lisenin öğrenci verilerinin veri madenciliği yoluyla analizi yapılarak öğrencilerin ayrı ayrı profillerini oluşturmak amaçlanmıştır. Öğrencilerin performansını doğrudan etkileyebilecek sağlık koşulları, finansal kaynaklar, yaşam standartları ve ailelerin eğitim düzeyi gibi bir dizi özellik dikkate alınmıştır. Bu amaçla, veri tabanındaki 443 öğrencinin verileri analiz edilerek bir veri ambarı oluşturulmuştur. İlişki kuralı madenciliğinin popüler algoritmalarından biri olan Apriori algoritması veri analizi için kullanılmıştır. Apriori algoritması, %90'ın üzerinde doğruluğa sahip 72 kural üretebilmiştir. Üretilen kuralların öğrencilerin profilini oluşturmada yardımcı olabileceği, okul yönetimi, öğretmenler, veliler ve öğrencilerin çalışmalarına katkı sağlayabileceği düşünülmektedir.

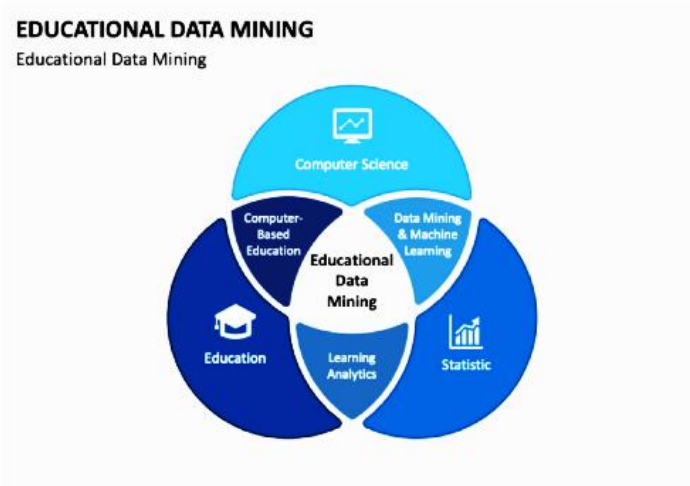
Kullanılan veri madenciliği teknikleri gelişen yapay zekâ teknolojisi sayesinde makine öğrenmesi algoritmalarını ve geniş içerikte eğitim verilerini kullanarak öğrenci performansını tahmin etmek için kullanılan matematiksel modellere dönüşmüştür. Bunlar sınıflandırma ve regresyon modelleri olarak ikiye ayrılmaktadır. Yaygın olarak kullanılan sınıflandırma ve regresyon algoritmaları; Karar Ağaçları, Destek Vektör Makineleri

(SVM), K-En Yakın Komşu (KNN) ve Yapay Sinir Ağları (ANN)'dır. Makine öğrenmesi algoritmaları, öğrenci performansını tahmin etmek için kullanılan önemli yöntemlerin başında gelir. Bu algoritmalar, öğrencilerin karakteristiklerini ve geçmiş performanslarını temsil eden veri setlerini analiz ederek tahmin yapar. Öğrenci başarısında yaygın olarak kullanılan makine öğrenmesi algoritmaları ayrıntılı olarak bu bölümün geri kalan kısmında incelenecektir.

Öğrenci başarısının tahmin etmede yaygın olarak kullanılan makine öğrenmesi algoritmaları arasında destek vektör makineleri (SVM), karar ağaçları, rastgele ormanlar, yapay sinir ağları ve k-NN yer almaktadır. Bu algoritmalar, öğrenci özniteliklerini kullanarak başarı tahmini yapmak için veri odaklı yaklaşımlar sunmaktadır. Her bir algoritmanın avantajları ve kısıtlamaları bulunmaktadır, bu nedenle kullanılan veri seti ve probleme bağlı olarak en uygun algoritmanın seçilmesi önemlidir.

Başer vd. (2020) tarafından Ortaöğretimde Öğrenim Gören Öğrenci Performanslarının Veri Madenciliği Yöntemleri ile Tahmin Edilmesi üzerine yapılan çalışmada okul raporları, anketler ve UC Irvine Machine Learning Repository veri tabanından elde edilmiş 31 öznitelikten oluşan veri kümesi kullanılmıştır. Öğrenci veri seti içeriğindeki özniteliklerden başlıcaları ders notları, devamsızlık, demografik unsurlar ile sosyal ve kültürel aktiviteleridir. Bu çalışmadaki analizler veri madenciliği yöntemlerinden olan Iterative Classifier, OneR, LogitBoost ve Yapay Sinir Ağları kullanılarak yapılmıştır. Bunnlardan en iyi performans gösteren yöntem, OneR Veri Madenciliği yöntemi olmuş ve bu yöntemle 5, 7 ve 10 katlı çapraz doğrulama deneyleri yapılmıştır. Deneyler sonucunda, doğruluk değerinin %92,15'i, duyarlılık değerinin %96'sı ve özgüllük değerinin %86,6'sı elde edilmiştir. Öğrencileri başarı düzeylerinin artırılması için öncelikle her birinin kişisel performansına etki eden faktörlerin tespit edilmesi ve gerekli iyileştirmelerin yapılması önerilmektedir. Bu sayede öğrenciler önerilen alanlar dışındaki konulara yoğunlaşmak durumunda kalmayacaktır. Verimli bir süreç yönetimiyle sadece performansın artması değil iş ve zaman kaybının da azalması öngörülmektedir. Gelecekte farklı veri madenciliği yöntemleriyle de deneylerin tekrarlanabileceği ve modellerin geliştirilebileceği beyan edilmektedir.

Eğitsel veri madenciliğinin gelişim alanlarından biri olan akademik kurumlardaki öğrenci performansının tahmini alanında veri madenciliği teknikleri kullanılmak suretiyle zengin çeşitlilikte dijital veri üretilmektedir. Öğrenci başarısının tahmin edilmesi ile ilgili yapılan bir başka araştırmada özellikle 10. ve 12. sınıf ile önceki dönem notları kullanılarak bir tahmin sistemi önerilmektedir (Dhilipan *et al.* 2021). Çalışma, Binomiyal mantıksal regresyon, Karar ağacı, Entropi ve KNN sınıflandırıcıları kullanılarak değerlendirilmiştir. Bu sistem, öğrencinin son notunu tanımasına yardımcı olarak akademik davranışlarını geliştirmeleri için öğrencinin daha yüksek bir puan elde etmesini amaçlamaktadır. Bu çalışmada kullanılan veri tabanındaki veri boyutunun büyüklüğü nedeniyle modelin tahmin verisi üretmekte zorlandığı görülmüştür. Öte yandan, öğrencilerin başarısını değerlendirmek ve takip etmek için belirlenmiş bir çerçevenin eksikliği de göz önüne alındığında öğrencilerin performansını tahmin etmede en uygun yöntemi belirlemek için bilinen yöntemler üzerine yapılan araştırmalar hala yetersiz kabul edilmektedir. Bu kapsamdaki çalışmalar, tahmin algoritmaları kullanarak öğrenci verilerindeki en ilgili öznitelikleri sınıflandırmaya odaklanmıştır. Eğitim verileri ile makine öğrenimi yöntemleri kullanılması öğrencilerin performansını daha fazla artırma potansiyeline sahip olunduğunu göstermektedir. Öğrenciler, eğitimciler ve akademik kurumlar bu sonuçlardan faydalanabilir ve etki yaratabilirler. Bu analizler için, ikili lojistik regresyon, karar ağacı, entropi ve KNN sınıflandırıcısı kullanılmıştır (Dhilipan *et al.* 2021). Şekil 2.2.'de Eğitsel veri madenciliği, Eğitim-Öğretim, İstatistik ve Bilgisayar bilimlerinin birbiri ile olan ilişkisini görülmektedir.



Şekil 2.2 Makine öğrenmesi ve eğitsel veri madenciliği (Pallathadka *et al.* 2021)

Bugünün rekabetçi dünyasında, eğitim kurumlarının öğrenci performansını tahmin etmesi, bireyleri yeteneklerine göre sınıflandırması ve gelecekteki testlerinde performanslarını artırmaya çalışması oldukça önemlidir. Şekil 2.2'de gösterildiği gibi makine öğrenmesi eğitimsel veri madenciliği alanında önemli bir role sahiptir. Eğitim sektörüne tahmin yapma kabiliyeti sunar. Sadece tahmin modelleri oluşturmada değil, birbirine bağımlı olan ders grubunu belirlemede de yardımcı olmaktadır. Bu bilgi, öğrencilerin başarılı olmalarında kendileri için hangi derslerin daha önemli olduğunu belirlemelerinde yardımcı olmaktadır. (Pallathadka *et al.* 2021). Bu uygulamalar, öğrenciler ve öğretmenlerden gelen verilerin analizi ile ilgilenir. Analiz, kategorizasyon veya tahmin için kullanılabilir. Yapılmış bu deneysel çalışmada UCI makine öğrenimi öğrenci performansı veri seti kullanılmıştır. Algoritmalardaki doğruluk ve hata oranı gibi belirli parametreler analiz edilmektedir. Bulgulara göre, öğrencilerin yetenekleri ve ilgileri performanslarıyla bağlantılı olabileceği düşünülmektedir. Bu tür analiz sonuçları öğretmenlerin öğrencilerine ihityaçları doğrultusunda odaklanmalarına olanak tanır. Öte yandan eğitim kurumları öğretmenin başarısını sıklıkla öğrencilerinin performanslarına göre ölçmektedir. Bu nedenle, öğrenci analizleri öğretmenler için de önem kazanmaya başlamıştır. Bu tür analizler, kurumların eğitim kalitesini artırmasına da yardımcı olur.

Çeşitli makine öğrenimi algoritmalarını kullanılarak öğrenci performans verilerinin sınıflandırılması ve tahmini üzerine yapılan bir başka çalışmada; Nave Bayes, ID3, C4.5 ve SVM gibi makine öğrenme algoritmaları üzerine araştırmalar yapılmıştır. Bu deneysel incelemede, UCI makine öğrenci performansı veri seti kullanılmaktadır. Algoritmalar, doğruluk ve hata oranı gibi özelliklere dayanarak değerlendirilmiştir. SVM, öğrenci performansı veri setini sınıflandırmak için en doğru teknik olduğu görüldüğü söylenmektedir (Pallathadka *et al.* 2021).

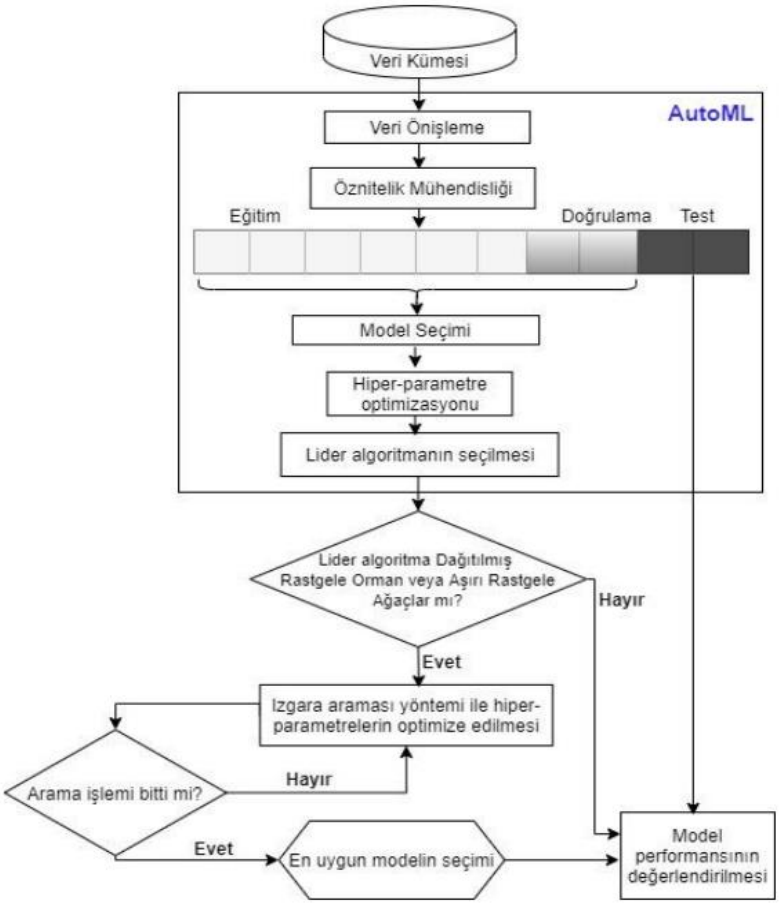
Eğitimsel veri madenciliğinin çalışma alanları arasında öğretmenlere geri bildirim sağlama, öğrencilere ders önerileri sunma, istenmeyen öğrenci davranışlarını belirleme ve öğrencilerin akademik performanslarını artırma gibi konular yer almaktadır. Bu alanda farklı makine öğrenmesi teknikleriyle yürütülen başka bir çalışmada öğrencilerin akademik performanslarının tahmin edilmesi için AutoML tekniği kullanılmıştır. Şekil 2.3'te AutoML akış şeması gösterilmektedir. AutoML yöntemi veri ön işleme, model

sınıfının seçimi, tahmin edicinin seçilmesi ve eğitimi, performansın değerlendirilmesi ve sonuçların yorumlanması işlemlerini otomatik olarak gerçekleştirerek belirli görevlerde en iyi performansı elde etmeyi amaçlamaktadır. Çünkü doğru modeller oluşturularak bu alanlarda yapılacak iyileştirmelerle eğitimin kalitesi artırılabilir. Doğru modelleri oluşturmak için uygun makine öğrenimi algoritmalarının seçimi, öğretmenler ve veri bilimcileri için son derece önemlidir. Şekil 2.3'te AutoML iş akış şeması gösterilmektedir.



Şekil 2.3 Auto ML iş akış şeması (Aghalarova and Keser 2022)

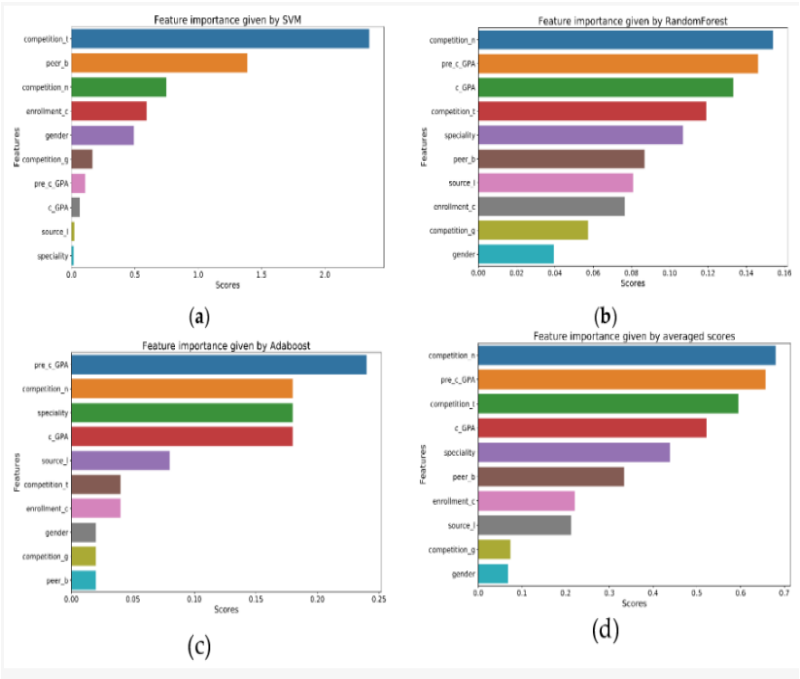
Kullanıcı müdahalesi olmadan veri önişleme, öznitelik seçimi, model seçimi ve modelin hiper-parametrelerinin ayarlanması bu yöntem ile otomatik olarak gerçekleştirilmektedir. Çalışmada, kullanılan gerçek dünya veri kümesi için en iyi algoritma olarak Dağıtılmış Rastgele Orman algoritması belirlenmiştir. Ayrıca, algoritmanın hiperparametreleri grid arama yöntemi kullanılarak optimize edilmiştir. Deney sonuçlarında, Dağıtılmış Rastgele Orman algoritmasının varsayılan hiperparametreleri için doğruluk ve f-skoru değerleri sırasıyla %77.50 ve %80.01 olarak elde edilmiştir. Grid arama ile hiperparametre için hesaplanan en uygun doğruluk ve f-skor değerleri ise sırasıyla %82.30 ve %82.50'dir görülmektedir (Aghalarova and Keser 2022). Şekil 2.4'de AutoML yönteminin akış şemasını göstermektedir.



Şekil 2.4 Önerilen yöntemin akış diyagramı (Aghalarova and Keser 2022)

Eğitim verisi madenciliği en gelişmiş teknikleri, yöntemleri ve uygulamalarını kullanarak öğrenme ortamını ilerletmede kritik bir rol oynar. Son dönemde geliştirilen yöntemler, öğrenci öğrenme ortamını keşfederek, eğitim verileri üzerinde makine öğrenimi ve veri madenciliği teknikleri kullanarak öğrenci performansını analiz etme imkânı sunar. Modern akademik kurumlar son derece rekabetçi ve karmaşık bir ortamda faaliyet göstermektedir. Kurumsal performans, hizmet kalitesi, başarı tahmini ve değerlendirme stratejileri ile gelecek planlama konularındaki rehberlik çalışmları akademik kurumlarının karşı karşıya kaldığı yaygın zorluklardan bazılarıdır. Bu kurumlarda öğrencilerin deneyimledikleri sorunları aşmak için öğrenci müdahale planları uygulanmalıdır (Albreiki *et al.* 2021). Son zamanlarda veri toplama sistemleri ve performans ölçme araçlarındaki gelişmeler sayesinde, eğitim analizleri daha etkili ve hızlı biçimde yapılabilmektedir.

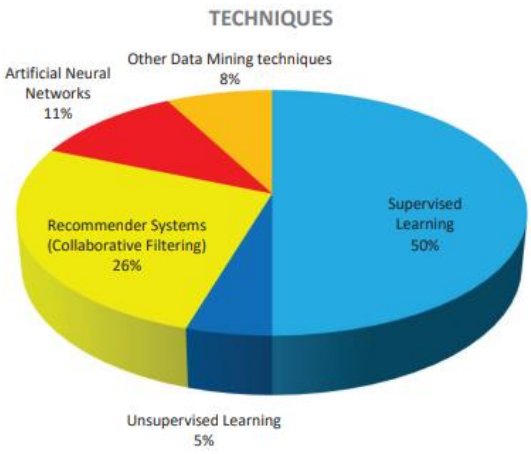
Gelişmiş veri madenciliği ve makine öğrenmesi büyük veri analitiği alanını doğuran teknikler olarak bunları analiz etmede ve izlemede de önerilen yöntemlerdir. Bu sayede sorunları çözmek için öğrencilere, öğretmenlere ve eğitimcilerine zamanında geri bildirim sağlayan çözümler üretebilecektir. Gelecekteki araştırmalar, ML tabanlı performans tahmin yöntemini pratik bir şekilde uygulayabilecek etkili bir topluluk (ensemble) yöntemiyle öğrencilerin performanslarını tahmin etmeye ve öğrencilere erken aşamalarda yardımcı olmaya yönelik müdahale yöntemlerini sunacaktır. (Albreiki *et al.* 2021) tarafından yapılmış olan çalışmada doğru tahmin performansına sahip birkaç algoritma (SVM, RF ve AdaBoost) Temel-öğrenici olarak uygulanmış ve aşırı uyum riskini azaltmak için nispeten basit bir algoritma Meta-öğrenici olarak benimsenmiş ve etiket sızıntısını da önlemek için 10 × 10 katlı çapraz doğrulama kullanılmıştır. Hankou Üniversitesi'nden toplanan verilere dayanan ampirik araştırma, önerilen ensemble modelinin diğer modellere göre daha iyi tahmin performansına sahip olduğunu göstermektedir. Ayrıca, Şekil 2.5'te verilen grafikte Support Vector Machine (SVM), Random Forest (RF) ve Adaboost algoritmalarında Tahmin değerine Özniteliklerin bağınmsız olarak etkileri ve ortalama değerleri gösterilmektedir.



Şekil 2.5 Destek vektör makinesi (SVM) (a), Rastgele orman (b), AdaBoost (c) tarafından verilen özellik önem düzeyleri ve bu üç modelden alınan puanların ortalama değerleri (d) (Albreiki *et al.* 2021)

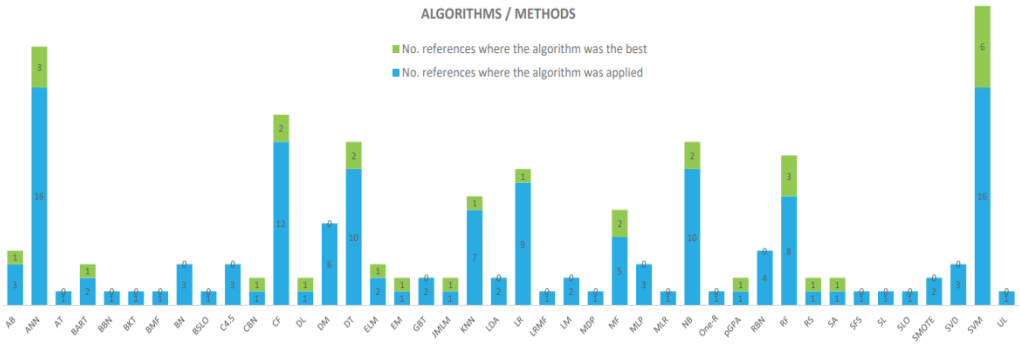
Öğrencilerin performanslarının tahmin edilmesi son zamanlarda popüler bir araştırma konusu haline gelmiştir. Uygulanan tahmin modellerinin çoğu makine öğrenmesi yöntemiyle oluşturulmaktadır. Bu modeller tahminlerin doğruluğuna odaklanırken yorumlanabilirliğe daha az önem göstermektedirler. (Yan *et al.* 2020) çalışmalarında öğrenci performansını iki simetrik kategorik sınıfa ayırmaktadır. Doğruluğu artırmak için destek vektör makinesi (SVM), rastgele orman (RF) ve AdaBoost gibi üç makine öğrenimi algoritması oluşturmakta ve ardından yığınlama (Stacking) tekniğiyle lojistik regresyon entegre edilmektedir. Önemli değişkenleri belirlemek için bir öznitelik önem analizi uygulanır. Deneysel veriler, Hankou Üniversitesi'ndeki dört akademik yıl verisinden oluşmaktadır. Bu analizde beş değerlendirme metriği (kesinlik, duyarlılık, F1, hata ve alıcının işletme karakteristik eğrisi altındaki alan (AUC)) üzerine karşılaştırmalı çalışmalar yapıldığında önerilen modelin genel olarak karşılaştırılan modellerden daha iyi performans gösterdiği belirtilmektedir. Analizlerde belirlenen önemli değişkenler yorumlanabilir olduğundan potansiyel olarak müdahaleye ihtiyaç duyan öğrencileri seçmek için bu çalışma bir rehber olabilecektir (Yan *et al.* 2020).

Rastrollo-Guerrero *et al.* (2020) tarafından eğitsel veri madenciliği alanında yapılmış çalışmalar üzerine geniş bir araştırma yapılmak suretiyle kullanılan tüm farklı teknikleri ortaya koymak için neredeyse 70 makalenin analiz edildiği görülmektedir. İncelenen makalelerde kullanılan tekniklerin dağılımı Şekil 2.6'da gösterilmektedir.



Şekil 2.6 Kullanılan tekniklerin çalışmalara göre oranı (Rastrollo-Guerrero *et al.* 2020)

Bu incelemede toplanan verilere dayanarak, öğrencilerin davranışlarını tahmin etmek için en yaygın kullanılan teknik olarak doğru ve güvenilir sonuçlar sağladığı için denetimli öğrenme tekniği olduğu özellikle SVM algoritması yazarlar tarafından en çok kullanılan ve en doğru tahminleri sağlayan algoritma olduğu belirtilmektedir. Şekil 2.7'de Yapay Sinir Ağları ise daha az kullanılan bir teknik olmasına karşın öğrencilerin performansını tahmin etmede büyük bir kesinlik elde ettiğini göstermektedir.

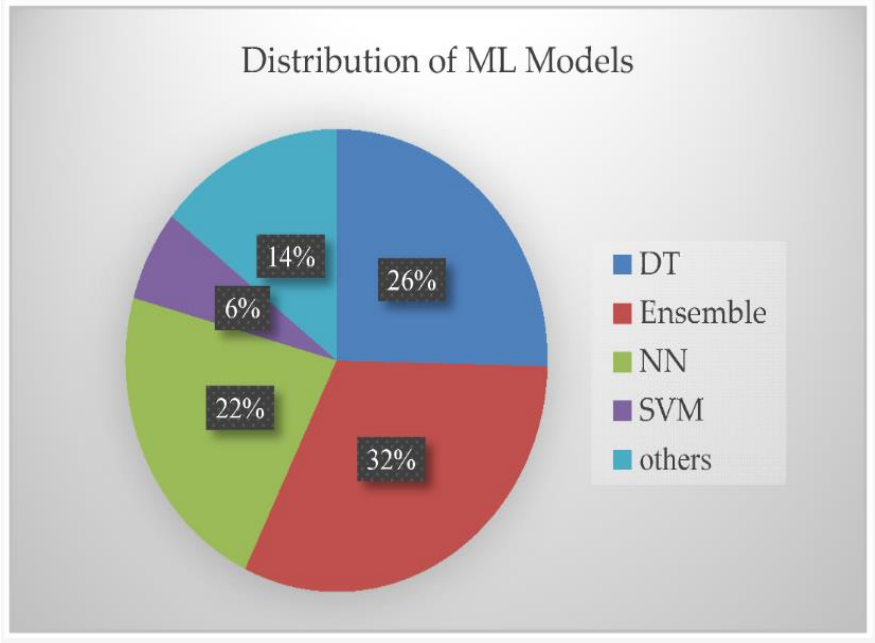


Şekil 2.7 Literatür taramasında ele alınan teknikler, hedefler ve algoritmalar hakkında temel istatistikler (Ghorbani and Ghousi 2020)

Yine benzer bir çalışma olarak (Ghorbani and Ghousi 2020) tarafından yayınlanmış makalede dengesiz veri kümelerinin öğrenci performansının tahmin edilmesinde ciddi bir zorluk oluştuduğu belirtilmektedir. Bu konuda farklı örneklem yöntemleri arasında bir karşılaştırma bulunmadığı hususuna değinilmekte ve dengesiz veri kümesi sorununun çözümünü ele almaktadırlar. Portekiz ve İrandan alınan farklı veri setlerine Borderline SMOTE, Random Over Sampler, SMOTE, SMOTE-ENN, SVM-SMOTE ve SMOTE-Tomek gibi çeşitli örnekleme teknikleri uygulanarak sonuçları karşılaştırılmıştır. Makalede, Rastgele Orman, K-En Yakın Komşu, Yapay Sinir Ağı, XG-boost, Destek Vektör Makinesi (Radyal Temel Fonksiyon), Karar Ağacı, Lojistik Regresyon ve Naif Bayes gibi çeşitli makine öğrenmesi sınıflandırıcılarını kullanıldığı ve model doğrulama teknikleri olarak da Rastgele Ayırma ve Karışık 5-kat çapraz doğrulama yöntemleri kullanıldığı bildirilmektedir. Farklı değerlendirme metriklerini kullanarak elde edilen sonuçlarda daha az sayıda sınıf ve nominal özniteliklerin modellere daha iyi performans sağlayacağını göstermiştir. Öte yandan yine aynı çalışmada istatistiksel bir anlamlılık testi olan Friedman testi sonuçlarına bakıldığında SVM-SMOTE yönteminin diğer

örnekleme yöntemlerine göre daha etkili olduğunu doğrulamaktadır. Ayrıca Rastgele Orman (RF) sınıflandırıcısın ile SVM-SMOTE yöntemi kullanılarak diğer tüm modellere göre en iyi sonucu elde ettiği görülmektedir.

Balaji *et al.* (2021) tarafından yapılan bir çalışmada ise Ensemble modellerin öğrenci başarısı tahmin etmede en yaygın kullanılan teknik olduğunu ortaya koymaktadır. Şekil 2.8'de ML Algoritmalarının oransal olarak kullanım sıklığı gösterilmektedir.



Şekil 2.8 Kullanılan makine öğrenimi yöntemlerinin grafik gösterimi (Balaji *et al.* 2021)

Araştırmanın başlangıcında 2700 makalenin değerlendirilmeye alındığı çeşitli dâhil etme ve dışlama kriterleri kullanılarak makalelere kalite puanları verildiği ve bu sayede ileri analiz için 56 makaleye kadar sayının filtrelendiği belirtilmiştir. Araştırmacılar tarafından önerilen gelecek yönelimleri kullanılan özelliklerin incelenmesi, kullanılan veri tabanı, uygulanan algoritmalar ve özelliklerin sınıflandırılması konularına göre belirlenen özellikler demografik, akademik ve davranışsal özellikler olarak sınıflandırılmış ve nihayetinde yalnızca bu özelliklere sahip 34 makale üzerinde istenilen incelemenin gerçekleştiği belirtilmektedir. Araştırma neticesinde her ML tabanlı algoritmaların aynı prensipte çalışsalar da kullanılan veri türlerine bağlı olarak farklı davrandığı

belirtilmektedir. Bu nedenle ML algoritmasının sınıflandırma veya tahmin sonuçları için modelin yeterince eğitilmelisi ve öyle kullanılması önerilmektedir. (Balaji *et al.* 2021).

Eğitsel veri madenciliği kapsamında gerçekleştirilen bir başka çalışmada, lisans öğrencilerinin final sınavı notlarını tahmin etmek amacıyla makine öğrenimi algoritmalarına dayanan yeni bir model önerilmektedir (Yağcı 2022). Bu modelde, kaynak veri olarak öğrencilerin vize notları kullanılmaktadır. Makine öğrenimi algoritmaları arasında yer alan rastgele ormanlar, en yakın komşu, destek vektör makineleri, lojistik regresyon, Naïve Bayes ve k-en yakın komşu algoritmalarının performansları hesaplanmış ve karşılaştırılmıştır. Veri seti 2019-2020 güz döneminde Türk Dili-I dersini alan 1854 öğrencinin akademik başarı notlarını içermektedir. Sonuçlar önerilen modelin %75'e varan oranda sınıflandırma doğruluğu elde ettiğini göstermektedir. Tahminler sadece üç tür parametre kullanılarak yapılmıştır. Bunlar vize sınavı notları, bölüm verileri ve fakülte verileridir. Bu tür veriye dayalı çalışmalar yükseköğrenimde öğrenme analizi çerçevesinin oluşturulması ve karar verme süreçlerine katkı sağlaması açısından çok önemlidir. Ayrıca bu çalışma başarısızlık riski yüksek öğrencilerin erken tespitine olanak sunarken en etkili makine öğrenimi yöntemlerinin belirlenmesine de katkı sunmaktadır (Yağcı 2022).

Eğitsel Veri Madenciliği (EVM) alanında yapılan bir diğer çalışmada özellikle sınıflandırma problemlerinin çözümünde yaygın SVM algoritmasının olarak tercih edildiği ancak literatür bakımından doğrusal destek vektör makinelerinin öğrenci performansını öngörmede alanında kullanımındaki bir boşluğu işaret etmektedir. Öğrenci performansının tahminini iyileştirecek algoritmayı belirlemek için doğrusal destek vektör makinelerinin performansı en son klasik makine öğrenmesi algoritmalarının performanslarıyla karşılaştırılmıştır. Bu nicel çalışmada 1000 alfa-sayısal öğrenci kaydından oluşan genel olarak erişilebilir bir veri kümesinde özellik seçimi kullanılarak deneyler yapıldığından bahsedilmektedir. Bu çalışmayı farklı kılan en temel unsur ise ırk, cinsiyet ve öğle yemeği gibi özelliklerin matematik dersi performansını etkilediği, öğle yemeğine erişimin ise okuma ve yazma performansını etkileyen temel faktör olduğunu göstermesidir. Ayrıca bu çalışmada doğrusal destek vektör makineleri 10 farklı makine öğrenmesi algoritmasıyla karşılaştırıldığında öğrenci performansını öngörmede üstün bir

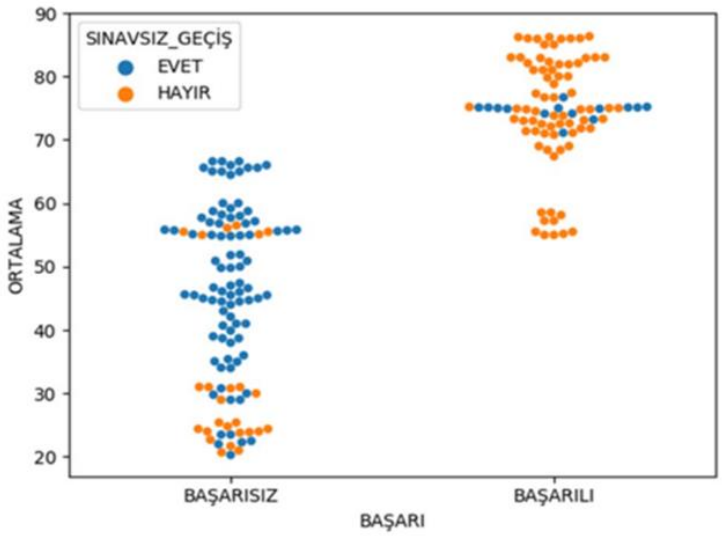
performans sergilemiştir (Naicker *et al.* 2020). Çalışma sonuçları bakımından öğrenci performansını etkileyen faktörleri sınıflandırmak ve tahmin etmek için doğrusal SVM'nin diğer algoritmalardan daha yüksek bir doğrulukla performans gösterdiğini ortaya koymaktadır.

Yapılan bir başka çalışmada ise 22 öznitelikten oluşan bin satırlık bir öğrenci veri kümesi kullanılarak öğrenci performansı üzerine tek tabanlı ve topluluk (Ensemble) makine öğrenmesi algoritmaları ile analizler yapılmıştır (Singh 2020). Çalışmada modellerin tahmin performansını kıyaslamak ve daha önceki tahmin performanslarına göre daha yüksek doğruluk elde edilmek istenmiştir. Bunu yaparken Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbors (KNN) ve Extra Tree (ET) olmak üzere dört farklı makine öğrenimi algoritması uygulanmış ve daha sonra Bagging ve Boosting ensemble yöntemleri kullanılarak her bir temel öğrenme algoritmasının sonuçları birleştirilmiş ve model bu şekilde oluşturulmuştur. Bagging ve Boosting ensemble teknikleri kullanılarak elde edilen bu model ile ortak modelin en iyi sonuçları kullanılmıştır. Tüm algoritmalarının ve ensemble tekniklerinin sonuçları doğruluk, hassasiyet, özgüllük ve f1-puanı gibi çeşitli faktörlerle test edilmiştir. Sonuçlar karşılaştırıldığında Boosting ensemble tekniği ile yapılan modelin Bagging modele kıyasla daha iyi sonuç verdiği görülmüştür. Geliştirilen model seçilen ders için performans tahmini yapmak için başvuru yapan öğrencilere uygulanabilir ve hem öğrenciler hem de kurumlar için faydalı olabilir (Singh 2020).

Gök (2017) tarafından makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi üzerine yapılmış olan bir araştırmada özellikle ülkemizde eğitim alanında gerçekleştirilen projeler ve fiziksel yatırımlara rağmen öğrencilerin ulusal ve uluslararası sınavlardaki başarıları gerekli düzeyde artmadığına vurgu yapılmaktadır. Sebeplerinin makine öğrenmesi yöntemleriyle analiz edilmesi ve erken müdahale edilmesi saysesinde başarının artacağı düşünülmektedir. Bu çalışmada özellikle Türkçe ve Matematik dersleri için bu başarısızlık durumu daha ciddidir. Ortaokul öğrencileri arasında akademik başarıyı etkileyen birçok ailevi, demografik ve okul temelli faktör bulunmaktadır. Ancak bugüne kadar hangi faktörlerin daha etkili olduğu ve başarıyı doğrudan etkilediği net bir şekilde belirtilmemiştir. Bu çalışmada belirlenen kriterlere göre hazırlanan 24 soruluk bir

anketin ortaokul 6, 7 ve 8. sınıf öğrencilerine uygulanmasından elde edilen veri setinin kullanıldığı bildirilmektedir. Anket sonuçlarından elde edilen veriler üzerinde regresyon ve sınıflandırma makine öğrenme modelleri kullanarak Türkçe, Matematik dersleri ve genel başarı ortalamalarının notları ve puanlarının tahmin edilmiş. GB ortalaması tahmininde rastgele orman regresyon yöntemi, not tahmininde ise KÖAK öznitelik seçme yöntemi ile lojistik sınıflandırma algoritmasının birlikte uygulandığı modelin en iyi performansı gösterdiği görülmüştür.

Bingöl Üniversitesi Teknik Bilimler Meslek Yüksekokulu ve Genç İlçe Meslek Yüksekokulu öğrencilerinin akademik performansları üzerinde yapılan çalışmada denetimli öğrenme (supervised learning) ve denetimsiz öğrenme (un-supervised learning) algoritmaları kıyaslanmakta ve veri görselleştirme (data visualization) yapılarak multidisipliner bir çalışma ortaya konulmaktadır. Ön-lisans programı olarak geçen Meslek Yüksekokulu öğrencilerinin bu okullara alım (meslek liselerinden doğrudan sınavsız geçiş) şeklinin başarı üzerindeki etkileri incelenmektedir. Çalışma sonucunda, öğrencilerin ortaöğretim eğitimini aldıkları illerde ön lisans programlarına devam etmelerinin öğrenci başarısını etkileyen en önemli faktörlerden biri olduğu ortaya çıkmıştır (Aydoğan ve Karcı 2019). Şekil 2.9'de öğrencilerin Meslek Yüksek Okuluna kayıt şeklinin başarılarına etkilerinin dağılımı gösterilmektedir.



Şekil 2.9 Kayıt şekline göre öğrenci başarı dağılımları (Aydoğan ve Karcı 2019)

Günümüz eğitim sisteminde öğrenci başarısı aynı zamanda öğretmenlerin de performanslarını belirleyen temel bir performans göstergesi olarak değerlendirilmektedir. Öğretmenlerin öğrenci merkezli eğitim sistemindeki etkinliğini arttırmada kişiselleştirilmiş çözümlere ihtiyaç duyulmaktadır. Özellikle son on yılda araştırmacılar eğitim alanında birçok makine öğrenmesi tekniği varyasyonu önermişlerdir. Özellikle kullanılan faklı veri setlerinin oluşturduğu dengesiz veri kümeleriyle başa çıkmak adına öngörüsel doğruluğun performansını artırarak derslerindeki nihai öğrenci notlarını tahmin etmeye çalışmışlardır. Makine öğrenmesi tekniklerinin analizinin yapıldığı diğer bir çalışmada 1282 gerçek öğrenci ders notu veri kümesi kullanılmıştır. Çalışmada Karar Ağaçları (Decision Tree), Destek Vektör Makinesi (Support Vector Machine), Naïve Bayes, K-En Yakın Komşu (K-Nearest Neighbor), Logistic Regression ve Rastgele Orman (Random Forest) algoritmaları ile Synthetic Minority Over-Sampling Technique (SMOTE) adı verilen çok sınıflı bir tahmin modeli kullanılmaktadır. Önerilen model aşırı uyumlanma ve dengesiz veri sorununu ortadan kaldırmak için çok sınıflı sınıflandırma algoritmaları ile örneklem arttırıcı özellik seçim yöntemleri birlikte kullanılmaktadır. Elde edilen sonuçlar, önerilen modelin RF ile entegrasyonunun, %99,5'lik en yüksek f-ölçüsü ile önemli bir iyileştirme sağladığını göstermektedir. Bu önerilen model, öğrenci notu tahmini için dengesiz çok sınıflı sınıflandırma için tahmin performans modelini geliştirebilecek umut verici sonuçlar göstermektedir (Bujang *et al.* 2021). Yapılan analizler elde edilen skorlar önerilen teknikler kadar bunların görselleştirilmesi de çalışmaların anlaşılırlığını arttırmak bakımından önem arz etmektedir.

Eğitimsel veri madenciliği, eğitim ortamlarında depolanan büyük miktardaki eğitsel verinin analiz edilmesi ve karar verme süreci için öğrencilerin akademik performansını erken dönemde tahmin etme ve eğitimsel verideki gizli desenler ile stratejik bilgileri ortaya çıkartmaya yarayan analitik bir araçtır. Ancak bu tip veri kümelerinde sıklıkla rastlanan durum dengesiz veri kümesi sorunudur. Dengesiz bir veri kümesi farklı sınıflara veya kategorilere ait örneklerin sayısal olarak dengesiz dağıldığı durumu ifade eder. Örneğin bir sınıfa ait örneklerin sayısı diğer sınıflara göre çok daha fazlaysa veya çok daha azsa bu durum dengesiz bir veri kümesi olarak adlandırılır. Bu da model performansına olumsuz etki eden ciddi bir zorluktur. Ders notlarına göre öğrencilerin akademik performanslarının tahmini üzerine yapılan bir çalışmada derin yapay sinir ağı,

karar ağacı, rastgele orman, gradyan artırma, lojistik regresyon, destek vektör makinesi ve K-en yakın komşu gibi makine öğrenmesi algoritmaları kullanılmıştır. Bu çalışmada öne çıkan yenilik ise modellerin performansını artırmak için SMOTE, ROS, ADASYN ve SMOTE-ENN gibi yeniden örnekleme teknikleri kullanılması suretiyle dengesiz veri kümesi sorununu çözülmeye çalışılmasıdır. Bunun için çeşitli yeniden örnekleme yöntemleri kullanarak model doğrulaması için iki yaklaşım kullanıldığı görülmektedir. İlk olarak rastgele uzatma ve tabakalı 5 kat çapraz doğrulama ile dengesiz bir veri seti kullanmanın modellerde sınıflandırıcıların hiçbirinin iyi performans göstermediği görülmüş ancak dengeli veri setinde Random Hold-Out yönteminin de kullanılması ile daha iyi sonuçlar elde edildiği gözlemlenmiştir. DNN, SMOTE yöntemini aşırı örnekleme yöntemi olarak kullanırken %89 doğruluk, %89 F1 skoru ve %89 duyarlılıkla diğerlerinden daha iyi performans göstermiştir. Uygulanan bu tekniğin başarısı dengsiz veri sorununu çözmede yeni yaklaşımlar oluşmasına katkı sağlayacaktır.

Öte yandan özellikle Covid-19 salgını ile birlikte hayatımızdaki yerini alan uzaktan (çevrimiçi) eğitim kavramı pek çok tartışma konusunu da beraberinde getirmiştir. Uzaktan Eğitim Teknolojilerindeki ilerlemeler uzaktan öğrenme ortamlarında resimler, şekiller, videolar ve etkileşimli içerik sunmaya yardımcı çeşitli Web2 (etkileşimli uzaktan eğitim) aracının ortaya çıkmasını sağlamıştır. Tüm bu gelişmelere rağmen çevrimiçi eğitim sisteminin verimliliği hala tartışma konusudur. Çünkü eğitime erişim engellerinin tamamen ortadan kaldırılamamış olması, çevrimiçi ölçme değerlendirme araçlarının yeterlilikleri ve ders içi uygulamaların etkinliği elde edilen öğrenci başarılarının görece olup olmadığı gibi henüz tam olarak çözülememiş sorunlar mevcuttur. Bunu çözmek adına eldeki farklı parametreleri içeren veri setleri ile çalışmalar yapılmaktadır. Araştırmacılar çevrimiçi bir eğitimdeki öğrenci başarı performansının önceki oturumlardaki performanslarına ve katılım düzeyine bağlı olabileceğini iddia etmektedir. Ancak literatürde önceki öğrenci performans ve katılım değerlendirmelerinin sonraki değerlendirmelere etkilerini gösterecek yeterli ölçümler bulunmadığı görülmektedir. Bu konudaki farklı bir çalışmada öğrencilerin değerlendirme notları ile son öğrenci performansları olmak üzere iki tahmin modeli tasarlanmıştır. Sonuçlar her iki model için de elde edilebilir sonuçlar ortaya koyduğunu göstermektedir. Değerlendirme notları modelinde en düşük RSME değeri RF ile elde edilen 8.131 değeri olurken son öğrenci

performansı modelinde ise en yüksek doğruluğu 0.086 değeri ile elde etmiştir (Alshabandar *et al.* 2020).

Eğitim verileri kullanarak öğrencilerin başarısı kadar başarısızlığını ya da okulu bırakma olasılıklarını erken ve doğru bir şekilde tahmin etmek öğrenme analitiği araştırma alanının yaygın bir konusu olarak kabul edilir (Koç 2023). Bu konuda özellikle sınıflama algoritmaları kullanılarak yapılmış çalışmalar mevcut olup sonuçlar birçok makine öğrenmesi algoritmasının kıt bir eğitim veri setine de başarıyla uygulanabileceğini göstermektedir. Ancak aynı anda olası bırakma durumlarını tahmin etmek ve yararlı müdahale mekanizmaları tasarlayacak modeli oluşturmak için sınıflandırma performans ölçütlerinin dikkatle düşünülmesi gerektiği ortaya konulmuştur (Kabathova and Drlik 2021).

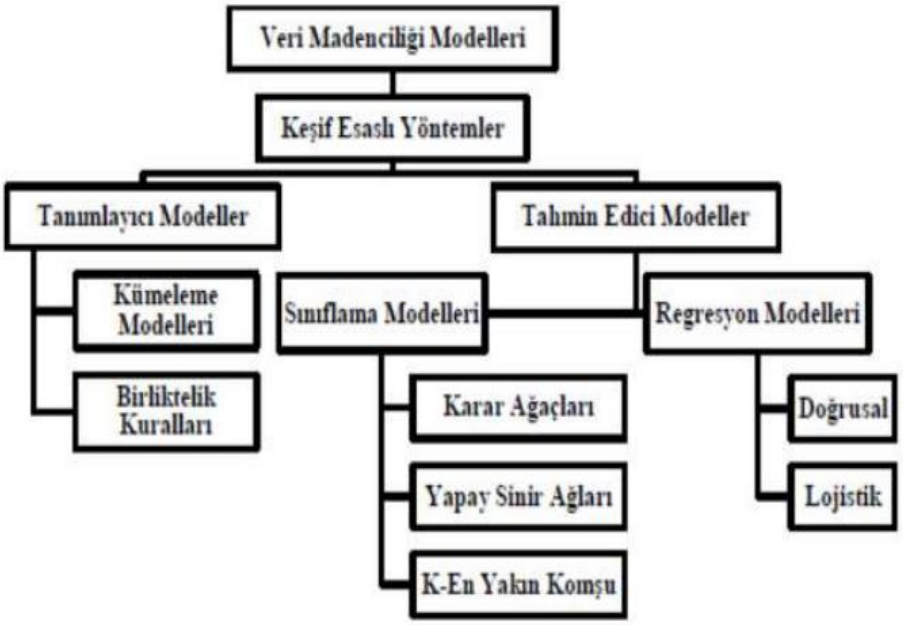
Bu noktadan hareketle bu çalışmada yukarıdaki çalışmalrda kullanılan benzer içerik ve kapsamdaki eğitsel veri seti kullanılarak çoğunlukla sınıflandırma çıktısı üzerine yapılmış araştırmalara LGS sınav başarısı üzerine ve öğrencinin yüzdelik dilimini tahmin etmeye yönelik yeni bir halka eklemeye gayret edilmiştir. Modelde Regresyon algoritmaları tercih edilmesiyle sınav sonucuna en yakın tahminin üretilmesi öngörülmüştür. Bölüm 3'te literatürde yaygın olarak kullanılan özellik seçim yöntemleri ve makine öğrenmesi yöntemleri detaylı olarak açıklanmıştır.

2.3 Özellik Seçim Yöntemleri ve Makine Öğrenmesi Algoritmaları

Her geçen gün artan bilgi ve büyük ölçekteki veri çeşitliliği bunların anlamlı bir şekilde işlenme ihtiyacını ortaya çıkarmıştır. Verilerin çeşitli istatistiksel yöntemlerle düzenlenmesi ve yorumlanması sonucunda ortaya çıkan üst seviye bilgi veri madenciliği olarak adlandırılır (Çağıltay 2010). Başka bir ifadeyle veri madenciliği büyük veri setlerinden bilgi elde etmek için veriyi işleme ve analiz etme sürecidir. Veri madenciliği, bilinen verileri ve analiz sonuçlarını kullanarak gelecekle ilgili tahminlerde bulunmaktadır (Yurdakul 2015). Bu nedenle veri inceleme ve analizi yapabilen yeni teknolojilerin kullanılması kaçınılmaz hale gelmiştir. Günümüzde çok büyük boyutlara ulaşan veriler için mevcut yöntemler ve teknikler yetersiz kalmış ve yeni teknolojilere

ihtiyaç duyulmuştur. Bilgisayar teknolojilerinin kullanımıyla veri toplama, depolama ve işleme yaygınlaşmıştır. Bilgi teknolojilerinde hızlı bir gelişme yaşanmaktadır ve bu gelişmeler, yeni teknolojilerin ortaya çıkması ve maliyetlerin düşmesi gibi önemli faktörlere katkı sağlamaktadır. Bu durum insanların daha hızlı, kullanışlı ve ekonomik bilişim teknolojilerine erişebilmesini sağlamaktadır (Kayhan 2019). Küreselleşen dünyada veriyi toplamak kadar bu veriyi anlamlı hale getirmek ve bunu kullanarak yeni bilgiler elde etmek önemli hale gelmiştir (Can and Gerşil 2017).

Toplanan verilerin çok fazla olması başlangıçta anlamlı sonuçlara ulaşmayı zorlaştırabilir. Veriler arasındaki ilişkilerin algoritmalar aracılığıyla incelenmesi veri anlamlandırmada fayda sağlar. Verilerin anlamlandırılması karmaşık bir süreç olup farklı yöntemler ve teknikler gerektirir (Yurdakul 2015). Veri madenciliğinin temel amacı geçmiş verilere dayanarak davranışları tahmin etmeye yönelik karar verme modelleri oluşturmaktır. İstatistiksel analiz teknikleriyle birlikte yapay zekâ algoritmalarının kullanımı büyük veri yığınlarından fark edilemeyen bilgilerin ortaya çıkarılmasına yardımcı olmaktadır (Yeleğin 2012). Veri setleri işlenerek anlamlı bilgi ortaya çıkarılır. Bu amaçla birçok model kullanılmaktadır. Kullanım amacına ve bağlamına bağlı olarak birçok algoritma bulunmaktadır. Şekil 2.10'de Veri Madenciliği Modelleri "Tahmin Edici" ve "Tamamlayıcı" olarak ikiye ayrılabilir. Tahmin edici modeller sınıflandırma ve regresyon olarak iki gruba ayrılır. Veriler kategorik ise sınıflandırma modeli sürekli değerler ise regresyon modeli kullanılır. Tahmin edici modellerde veri setleri incelenerek geleceğe dair eğilimler analiz teknikleriyle tahmin edilebilir ya da farklı sınıflara ayrılarak çıkarımlar elde edilebilir (Düzgün 2017).



Şekil 2.10 Veri madenciliği modelleri (Düzgün 2017)

2.4 Özellik Seçim Yöntemleri

Genel kanaat olarak makine öğrenimindeki başarının, seçilen algoritmalara ve yüksek performanslı işlemcilere dayandığı düşünülse de aslında doğru seçilmiş (feature selection) ve ön işlemeden geçirilmiş (pre-processing) veri setinin başarımdaki farkı yaratan gerçek unsur olduğu bilinmemektedir. Öznitelik seçimi makine öğreniminde özellikle öznitelik sayısı fazla olduğu durumlarda daha da önemli hale gelir. En iyi sonucu elde etmede veri setindeki her özniteliğin kullanılması gerekmeyebilir. Ortak bağımlılığı yüksek minimum sayıdaki öznitelikleri kullanarak modelin hızını ve doğruluğunu arttırmak mümkündür. Özellikle borsa, bahis gibi büyük veriye (Bigdata) dayanan uygulamalarda bağımlı değişkene en fazla etki eden bağımsız değişkenleri belirlemek modelin eğitim ve tahmin performansını doğrudan etkilemektedir. Feature Selection yöntemleri üç grupta incelenebilir.

2.4.1 Filter-based methods

Filtre yöntemleri özelliklerin alaka düzeyini bağımlı değişkenle olan korelasyonları ile ölçer. Genellikle bir ön işleme adımı olarak kullanılır. Özelliklerin seçimi herhangi bir makine öğrenimi algoritmasından bağımsızdır. Öznitelikler sonuç değişkeni ile diğer değişkenlerin korelasyonlarında elde edilen istatistiksel puanlara göre seçilir. Kullanılan Filtre tabanlı özellik seçim yöntemleri şunlardır;

- Pearson Korelasyonu: Pearson korelasyonu istatistiksel analizde iki değişken arasındaki ilişkinin gücünü ve yönünü ölçmek için kullanılan bir parametrik bir ölçüdür. Pearson korelasyonu, değişkenler arasındaki doğrusal ilişkiyi değerlendirmek için Pearson korelasyon katsayısı olarak ifade edilen bir sayısal değer sağlar. Bu katsayı -1 ile +1 arasında bir değer alır, negatif değerler negatif doğrusal ilişkiyi, pozitif değerler ise pozitif doğrusal ilişkiyi temsil ederken, sıfır değeri de herhangi bir ilişki olmadığını gösterir (Pearson 1895).

- LDA: Lineer diskriminant analizi (LDA) çoklu değişkenlerin sınıflandırılması için kullanılan istatistiksel bir yöntemdir. Bu yöntem veri kümesindeki değişkenler arasındaki ilişkiyi kullanarak farklı sınıflara ait örnekleri birbirinden ayırmayı hedefler. LDA özellikle boyut azaltma yöntemleri arasında yaygın olarak kullanılmaktadır. LDA sınıflar arasındaki ayrımı maksimize eden lineer bir dönüşüm bulmaya çalışır. Bu dönüşüm sınıf içi değişkenliği minimize etmek ve sınıflar arasındaki ayrımı maksimize etmek için optimum olarak hesaplanır. Elde edilen dönüşüm veri noktalarını yeni bir uzayda temsil eder ve bu uzayda sınıfların birbirinden ayrılması daha kolay hale gelir (Fisher 1936).

- PCA: Principal Component Analysis (PCA) çok değişkenli veri analizinde kullanılan bir yöntemdir. PCA veri setindeki değişkenler arasındaki ilişkileri anlamak boyut azaltma ve veri özelliklerini vurgulama amacıyla kullanılır. PCA veri setindeki değişkenlerin birbiriyle olan ilişkilerini temel bileşenlere dönüştürerek gerçekleştirilir. Bu temel bileşenler veri setindeki değişkenlerin varyansını en iyi açıklayan doğrultular olarak tanımlanabilir. İlk temel bileşen veri setindeki en büyük

varyansa sahip doğrultuyu temsil ederken sonraki temel bileşenler sırasıyla azalan varyansa sahip doğrultuları temsil eder. PCA veri setindeki boyutu azaltma yeteneği sayesinde büyük ölçekli veri analizinde sıklıkla kullanılır. Özellikle yüksek boyutlu veri setlerinde veri setinin boyutunu azaltarak daha basit ve anlaşılır bir yapı elde etmek ve gereksiz değişkenlerin etkisini azaltmak hedeflenir (Jolliffe 2002).

- ANOVA: ANOVA (Analysis of Variance) istatistiksel bir yöntemdir ve gruplar arasındaki varyansın analiz edilmesi amacıyla bir veya daha fazla faktörün etkisini değerlendirmek için kullanılır. ANOVA gruplar arasındaki farklılıkları belirlemek ve gruplardaki varyansın hangi faktörlerden kaynaklandığını anlamak için yaygın olarak kullanılan bir araştırma tasarımı ve analiz aracıdır. ANOVA'nın temel amacı istatistiksel olarak anlamlı farkların var olup olmadığını belirlemektir. ANOVA sonuçları p-değerleri ve güç analizi gibi istatistiksel ölçütler kullanılarak yorumlanabilir. ANOVA analizi genellikle sosyal bilimler, psikoloji, ekonomi, tıp ve diğer araştırma disiplinlerinde kullanılan istatistiksel analiz yöntemlerinden biridir (Keenan *et al.* 2012).

- Chi2 (Ki-Kare) Testi: Chi-kare testi (Chi2 testi) istatistiksel analizlerde kullanılan bir yöntemdir. Bu test iki kategorik değişken arasındaki ilişkiyi belirlemek amacıyla kullanılır. İlk olarak gözlenen frekanslar ile beklenen frekanslar arasındaki farkın ölçüsü olan Chi-kare istatistiği hesaplanır. Ardından, bu istatistik serbestlik derecesine bağlı olarak Chi-kare dağılımıyla karşılaştırılır ve ilişkinin istatistiksel olarak anlamlı olup olmadığı belirlenir (Howell 2014).

Filtre yöntemleri çoklu bağlantıyı ortadan kaldırmamaktadır. Bu nedenle, verileriniz için modelleri eğitmeden önce, özelliklerin çoklu bağlantıları da göz ardı edilmemelidir.

2.4.2 Wrapper (sarmalayıcı) based yöntemler

Bu yöntemde özelliklerin bir alt kümesini kullanmaya ve bunları kullanarak bir model eğitmeye çalışılır. Önceki modelden elde ettiğimiz çıkarımlara dayanarak alt kümeye

özellikler eklemeye veya çıkarmaya karar veririz. Filtreleme yöntemleri modellerin eğitimini içermediğinden sarmal yöntemlere kıyasla çok daha düşük maliyetli ve hızlıdır (Jović *et al.* 2015). En sık kullanılan sarmal yöntemler şunlardır;

- İleri Seçim (Forward Selection) modelde seçili hiçbir özniteliğin bulunmadığı durum ile başlayan yinelemeli bir yöntemdir. Her değişken eklendiğinde performans ölçülür ve yineleme yeni bir değişkenin eklendiği halde modelin performansını iyileştiremeyene kadar sürer. Forward Selection tekniği en sık tercih edilen Sarmal yöntemlerden biridir (Jović *et al.* 2015).

- Geriye Doğru Eleme **(**Backward Elemination**)** tüm özelliklerle başlanır ve her yinelemede modelin performansını artıran en önemsiz özelliği kaldırıler. Özelliklerin kaldırılmasında herhangi bir gelişme görülmeyene kadar bu işlem tekrar edilir.s (Jović *et al.* 2015).

- Özyinelemeli Özellik Eleme En iyi performans gösteren özellik alt kümesini bulmayı amaçlayan açgözlü bir optimizasyon algoritmasıdır. Tekrar tekrar modeller oluşturur ve her yinelemede en iyi veya en kötü performans gösteren özelliği bir kenara bırakır. Tüm özellikler tükenene kadar kalan özelliklerle bir sonraki modeli oluşturur. Daha sonra, özellikleri ortadan kaldırılma sırasına göre sıralar (Yan ve Zhang 2015).

2.4.3 Gömülü yöntemler

Filtre ve Sarmal yöntemlerinin niteliklerini birleştirir. Kendi yerleşik özellik seçim yöntemlerine sahip algoritmalar tarafından uygulanır.

2.4.4 Filter ve wrapper yöntemleri arasındaki temel farkları

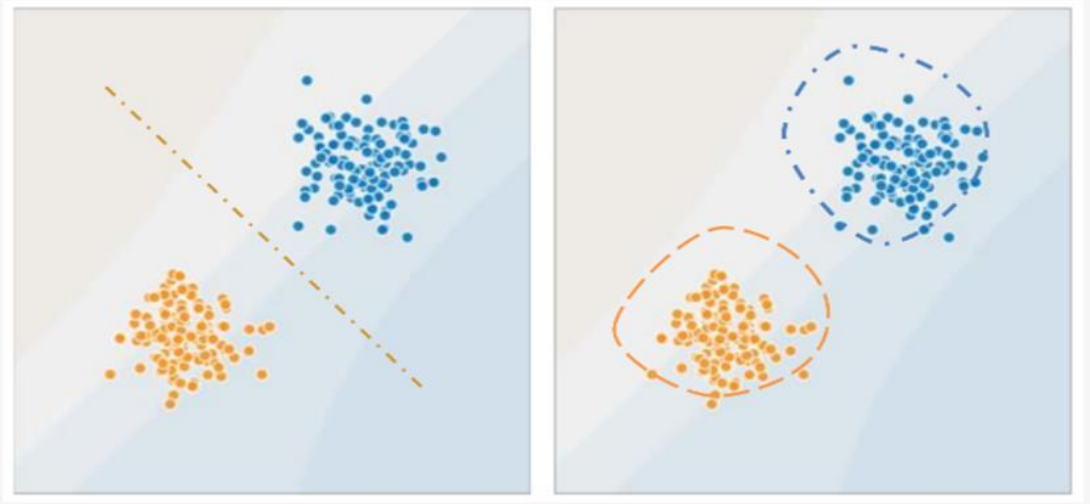
- Filtre yöntemleri özelliklerin alaka düzeyini bağımlı değişkenle olan korelasyonlarıyla ölçerken, sarmalayıcı yöntemler üzerinde bir model eğiterek bir özellik alt kümesinin kullanışlılığını ölçer.
- Filtreleme yöntemleri modellerin eğitimini içermediğinden sarmalayıcı yöntemlere kıyasla çok daha hızlıdır. Öte yandan, sarma yöntemleri de hesaplama açısından çok pahalıdır.
- Filtre yöntemleri bir özellik alt kümesinin değerlendirilmesi için istatistiksel yöntemler kullanırken, sarmalayıcı yöntemler çapraz doğrulama kullanır.
- Filtre yöntemleri birçok durumda özelliklerin en iyi alt kümesini bulmakta başarısız olabilir ancak sarmalayıcı yöntemler her zaman özelliklerin en iyi alt kümesini sağlayabilir.
- Sarmalayıcı yöntemlerdeki özelliklerin alt kümesini kullanmak modeli filtre yöntemlerindeki özelliklerin alt kümesini kullanmaya kıyasla fazla uydurmaya daha yatkın hale getirir (Jović *et al.* 2015).

2.5 Makine Öğrenmesi Algoritmaları

Makine öğrenimi, bir bilgisayar sisteminin veriye dayalı deneyimlerle öğrenmesini sağlayan bir alan olarak tanımlanabilir. Kullanılan veri setine ve istenen çıktıya uygun olarak sınıflandırma, regresyon ve etiketleme olmak üzere üç temel makine öğrenimi algoritması vardır (Nasteski 2017).

- Etiketli veriler: Makine öğreniminde veri etiketleme, ham verileri (resimler, metin dosyaları, videolar vb.) tanımlama ve bir makine öğrenimi modelinin ondan öğrenebilmesi için bağlam sağlamak üzere bir veya daha fazla anlamlı ve bilgilendirici etiket ekleme işlemidir. Çıktılar bu bağlama göre türetilir. Örneğin; Görüntü analiz (Fotoğraftaki nesneleri tanıma, hareketli görüntüde analiz vs.)

- Sınıflandırma: Veri örneğini farklı sınıflara veya kategorilere atama işlemidir. Örneğin bir e-postanın spam veya spam olmayan olarak sınıflandırılması veya bir görüntünün farklı nesnelere ait olduğunun belirlenmesi gibi durumlar sınıflandırma problemleri olarak ele alınabilir. Sınıflandırma algoritmaları, eğitim verilerindeki örneklerden çıkarılan özelliklere dayanarak bir model oluşturur ve bu modeli kullanarak yeni örnekleri sınıflandırır. Genelde ayrık değerleri tahmin etmek için kullanılır. Örneğin; {1,0}, {Doğru, Yanlış}. Sınıflandırma algoritmalarının çalışma mantığı Şekil 2.11'de gösterilmektedir.



Şekil 2.11 Sınıflandırma algoritmaları (Balaji *et al.* 2021)

- Regresyon: Regresyon analizi, veriler arasındaki ilişkileri anlamak ve gelecekteki değerleri tahmin etmek için kullanılan bir araçtır. Örneğin; ev fiyatları, borsa, yılsonu satış vs. regresyon problemleridir. Başka bir ifadeyle, hedef (bağımlı) değişken ve yordayıcılar (bağımsız değişkenler) arasındaki doğrusal ilişki problemleridir. Bu tür problemlerde hedef değişken sürekli bir değer tutar. Regresyon modelleri genelde bu değeri tahmin için kullanılır. Regresyon modelleri doğrusal regresyon, DT, rastgele orman ve SVM gibi algoritmaları içerir. Regresyon problemlerinin performans metrikleri, ortalama mutlak hata (MAE), ortalama karesel hata (MSE), ortalama karesel hatanın kökü (RMSE) ve R-kare hatası olarak tanımlanır. Sıfır MAE değeri hata olmadığını veya mükemmel tahminleri gösterir. Sıfır olarak tahmin edilen bir MSE, tahmin edicinin parametrenin gözlemlerini mükemmel bir doğrulukla tahmin

ettiği anlamına gelir. Ortalama hata karesi hatası (RMSE), tahmin edilen ve gerçek gözlem arasındaki farkların karelerinin ortalamasının karekökünü alarak hatanın ortalama büyüklüğünü ölçer. RMSE her zaman MAE'den büyük veya ona eşit olacaktır; aralarındaki fark ne kadar büyükse, örneklemdeki bireysel hatalardaki varyans da o kadar büyük olur. RMSE = MAE ise, tüm hatalar aynı büyüklüktedir. R-kare puanı, bağımsız değişkenlerden tahmin edilebilen bağımlı değişkendeki varyansın oranıdır. Belirleme katsayısı olarak bilinir. R2'nin değeri 0 ile 1 arasındadır, burada 0 uyum yok ve 1 mükemmel uyum anlamına gelir (Balaji *et al.* 2021).

Kullanılan Makine öğrenmesi modelin yapısına göre de ikiye ayrılır;

- Tek Tabanlı Makine Öğrenmesi
- Topluluk Öğrenme Algoritmaları

2.5.1 Tek tabanlı makine öğrenmesi

Makine öğrenimi modeli belirli kalıp türlerini tanımak üzere eğitilmiş bir dosyadır. Oluşturulan model, veriler üzerinde akıl yürütmesi ve bunlardan öğrenme yapabilmesi için kullanabileceği bir algoritma ve veri kümesi iel eğitilir. Daha sonra model performansı, eğitim için kullanılmamış test verileri ile kontrol edilir. Makine öğrenimi problemlerinde geleneksel temel bir yaklaşım, geçerli bir çözümün uygulanmasında tek bir belirlenmiş algoritma veya model kullanmaktır. Bu geleneksel yöntem bu algoritmanın geliştirilmesi kapsamındaki tüm süreçlerin bir ardışık düzeni oluşturmaya dayanır.

- Linear Regression: Lineer regresyon bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkiyi tanımlayan ve bu ilişkiyi doğrusal bir modelle ifade eden bir makine öğrenmesi algoritmasıdır. Bu algoritma veri setindeki değişkenler arasındaki ilişkiyi temsil etmek ve bağımlı değişkenin tahminini yapmak için doğrusal bir denklem kullanır. Örneğin, bir evin fiyatını etkileyen faktörleri analiz etmek için lineer regresyon kullanılabilir (Maulud and Abdulazeez 2020).

- SVR (Support Vector Regressor): Destek Vektör Regresyonu veri noktalarını bir hiper-süzgece en iyi şekilde uydurarak regresyon problemi olarak çözen bir destek vektör makinesi (SVM) algoritmasıdır. SVR, bir doğrusal regresyon modelinin aksine veri noktalarını sınırlayan bir pay ile bir regresyon doğrusu oluşturur. Bu pay, hedef değişkenin regresyon doğrusuna ne kadar yakın olması gerektiğini belirler. SVR, aykırı değerler ve doğrusal olmayan ilişkiler olduğunda kullanılabilen non-parametrik regresyon yöntemidir (Brereton and Lloyd 2010).

- K-Neighbors Regressor: K-Komşu Regresyonu, bir regresyon problemi için kullanılan bir makine öğrenmesi algoritmasıdır. Bu algoritma bir veri noktasının tahminini yaparken o noktanın en yakın K komşusunu kullanır. K komşunun değerlerini kullanarak tahmin edilecek noktanın değerini hesaplar. K-Neighbors Regressor basit ve esnek bir algoritma olup özellikle komşuluk ilişkilerinin önemli olduğu durumlarda etkilidir (Ray 2019).

- ElasticNet Regressor: Lineer regresyonun gelişmiş bir versiyonudur. Genelde düşük boyutlu veri setlerinde aşırı uyum (overfitting) problemini ele almak için kullanılır. Bu regresyon yöntemi hem L1 hem de L2 düzenlileştirmelerini birleştirerek bir hiperparametre ile kontrol edilen bir karışım modeli kullanır. ElasticNet Regresyonu, değişken seçimi yaparak modelin karmaşıklığını azaltır ve çoklu korelasyonlu değişkenler arasında dağılımı dengelemektedir (Emmert and Dehmer 2019).

- ANN (Artificial Neural Network): Yapay Sinir Ağı, biyolojik sinir ağlarını taklit eden bir makine öğrenmesi modelidir. ANN, birçok düğüm veya nöron ve bu düğümler arasındaki bağlantılardan oluşur. Her bir düğüm girdi verilerini alır bu girdileri ağırlıklarla çarpar ve bir aktivasyon fonksiyonu kullanarak bir çıktı üretir. ANN, geniş bir uygulama yelpazesine sahiptir ve sınıflandırma, regresyon, desen tanıma gibi birçok görevde kullanılabilir (Ray 2019).

2.5.2 Topluluk öğrenme algoritmaları

Modeli eğitirken, tek bir algoritma ile eğitmek yerine, birden fazla algoritma ile eğitmeyi amaçlayan makine öğrenmesi yöntemidir. Modellerin birlikte daha doğru karar vermesi beklenmektedir. Bir başka deyişle ensemble yöntemler; tek bir modele kıyasla daha güçlü ve genellenebilir sonuçlar elde etmek amacıyla birden fazla temel modelin tahmin sonuçlarını birleştirir. Bu yöntemlerin başarısı iki ölçüte göre olur; temel öğrenicilerin (base learner) öğrenme başarısı ve birbirlerinden farklılıklarıdır.

- Decision Tree Regression: Karar Ağacı Regresyon problemleri için de kullanılan bir makine öğrenmesi algoritmasıdır. Bu algoritma veri setini bir ağaç yapısı şeklinde temsil eder. Her iç düğüm bir özelliği veya özelliğe dayalı bir karar kuralını temsil eder. Yaprak düğümleri ise hedef değişkenin tahminini içerir. Karar ağacı regresyonu basit, anlaşılır ve yorumlanabilir bir model sunar. Ayrıca non-lineer ilişkileri de yakalayabilme yeteneğine sahiptir (Ray 2019).

- Random Forest Regression: Rastgele Orman Regresyonu, bir regresyon problemi için kullanılan bir makine öğrenmesi algoritmasıdır. Bu algoritma birden fazla karar ağacının bir araya gelerek oluşturduğu bir orman yapısını kullanır. Her bir ağaç veri setinin rastgele bir örneğini kullanarak eğitilir ve tahminler yapar. Ardından tahminlerin ortalaması veya ağırlıklı ortalaması alınarak final tahmin hesaplanır. Rastgele orman regresyonu karar ağacının yapabileceğinden daha iyi bir performansa ve genellemeye sahiptir (Ray 2019).

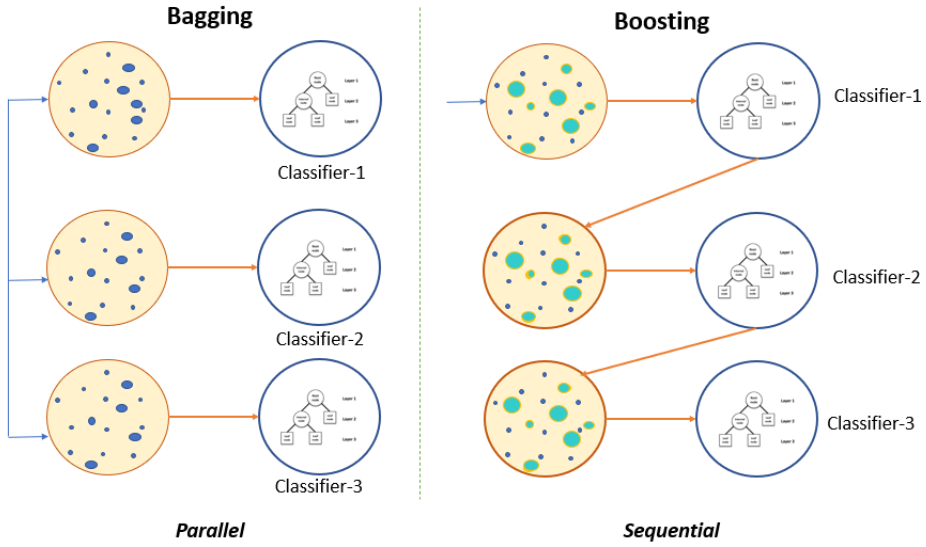
- Extra Trees Regression: Ekstra Ağaçlar Regresyonu, rastgele ormanın bir genişlemesidir. Bu sınıflandırıcı genellikle Aşırı Rastgele Ağaç Sınıflandırıcısı olarak adlandırılır. Farklı türden birbiriyle ilişkili karar ağaçlarının sonuçları, sonuçlarının elde edilmesini sağlamak için ormana yerleştirilir. Bu algoritma, her ağaçta özellik seçimi yapmak için rastgele kesme noktaları kullanır. Rastgele bir örnekleme yöntemiyle ağaçların oluşturulması, modelin daha da çeşitliliğini artırır. Ekstra Ağaçlar, rastgele ormandan daha fazla çeşitlilik ve genelleyebilme yeteneği sunar (Rastogi *et al.* 2023).

- AdaBoost Regressor: AdaBoost Regresyonu, bir regresyon problemi için kullanılan bir makine öğrenmesi algoritmasıdır. Bu algoritma, bir dizi zayıf tahminleyiciyi (genellikle karar ağaçları) birleştirerek güçlü bir tahminleyici oluşturur. Her bir tahminleyici, önceki tahminlerin hatalarına odaklanan ağırlıklar kullanılarak eğitilir. AdaBoost Regresyonu, hataları azaltarak ve veriye odaklanarak tahmin performansını geliştirir (Mayr *et al.* 2014).

- Gradient Boosting Regressor: Bu artırma algoritması, bir hata fonksiyonunu minimize etmek için bir dizi zayıf tahminleyiciyi bir araya getirir. Her bir tahminleyici önceki tahminlerin hatalarını düzeltmek için gradient iniş yöntemini kullanarak eğitilir. Gradyan Artırma Regresyonu, hata azaltma sürecini tekrarlayarak modeli iyileştirir ve daha iyi bir tahmin performansı elde eder (Mayr *et al.* 2014).

- X-Gradient Boosting Regressor: GradientBoosting Regressor algoritmasının geliştirilmiş bir versiyonudur. Bu algoritma özellikle büyük veri kümeleri üzerinde daha hızlı bir şekilde çalışmak için optimize edilmiştir. X-Gradyan Artırma Regresyonu, paralel hesaplama ve özellik blokları kullanarak daha etkili bir model oluşturur. Ayrıca daha fazla esneklik ve iyileştirilmiş tahmin performansı sunar (Kalipe *el al.* 2018).

- Bagging (Tobralama) Yönetemi: L. Breiman tarafından önerilen bu yöntem ile birden çok homojen algoritma bağımsız olarak eğitilir ve daha sonra modelin ortalamasını belirlemek için birleştirilir. Sınıflandırma ve regresyonda kullanılan makine öğrenimi algoritmalarının kararlılığını ve doğruluğunu artırmak için tasarlanmıştır. Varyansı azaltmaya ve aşırı uyumdan kaçınmaya yardımcı olur. Genellikle karar ağacı yöntemlerine uygulanır. Bagging, modelin çalışma mantığı kısaca şöyledir; Orijinal veri kümesinden eşit demetlerle birden çok alt küme oluşturulur ve değiştirilen gözlemler seçilir. Bu alt kümelerin her biri üzerinde bir temel model oluşturulur. Her model, her eğitim seti ile paralel ve birbirinden bağımsız olarak öğrenilir. Tüm modellerden elde edilen tahminler birleştirilerek nihai tahminler belirlenir. Şekil 2.12'te Bagging ve Boosting modellerin iş mantığı arasındaki farklılık gösterilmektedir.



Şekil 2.12 Bagging ve Boosting model arasındaki fark (Chauhan 2021)

- Boosting (Artırma) Yöntemi: Boosting yaklaşımı ise birden çok homojen algoritmayı sırayla eğittiğimiz bir topluluk tekniğidir. Bu bireysel algoritmalar en iyi sonuçlara sahip nihai bir modeli oluşturur. Bir algoritmanın performansı önceden oluşturulmuş algoritmanın performansından etkilenir. Zayıf sınıflandırıcıların sayısından güçlü bir sınıflandırıcı oluşturmaya çalışan bir topluluk modelleme tekniğidir. Zayıf modelleri seri halinde kullanarak bir model oluşturarak yapılır. İlk olarak eğitim verilerinden bir model oluşturulur. Daha sonra birinci modelde var olan hataları düzeltmeye çalışan ikinci model kurulur. Bu prosedür eğitim veri setinin tamamı doğru bir şekilde tahmin edilene veya maksimum model sayısı eklenene kadar devam ettirilir ve modeller eklenir. En bilinen ve yaygın olarak kullanılan Boosting algoritması AdaBoost'dur. Adaboost ikili sınıflandırma amacıyla geliştirilen gerçekten başarılı ilk yükseltme algoritmasıdır. AdaBoost Adaptive Boosting'in kısaltmasıdır ve birden çok "zayıf sınıflandırıcıyı" tek bir "güçlü sınıflandırıcıda" birleştiren çok popüler bir güçlendirme tekniğidir. Kısaca çalışma mantığı şu şekildedir; veri setindeki veri noktalarının her birine eşit ağırlık atanır ve modele girdi olarak alınır. Yanlış sınıflandırılmış veri noktaları belirlenir ve bunların veri noktalarının ağırlığını artırılıp doğru sınıflandırılmış veri noktalarının ağırlıklarını azaltılır. Bu şekilde en iyi başarım elde edilinceye kadar işlem tekrar edilerek tüm veri noktalarının ağırlıklarını normalleştirilir. Çizelge 2.1'te Bagging ve Boosting modeller arasındaki farklar gösterilmektedir.

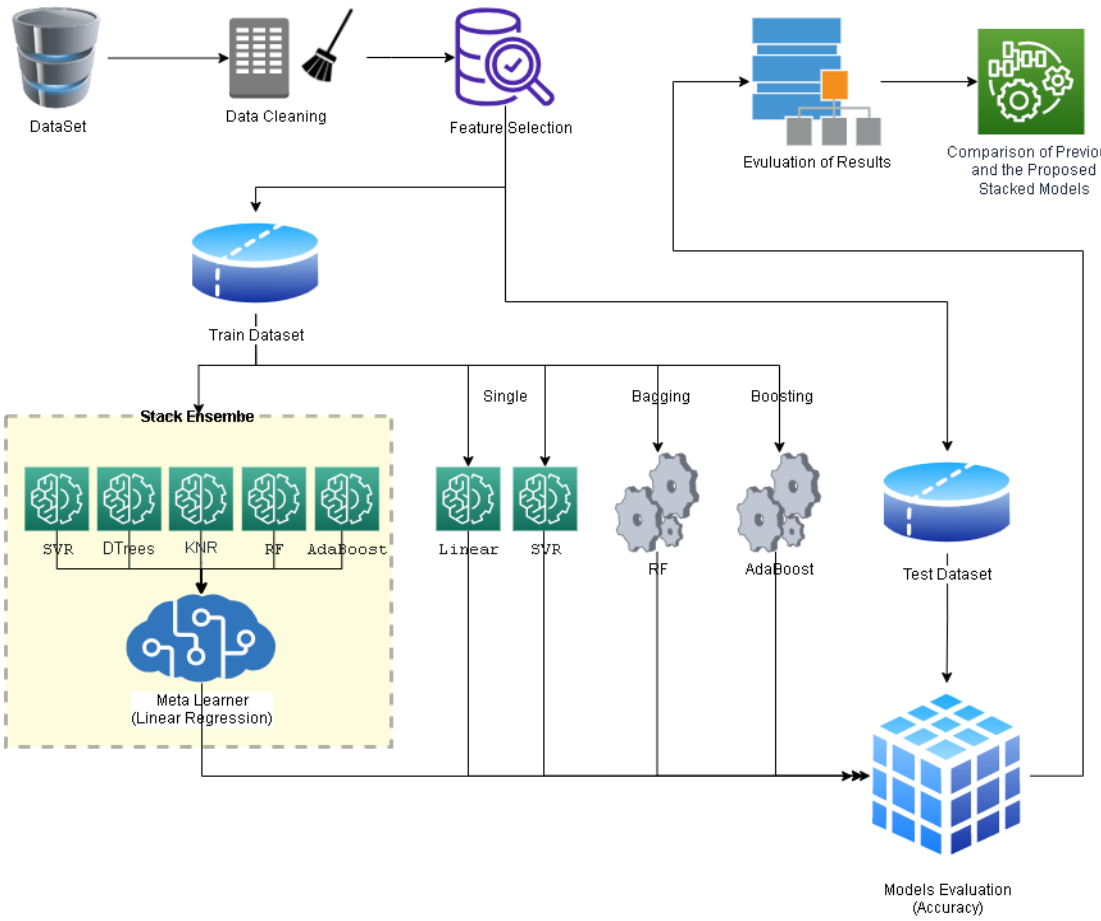
Çizelge 2.1 Bagging ve boosting model farkları

BAGGING	BOOSTING
Aynı türe ait tahminleri birleştirmenin en basit yolu	Farklı türlere ait tahminleri birleştirmenin bir yolu
Sapmayı değil varyansı azaltmayı hedefleyin.	Sapmayı azaltmayı hedefleyin, varyansı değil.
Her model eşit ağırlık alır.	Modeller performanslarına göre ağırlıklandırılır.
Her model bağımsız olarak inşa edilmiştir.	Yeni modeller, daha önce oluşturulmuş modellerin performansından etkilenir
Farklı eğitim verisi alt kümeleri, tüm eğitim veri kümesinden değiştirmeli satır örneklemesi ve rastgele örnekleme yöntemleri kullanılarak seçilir.	Her yeni alt küme, önceki modeller tarafından yanlış sınıflandırılan öğeleri içerir.
Torbalama, aşırı uyum sorununu çözmeye çalışır.	Boosting, önyargıyı azaltmaya çalışır.
Sınıflandırıcı kararsızsa (yüksek varyans), torbalama uygulayın.	Sınıflandırıcı kararlı ve basitse (yüksek yanlılık), güçlendirmeyi uygulayın.
Bu temelde sınıflandırıcılar paralel olarak eğitilir.	Bu temelde sınıflandırıcılar sıralı olarak eğitilir.
Örnek: Rastgele orman modeli, Bagging'i kullanır.	Örnek: AdaBoost, Yükseltme tekniklerini kullanır

- Stacking Yöntemi: Başka bir yaklaşım ise farklı öğrenme algoritmalarını tek bir veri kümesine uygulamaktır. Daha sonra farklı sınıflandırıcıların tahminleri birleştirilerek ve bir meta-seviye-sınıflandırıcı tarafından kullanılarak nihai bir hipotez üretilir. Bu teknik "stacking" olarak adlandırılmaktadır (Sikora 2014). Stacking modeli, makine öğrenimi alanında kullanılan bir ensemble (kombinasyon) yöntemidir. Ensemble yöntemleri, birden fazla öğrenme modelini bir araya getirerek daha güçlü ve daha doğru tahminler elde etmeyi hedefler. Stacking modeli, temelde iki seviyeden oluşur: Level 0 ve Level 1. Level 0 (Base Learner): Level 0, farklı öğrenme algoritmalarını veya modellerini içerir. Bu seviyede, çeşitli öğrenme modelleri (örneğin, karar ağaçları, destek vektör makineleri, k-NN vb.) aynı veri setine uygulanır ve her bir modelin tahminleri elde edilir. Bu modeller, veri setinin farklı özelliklerini ve yapılarını yakalamada farklı yaklaşımları temsil eder. Level 1 (Meta Learner): Level 1, Level 0'dan gelen tahminlerin birleştirildiği ve birleştirilen tahminlerin kullanılarak nihai bir tahminin oluşturulduğu seviyedir. Starking modeli, birden fazla modelin birleştirilmesiyle genellikle daha yüksek performans sağlayabilen ve daha iyi genelleme yapabilen bir ensemble yöntemidir. Level 0'daki farklı modeller, çeşitlilik sağlar ve farklı açılardan veriye yaklaşarak daha kapsamlı bir bilgi yakalayabilir. Level 1'deki üst seviye model ise bu farklı tahminleri bir araya getirerek daha güçlü bir sonuç elde etmeyi amaçlar (Breiman 1996).

3. MATERYAL VE METOT

Bu çalışmada yürütülen işlem basamakları ve uygulanan teknikler Şekil 3.1'deki akış diyagramında gösterilmektedir.



Şekil 3.1 Modelin akış diyagramı

3.1 Veri Seti

Çalışmanın veri seti, Ağrı, Çankırı ve Edirne illerindeki 8.sınıf (ortaokul son sınıf) öğrencilerinin Millî Eğitim Bakanlığı Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından 2021/2022 eğitim ve öğretim yılındaki LGS sınav verileri ile aynı öğrencilerin e-okul sisteminde kayıtılı ortaokul öğrenim süreçlerindeki eğitsel bilgilerden oluşmaktadır. Benzersiz bir OPAK_ID kod aracılığı ile tekil hale getirilmiş ve içerisinde öğrenci kimliğine ya da kişisel bilgilerine dair veri başlığı bulunmayan sadece eğitsel

verilerden oluşan bir veri setidir. Veri seti yapısal olarak iki grup veri içermektedir. Birinci grup sınav performansına dayalı verilerdir. Bunlar öğrencilerin Lise Giriş Sınavındaki derslere göre doğru yanlış cevap sayıları ile öğrencinin yüzdelik dilimi ve puanlarıdır. İkinci grup ise e-okul sisteminden alınan ve aynı dersler için ortaokul 5-8.sınıf arası not ortalamasını, devamsızlık sayısı, aile öğrenim ve gelir durumu, okulun bulunduğu yerleşim yeri okul türü gibi öğrenciyi ve öğrenim hayatını tanımaya yarayacak 25 farklı değişkenden oluşmaktadır.

Veri setindeki veri başlıklarının belirlenmesinde; hem bir eğitmen hem de Milli Eğitim Bakanlığında uzun yıllardır görev yapmış olmanın verdiği tecrübenin büyük katkısı olmuştur. Seçilen öznitelikler öğrencilerin başarılarına etki edebilecek temel veri başlıklarını içermektedir. Veri seti temin edildikten sonra veri temizliği ve feature selection çalışmaları ile orijinal veri setinde bulunan iki bağımlı değişkenden Sınav Puanı değişkeni ile bağımlı değişkeni etkilemeyen bağımsız değişkenlerden Opak_ID, Il Plaka Kodu ve ders bazlı doğru yanlış cevap sayıları vb. başlıkları çıkartılmıştır. Bunun yanında az sayıdaki öğrencinin özel durumlarına bağlı nedenlerden (Din Kültürü ve Ahlak Bilgisi dersi muafiyeti vb.) kaynaklanan veri eksiklikleri bulunmaktaydı. Bu öğrencilere ait veriler veri setinden çıkartılmak suretiyle veri dengesizliğinin önüne geçilmiştir. Bu çalışmaların ardından veri seti kullanıma hazır hale getirilmiştir. 29/09/2022 tarih ve 59378648 sayılı yazı ile Millî Eğitim Bakanlığı Bilgi İşlem Genel Müdürlüğünden resmi olarak temin edilerek çalışmada kullanılan Veri Setinin içeriği ve veri tipleri aşağıdaki Çizelge 3.1'de gösterilmektedir.

Çizelge 3.1 Veri Seti öznitelik listesi veri yapısı

#	COLUMN	NON-NULL COUNT	D-TYPE
1	YERLESIM	10686	int64
2	KURUM_TUR	10686	int64
3	OKUL_ONCESI_EGITIM	10686	int64
4	TASIMALI_MI	10686	int64
5	DEVAMSIZLIK	10686	float64
6	OZEL_EGITIM	10686	int64
7	BILSEM_OGR_MI	10686	int64
8	KARDES_SAYI	10686	int64
9	ODASI_VAR_MI	10686	int64

Çizelge 3.1 Veri Seti öznitelik listesi veri yapısı (Devam)

#	COLUMN	NON-NULL COUNT	D-TYPE
10	HANE_KISI	10686	int64
11	ANNE_OGRENIM	10686	int64
12	BABA_OGRENIM_DURUM	10686	int64
13	GELIR_DUZEYI	10686	int64
14	OKU_KITAP_SAY	10686	int64
15	CINSIYETI	10686	int64
16	NAKIL_MI_GELDI	10686	int64
17	SINIF_MEVCUDU	10686	int64
18	YATILILIK_DURUM	10686	int64
19	OKUL_TURKCE	10686	float64
20	OKUL_MAT	10686	float64
21	OKUL_FEN	10686	float 64
22	OKUL_TARIH	10686	float 64
23	OKUL DIN K	10686	float 64
24	OKUL_Y_DIL	10686	float64
25	GBS_YUZDE	10686	float64

3.2 Veri Tanım ve Tipleri

Modelde kullanılan veri tanımları ve dönüşüm tipleri aşağıdaki Çizelge 3.2'de gösterilmektedir.

Çizelge 3.2 Veri başlıkları ve içerdiği değerler skalası

ÖZ NİTELIK ADI	VERİ TİPİ	VERİ TANIMILER
Aile gelir durumu	1	Çok Düşük
	2	Düşük
	3	Orta
	4	İyi
	5	Çok İyi
Anne, baba mezuniyet bilgisi	0	İlkokul
	1	Ortaokul
	2	Lise
	3	Yüksek Okul
	4	Lisans
	5	Lisansüstü
	6	Doktora
	7	3 Yıllık Eğitim Enstitüsü
	8	Okuma Yazma Biliyor
	9	Okuma Yazma Bilmiyor

Çizelge 3.2 Veri başlıkları ve içerdiği değerler skalası (Devam)

ÖZ NİTELIK ADI	VERİ TİPİ	VERİ TANIMILER
	10	İlköğretim
Cinsiyet	0	Kız
	1	Erkek
Devamsızlık verisi	0-180	
Kardeş sayısı	0-18	
Özel eğitim öğrencisi mi?	0	Hayır
	1	Evet
Okul türü	226	Ortaokul
	241	Yatılı Bölge Ortaokulu
	151	İmam Hatip Ortaokulu
	541	Özel Türk Ortaokulu
Okul yerleşim yeri	0	İl, İlçe Merkezi
	1	Köy
	2	Belde
	3	Diğer (Mezra vs.)
Öğrenci taşımalı durum	0	Taşımalı Değil
	1	Taşımalı
Ders not ortalamaları	0 - 100	
Sınıf mevcudu	0 - 100	
Taşımalı eğitim öğrencisi mi?	0	Hayır
	1	Evet
Okul öncesi eğitim almış mı?	0	Hayır
	1	Evet
Bilsem öğrencisi mi?	0	Hayır
	1	Evet
Kendi odası var mı?	0	Hayır
	1	Evet
Yatılılık durumu	0	Hayır
	1	Evet
Hanedeki kişi sayısı	0 - 50	
Okula nakil yoluyla mı geldi?	0	Hayır
	1	Evet

3.3 Veri Ön İşleme

Veri setinde bulunan 35 özellikten ikisi bağımlı değişken olup diğer 32 tanesi bağımsız değişken biri de öğrenci ID'sidir. Öğrenci ID'si ile birlikte bağımlı değişkenlerden

Yüzdelik Dilim ve Sınav Puanı parametreleri için iki ayrı model planı oluşturulmuştur. Feature Selection işlemi için öncelikle yüzdelik dilim değişkeni modelin çıktı (output) bağımlı değişkeni olarak belirlenmiş ve buna göre yapılan Feature Selection işleminde sınav verilerinden oluşan doğru ve yanlış cevap girdilerinin bağımlı değişkene doğrudan etki ettiği ve diğer özelliklerin etkisinin ölçülemesini engellediği görülmüştür. Bu nedenle veri setinden sınav girdileri de çıkartılmıştır. Bu aşamadan sonra sınavın zorluğuna göre değişim gösteren Sınav Puanı değişkeni yerine yüzdelik değişkeni tek output olarak kullanılmasına karar verilmiştir. Ayrıca il ve ilçe bazlı analiz yapılmayacağı için ve farklı uyrukta öğrenci de olmadığı için İl, İlçe ve öğrenci uyruğu değişkenleri de veri setinden çıkartılarak toplam 24 bağımsız 1 bağımlı değişken içeren 25 değişkenli veri seti elde edilmiştir. Şekil 3.2 Veri başlıkları ve aldığı değerler gösterilmektedir.

```
In [20]: data_scaled2.describe()
```

Out[20]:

	0	1	2	3	4	5	10
count	10686.000000	10686.000000	10686.000000	10686.000000	10686.000000	10686.000000	
mean	0.242280	0.025478	0.584877	0.117911	0.078488	0.016002	
std	0.422987	0.147206	0.492766	0.322518	0.097334	0.125490	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.014535	0.000000	
50%	0.000000	0.000000	1.000000	0.000000	0.049419	0.000000	
75%	0.500000	0.000000	1.000000	0.000000	0.107558	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

8 rows × 25 columns

Şekil 3.2 Veri başlıkları ve aldığı değerler

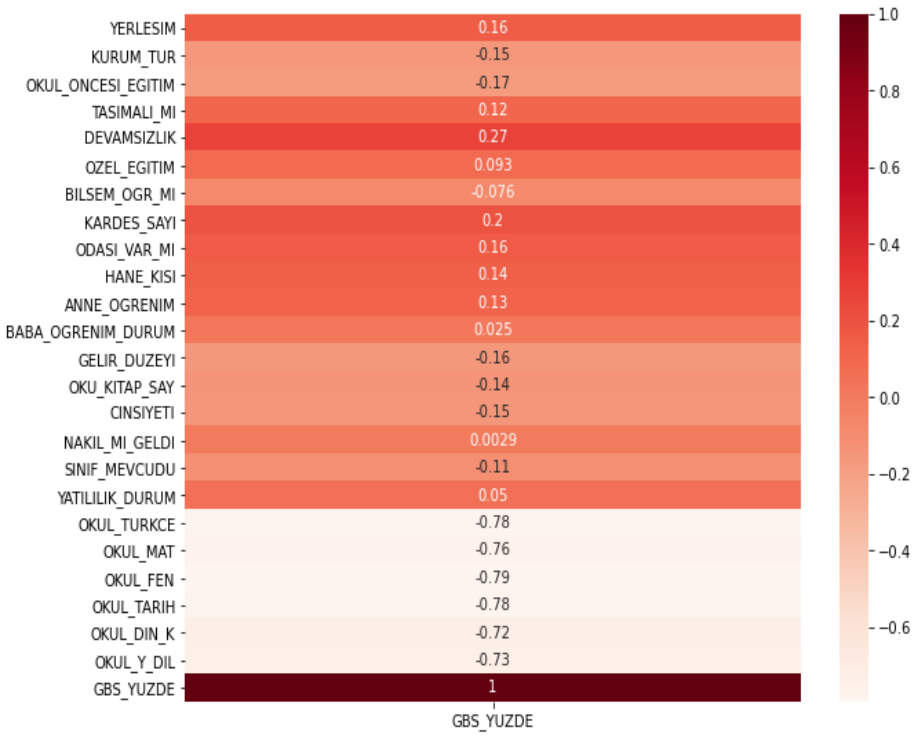
3.4 Öznitelik Seçimi

Veri kümesinin analizi sonrasında yapısı ve özellikleri incelenerek verilerin nitelikleri, yoğunluğu, boyutu ve dağılımı gibi özellikleri göz önünde öznitelik seçimi şu şekilde yapılmıştır.

- Özniteliklerin önemi belirleme: Feature Selection algoritmaları yardımıyla Tahmini en fazla etkileyen özellikler belirlenmiş etkisi olmayanlar veri setinden çıkartılmıştır.

- Özniteliklerin dönüştürülmesi: Bazı öznitelikler doğrudan kullanıma uygun olmadığı için (Char ve String ifade içeren) bu alanların numerik dönüşümleri yapılarak alınmıştır. Örneğin (Cinsiyet E:1, K:0)kadar karmaşık olabilir.
- Öznitelik seçimi yöntemlerinin kullanımı: Öznitelik seçimi yöntemleri, model performansını artırmak için önemli ön işleme aşamalarından biridir. Bu yöntemler arasında, filtre yöntemleri, sarma yöntemleri ve gömme yöntemleri gibi farklı teknikler yer alır.

Öznitelik seçimi, bazı veri setlerinde model performansını artırmanın yanı sıra, modelin daha anlaşılır ve yorumlanabilir olmasını sağlar. Şekil 3.3'de Veri Setindeki tüm parametrelerin Chi-Square Feature Selection metodu ile çıkartılmış ağırlıkları verilmektedir.



Şekil 3.3 Chi-square metodu ile yapılan öznitelik seçim sonuçları

Veri Setine farklı öznitelik seçim yöntemleri uygulanmış ve her bir Feature Selection metodundan farklı öznitelik setleri elde edilmiştir. Elde edilen bu öznitelik sonuçlarından

akademik çalışmalarda yaygın olarak kullanıldığı bilinen ANOVA ve Chi-Square öznitelik seçim yöntemlerinden elde edilen öznitelik veri seti modellere uygulanmıştır.

Bununla birlikte öznitelik seçimi yapılmamış 25 öznitelik içeren veri seti de modellere ayrıca uygulanmış ve alınan sonuçlar karşılaştırılmıştır. Çizelge 3.3'te Veri Setine uygulanan farklı öznitelik seçim yöntemleri ve seçilen Öznitelikler gösterilmektedir.

Çizelge 3.3 Öznitelik seçim yöntemlerine göre alınan sonuçlar

YÖNTEM	TEKNİK	SEÇİLEN ÖZNİTELİKLER
Filter based	Pearson korelasyonu	DEVAMSIZLIK KARDES_SAYI OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K OKUL_Y_DIL
	ANOVA ()	KURUM_TUR OKUL_ONCESI_EGITIM DEVAMSIZLIK KARDES_SAYI ODASI_VAR_MI CINSIYETI OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K OKUL_Y_DIL
	Ki-Kare (chi2)	YERLESIM KURUM_TUR DEVAMSIZLIK KARDES_SAYI ANNE_OGRENIM OKU_KITAP_SAY OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K OKUL_Y_DIL
Wapper	Forward Selection	YERLESIM DEVAMSIZLIK OZEL_EGITIM KARDES_SAYI SINIF_MEVCUDU OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K
	Backward Selection	YERLESIM DEVAMSIZLIK OZEL_EGITIM CINSIYETI SINIF_MEVCUDU OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K

Çizelge 3.3 Öznitelik seçim yöntemlerine göre alınan sonuçlar (Devam)

YÖNTEM	TEKNİK	SEÇİLEN ÖZNİTELİKLER
	Recursive Feature Elimination (RFE)	YERLESIM OKUL_ONCESI_EGITIM BILSEM_OGR_MI ODASI_VAR_MI GELIR_DUZEYI CINSIYETI NAKIL_MI_GELDI YATILILIK_DURUM OKUL_TURKCE OKUL_FEN
Emedded	Lasso	YERLESIM KURUM_TUR DEVAMSIZLIK KARDES_SAYI ODASI_VAR_MI HANE_KISI ANNE_OGRENIM GELIR_DUZEYI OKU_KITAP_SAY CINSIYETI SINIF_MEVCUDU OKUL_TURKCE OKUL_MAT OKUL_FEN OKUL_TARIH OKUL_DIN_K OKUL_Y_DIL

3.5 Veri Temizleme ve Veri Doğrulama

Bilgisayar biliminde veri doğrulama, kullanılan verinin istenilen veri kalitesine sahip olduğundan yani hem doğru hem de kullanışlı olduğundan emin olmak için verilerin veri temizleme işleminden geçmesini sağlama sürecidir. Girilen verinin aykırı değer (Outlier) içerip içermediğini tespit etmek varsa bunları optimize etmek gerekmektedir. Bu işlem için Standardizasyon işlemi yapılabilir ya da verinin Z-skoru hesaplanarak buna göre veri manuel olarak düzenlenebilir. Veri setindeki verilerin araştırma modeline uygunluğu kontrol edilerek eksik verisi olan öğrencilerin verileri çıkartılmıştır. Ayrıca Outlier (Aykırı) veriler olup olmadığı da kontrol edilmiş ve az sayıdaki (Muaf olunan derslerden kaynaklı) NULL veri içeren satırlar çıkarılmıştır. Aykırı değere rastlanmamıştır. Veri eksikliği araştırmanın sonuçlarını etkileyebileceğinden bu tür kontroller önemlidir. Bu sayede araştırmanın doğruluğu ve güvenilirliği arttırılmış ve hipotezlerin geçerliliği sağlanmıştır.

Öte yandan Python'da veriyi Standartlaştırmak için gerekli kütüphaneler kullanılarak modelde kullanmadan önce doğruluğu da sağlanmış güvenilir veri haline dönüştürmek mümkündür. Bunun için en yaygın kullanılan metotlar; Standardizasyon için StandartScaller, Normalizasyon için ise Normalizasyon ve MinMax Scaller metotlarıdır. Veri seti e-okul veri tabanından alındığı için veri yapısı gereği pek çok standarta sahip uygunluğu yüksek bir yapıdaydı. Eksik veriler de elimine edildikten sonra StandartScaler, Normalizasyon, MinMax Scaller ve Robust Scaller işlemleri uygulandıktan sonra oluşturulan modellerde kullanılmıştır. Veri seti uygunluğu yüksek olduğundan bu işlemlerin bu çalışmanın performans metriklerinde iyileşmeye etkisi olmamıştır. Şekil 3.4'te Veri setinin özeti gösterilmektedir.

```
In [19]: from sklearn.preprocessing import MinMaxScaler
         min_max_scaler = MinMaxScaler()
         x_scaled = min_max_scaler.fit_transform(data)
         data_scaled2 = pd.DataFrame(x_scaled)
         data_scaled2.head()
```

Out[19]:

	0	1	2	3	4	5	6	7	8	9	...	15	16	17	18	
0	0.0	0.002959	0.0	0.0	0.000000	0.0	0.0	0.0625	1.0	0.1	...	1.0	0.487179	0.0	0.5939	0.7
1	0.0	0.000000	1.0	0.0	0.130814	0.0	0.0	0.0625	1.0	0.1	...	1.0	0.538462	0.0	0.4606	0.4
2	1.0	0.044379	0.0	0.0	0.110465	0.0	0.0	0.6250	1.0	0.6	...	0.0	0.384615	1.0	0.9094	0.6
3	0.0	0.000000	0.0	0.0	0.005814	0.0	0.0	0.0000	0.5	0.3	...	1.0	0.923077	0.0	0.7250	0.4
4	0.0	0.044379	1.0	0.0	0.037791	0.0	0.0	0.1250	1.0	0.2	...	1.0	0.230769	1.0	0.6847	0.5

5 rows × 25 columns

Şekil 3.4 Veri seti özeti

3.6 Optimum Modeli Belirleme

Bu çalışmada tahmin edilmek istenen öğrencinin sınav skoru ya da farklı bir deyişle sıralama yüzdesi olacağı için ML Algoritmaları ile regression analizi yapılması gerekmektedir. İlk etapta en temel regresyon modeli olan Linear Regression algoritması daha sonra diğer tek tabanlı algoritmalar ve sırasıyla Bagging, Boosting ve Stacking topluluk öğrenme yöntemleri ile çok tabanlı algoritmalar kullanılmak suretiyle farklı tahmin modelleri ile analizler yapılmıştır. Çizelge 3.4'te kullanılan mimari ve ML algoritmaları ile en iyi skoru üreten Hyper-parametreler listelenmektedir.

Çizelge 3.4 Algoritmalara göre en iyi sonuç alınan hyper-parametre listesi

MİMARİ	TEKNİK	ALGORİTMA	HYPER-PARAMETRE
Tek Tabanlı	Single	ANN	{Dense= [100,50,10,1] Activation= ['relu', 'linear'] DropOut= 0.2 Optimizer=['rmsprop', loss='mse'] EarlyStopping: [monitor='val_loss', patience=50], epochs= 1000, batch_size= 512, verbose=1 }
		SVR	{kernel='linear'}
		K-Neighbors	{n_neighbors = 10}
		Linear Regression	{}
		ElasticNet	{alpha = 1}
Topluluk (Ensemble) Öğrenme Algoritmaları	Bagging	Linear Regression	{}
		SVR	{kernel='linear'}
		K-Neighbors	{n_neighbors=100}
		Random Forest	{n_estimators=100}
		ANN	{Dense= [100,50,10,1] activation= ['sigmoid', 'tanh', 'relu', 'linear'] optimizer='rmsprop', loss='mse', epochs=100, batch_size=512, EarlyStopping: [monitor='loss', patience=50]}
		Decision Tree	{max_depth=5}
		Extra Trees	n_estimators=100
	Boosting	AdaBoost	{'learning_rate': 0.1, 'n_estimators': 50}
		X-GBoost	{'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 200}
		Gradient Boosting	{'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 200}
	Stacking	Linear Regression	{}
		SVR	{kernel='linear'}
		K-Neighbors	{n_neighbors=100}
		Random Forest	{n_estimators=100}
		Decision Tree	{max_depth=5}
		Extra Trees	{n_estimators=100}
		Gradient Boosting	{n_estimators=100}
		X-GBoost	{n_estimators=100}
		AdaBoost	{n_estimators=100}

Bu çalışmada oluşturulan algoritmaların performanslarını değerlendirme aşamasında aşağıda belirtilen değerlendirme metrikleri kullanılmıştır.

3.6.1 Tek tabanlı (single based) algoritma metrikleri

Tek tabanlı ML regresyon algoritmalarında, modelin doğruluğunu ölçmek için kullanılan bazı yaygın değerlendirme metrikleri şunlardır:

- Mean Squared Error (MSE): Ortalama kare hatası olarak bilinen metrik bir tahmin modelinin gerçek değerlerle ne kadar uyumlu olduğunu ölçmek için kullanılan istatistiksel bir ölçüdür. MSE regresyon problemlerinde sıkça kullanılan bir hata metriğidir. MSE hesaplaması tahmin edilen değerler ile gerçek değerler arasındaki farkların karelerinin ortalamasını temsil eder. Bu farkların karelerinin alınması, negatif ve pozitif hataları dengeler ve büyük hataları vurgular. MSE değeri ne kadar küçük olursa, tahmin modelinin gerçek değerlere o kadar yakın olduğu anlaşılır (Nguyen 2020). Denklem (3.1)'de MSE hesaplama formülü görülmektedir.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - \hat{y}_i)^2 \quad (3.1)$$

- Root Mean Squared Error (RMSE): Ortalama kök kare hatası metriği, makine öğrenimi ve istatistiksel analizde kullanılan bir hata ölçüsüdür. RMSE tahmin edilen değerlerle gerçek değerler arasındaki farkları ölçerek tahmin modelinin doğruluğunu değerlendirmek için kullanılır. RMSE öncelikle hata karelerinin ortalama değerinin alınmasıyla hesaplanır. Daha sonra bu değerin karekökü alınarak RMSE elde edilir. Bu ölçüt, tahmin edilen değerlerin gerçek değerlerden ne kadar uzak olduğunu hesaplar ve bu hataları kareler ve ortalamalar üzerinden değerlendirir. RMSE genellikle regresyon modellerinin performansını değerlendirmede kullanılır. Daha küçük bir RMSE değeri, tahmin modelinin gerçek değerlere daha yakın olduğunu gösterirken, daha büyük bir RMSE değeri ise daha fazla hata olduğunu işaret eder (Moody 2019). Denklem (3.2)'de RMSE hesaplama formülü görülmektedir.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_j)^2}{n}} \quad (3.2)$$

- Mean Absolute Error (MAE): Bir regresyon modelinin tahminlerinin gerçek değerlerden ne kadar uzak olduğunu ölçen bir performans metriğidir. MAE tahminlerin gerçek değerlerden mutlak farklarının ortalamasını hesaplar. Akademik bir dilde kısa bir tanım olarak Mean Absolute Error (MAE), bir regresyon modelinin tahminlerinin gerçek değerlerden olan mutlak farklarının ortalamasını hesaplayarak modelin tahmin doğruluğunu ölçen bir performans metriğidir (Chugh 2020). Denklem (3.3)'de MAE hesaplama formülü görülmektedir.

$$MAE = \frac{\sum_{t=1}^{n} |u_t|}{n} \quad (3.3)$$

- R-squared (R-kare): Bir modelin verileri ne kadar iyi açıkladığını ölçer. R-kare tahmin edilen değerlerin toplam varyansının gerçek değerlerin toplam varyansına oranını ifade eder (Chugh 2020). Denklem (3.4)'te R-Squared hesaplama formülü görülmektedir.

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}} \quad (3.4)$$

- Adjusted R-squared (Düzeltilmiş R-kare): R-kare'nin düzeltilmiş bir versiyonudur ve modeldeki değişkenlerin sayısını da hesaba katarak daha adil bir karşılaştırma yapılmasına izin verir (Chugh 2020). Denklem (3.5)'te düzetltilmiş R-Squared hesaplama formülü görülmektedir.

$$R^2_{adj} = 1 - \frac{MS_{Residual}}{MS_{Total}} \quad (3.5)$$

- Mean Absolute Percentage Error (MAPE): Gerçek değerlere göre yüzde olarak tahmin edilen değerlerin mutlak farklarının ortalamasını ölçer (Kumar 2020). Denklem (3.6)'da MAPE'nin hesaplama formülü görülmektedir.

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|u_t|}{Y_t}}{n} * 100 \quad (3.6)$$

Bu metrikler, tek tabanlı ML regresyon algoritmaları ile elde edilen sonuçların doğruluğunu değerlendirmek için kullanılabilir. Hangi metriğin kullanılacağı, problemin doğasına ve uygulanacak olan algoritmanın özelliklerine bağlıdır.

3.6.2 Topluluk öğrenme algoritma metrikleri

Topluluk öğrenme (ensemble learning) ML regresyon algoritmaları, birden fazla regresyon modelinin birleştirilmesiyle oluşturulur ve daha doğru tahminler yapmayı amaçlar. Farklı topluluk öğrenme algoritma yöntemleri olan Bagging, Boosting ve Stacking yöntemler ile uygulanan regresyon modelleri ile elde edilen sonuçların doğruluğunu ölçmek için kullanılan bazı yaygın regresyon metrikleri, tek tabanlı regresyon algoritmaları için kullanılan metriklerle aynıdır. Ancak, topluluk öğrenme modelleri birden fazla modelin birleştirilmesinden oluştuğu için, bu modellerin birleştirilmesinin başarısını ölçmek için ek metrikler de kullanılabilir. Bu metriklerden bazıları şunlardır:

- Bagging Mean Squared Error (Bagging MSE): Bagging Regressor gibi, Bagging yöntemi kullanarak oluşturulan modellerin ortalama karesel hatasını ölçer.
- Bagging Root Mean Squared Error (Bagging RMSE): Bagging MSE'nin kareköküdür. Bu metrik, Bagging Regressor için kullanılan RMSE metriğiyle benzerdir.
- Bagging Mean Absolute Error (Bagging MAE): Bagging yöntemi kullanarak oluşturulan modellerin ortalama mutlak hatasını ölçer.

- Bagging R-squared (Bagging R-kare): Bagging yöntemi kullanarak oluşturulan modellerin ortalama R-kare değerini ölçer.
- Out-of-Bag R-squared (OBB R-kare): Bagging yöntemi kullanarak oluşturulan modellerin OOB verileri üzerindeki R-kare değerini ölçer. OOB verileri, modelin oluşturulurken kullanılmayan verilerdir.

4. BULGULAR VE TARTIŞMA

Bu bölümde öznitelik seçim sonuçları ve oluşturulan tek tabanlı ve topluluk makine öğrenmesi algoritmalarının analizleri gösterilmektedir. Çalışmada uygulanan öznitelik seçim teknikleri uygulanmış model ile öznitelik seçim uygulanmadan veri setinin tüm öznitelikleri kullanılarak elde edilen sonuçlar ve modeller arası performans metrikleri detaylı şekilde ele alınmıştır.

4.1 Öznitelik Seçim Sonuçları

Elde edilen Öznitelik seçim sonuçlarından yaygın olarak kullanılan Chi-Square ve ANOVA teknikleri ile belirlenen 12 alan Çizelge 5.1'de gösterilmektedir. Tüm yöntemlerde bağımlı değişkene en fazla etkiyi yapan parametrelerin ders başarı notları olduğu görülmüştür. Kullanılan tekniğe göre değişim gösterenler ise öğrencilerin sosyo-ekonomik ve çevresel faktörler olduğu görülmektedir. Çizelge 5.1'de de görüleceği gibi yaygın olarak kullanılan bu iki öznitelik seçim yönteminden 12 parametreden 9 tanesi aynı çıkmış sadece üç özellik değişim göstermiştir. Chi-Square ve ANOVA teknikleri ile belirlenen Öznitelikleri içeren Veri Seti ML algoritmalarına uygulandığında alınan skorlar birbirine yakın çıkmaktadır. Bu nedenle tümünde daha iyi skorlar elde edilen Chi-Square Feature Selection metodu ile elde edilen veri seti tercih edilmiştir. Öte yandan bulunduğu okula sonradan gelip gelmediği, BİLSEM (üstün yetenekli) öğrenci olma durumu, Taşımalı Eğitim kapsamında olup olmaması ve Baba öğrenim durumu bağımlı değişkene etkilerinin oldukça düşük olduğu görülmüştür. Çizelge 4.1'de ANOVA ve Chi-Square teknikleri ile belirlenen öznitelikler gösterilmektedir.

Çizelge 4.1 ANOVA ve chi-square teknikleri ile belirlenen öznitelik tablosu

CHI-SQUARE	ANOVA
YERLESIM	OKUL_ONCESI_EGITIM
KURUM_TUR	KURUM_TUR
DEVAMSIZLIK	DEVAMSIZLIK
KARDES_SAYI	KARDES_SAYI
ANNE_OGRENIM	ODASI_VAR_MI
OKU_KITAP_SAY	CINSIYETI
OKUL_FEN	OKUL_FEN
OKUL_TURKCE	OKUL_TURKCE
OKUL_TARIH	OKUL_TARIH
OKUL_MAT	OKUL_MAT
OKUL_Y_DIL	OKUL_Y_DIL
OKUL_DIN_K	OKUL_DIN_K

4.2 Tek Tabanlı Makine Öğrenme Modeli Sonuçları

Tek tabanlı algoritmalardan LinearRegression, ANN, SVR, K-Neighbors ve ElasticNET algoritmalarının kullanıldığı modelde her bir algoritma için ayrı ayrı tahmin verileri üretilmiş ve metrikler kayıt altına alınmıştır. Algoritmaların default parametrelerinin yanında farklı hyper parametreler ile de testler tekrarlanarak en iyi skor metrikleri Çizelge 4.6'da gösterilmişti. Ayrıca çalışma Feature Selection yapılmamış Tek tabanlı mimari tercih edilecek ise bu veri seti 24 öznitelik içeren veri seti ile de tekrarlandığında daha iyi skorlar elde edildiği görülmüştür. Tek Tabanlı Algoritmalar ile alınan skorlar Çizelge 4.2 ve 4.3'te ayrı ayrı gösterilmektedir.

Çizelge 4.2 Tek tabanlı makine öğrenmesi algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
ANN	0.770	176.17	13.27	10.35	20.65
SVR	0.724	211.81	14.55	11.33	38.29
K-neighbors	0.748	193.54	13.91	10.88	33.68
LR	0.729	207.75	14.41	11.54	50.54
ElasticNet	0.707	224.78	14.99	12.26	64.31

Çizelge 4.3 Tek tabanlı makine öğrenmesi algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
ANN	0.785	173.63	13.17	10.01	20.85
SVR	0.767	202.71	14.23	10.33	34.10
K-neighbors	0.762	186.69	13.66	10.79	44.02
LR	0.740	199.07	14.10	11.22	47.42
ElasticNet	0.717	217.05	14.73	12.03	62.21

Genel olarak bakıldığında en iyi performansı ANN (R2: 0.785) algoritması feature selection yöntemi uygulanmadan sağlamıştır. Diğer taraftan en düşük performansı ElasticNet algoritması sağlamıştır.

4.3 Topluluk Tabanlı Makine Öğrenme Modeli Sonuçları

4.3.1 Bagging sonuçları

Oluşturulan Bagging Modelde BaggingRegressor algoritması kullanılmış ve testlerde base estimator olarak kullanılan algoritmalar şunlarıdr; Extra Trees Regressor, Random Forest Regressor, Sequential Neural Network, Decision Tree Regressor, Linear Regression, Support Vector_Regressor algoritmaları ile Model derlenmiştir. Bu model de farklı hyper – parmetreler ve Feature Selection yapılmamış Veri Seti ile tekrarlar yapılarak Çizelge 4.4 ve 4.5'teki sonuçlar elde edilmiştir.

Çizelge 4.4 Bagging algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış)

ML ALGORITMASI	R2	MSE	RMSE	MAE
Extra trees	0.769	176.84	13.29	10.49
RFR	0.766	179.24	13.38	10.55
ANN	0.721	213.88	14.62	11.44
K-neighbors	0.753	189.63	13.77	10.82
DTR	0.744	195.94	13.99	11.07
LR	0.729	207.93	14.42	11.55
SVR	0.724	211.78	14.55	11.34

Çizelge 4.5 Bagging algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış)

MODEL (BAGGING)	R2	MSE	RMSE	MAE
Extra trees	0.781	168.23	12.97	10.26
Random forest	0.776	171.51	13.09	10.32
ANN	0.764	181.05	13.45	10.66
K-neighbors	0.760	184.35	13.57	10.64
Decision tree	0.746	194.68	13.95	11.04
Linear regression	0.741	198.58	14.09	11.18
SVR	0.738	201.19	14.14	10.98

Tek tabanlı makine öğrenmesi algoritmaları ile alınan sonuçlara genel olarak bakıldığında en iyi performansı ANN (R2: 0.785) algoritması feature selection yöntemi uygulanmadan sağlamıştır. Diğer taraftan en düşük performansı ElasticNet algoritması sağlamıştır.

Topluluk tabanlı makine öğrenme algoritmalarının Bagging tekniği ile tasarlanmış modelini ile elde edilen sonuçlara bakıldığında genel olarak en iyi performans gösteren algoritma Extra Trees (R2: 0.781) algoritması olmuştur. En düşük performansı ise Feature selection yapılmış veri seti ile ANN feature selection yapılmadan uygulanan veri seti ile SVR algoritmaları sağlamıştır.

4.3.2 Boosting sonuçları

Bu teknik ile oluşturulan üç farklı modelde; AdaBoostRegressor, XGBoostRegressor ve GradiantBossting Regresyon algoritmaları kullanılarak model tasarlanmış ve farklı parametreler denemiştir. Boosting model olrak tercih edien 3 farklı algoritma ile AdaBoostRegressor, XGBoostRegressor ve GradiantBossting modeller içerisinden en iyi skor R-Squared skoru GradiantBossting ile alınmıştır. Bossting algoritmalar ile elde edilen metrikler Çizelge 4.6 ve 4.7'deki sonuçlar elde edilmiştir.

Çizelge 4.6 Boosting algoritmalarının performans sonuçları (Chi-Square öznitelik seçimi uygulanmış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
Gradient boost	0.77	176.83	13.29	10.42	32.74
XGBoost	0.77	179.64	13.29	10.42	33.12
AdaBoost	0.72	209.35	14.46	11.66	50.27

Çizelge 4.7 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
Gradient boosting	0.782	167.57	12.94	10.13	30.14
XGBoost regressor	0.780	169.28	13.01	10.21	29.83
AdaBoost regressor	0.728	208.68	14.44	11.79	52.74

Boosting tipi topluluk makine öğrenmesi algoritmaları ile elde edilen sonuçlarda genel olarak en iyi performansı gösteren algoritma Gradient Boosting (R2: 0.782) algoritması olmuştur. Feature selection yapılmamış veri seti ile elde edilen bu skor yaklaşık olarak Bagging model ile alınan skora yakındır. Boosting algoritmalardan en düşük skor ise AdaBoost algoritma ile alınmıştır.

4.3.3 Stacking sonuçları

Stacking tekniği ile oluşturulan bir Ensemble Super Learner ve Stacking Regressor modellerinde tek tabanlı ve topluluk algoritmaları birlikte kullanılmıştır. Modellerde meta algoritma olarak Linear Regression, Base model algoritmaları olarak; Support Vector Regressor, K-Neighbors Regressor, Random Forest Regressor, Decision Tree Regressor, Extra Trees Regressor, GradientBoosting Regressor, XGBoost Regressor ve AdaBoost Regressor algoritmaları kullanılırmıştır. Stacking Modelde de Feature Selection yapılmış ve yapılmamış Veri Setleri ayrı ayrı kontrol edilmiştir. Oluşturulan Stacking ve SuperLearner modellerin R-Squared ve diğer performans metrikleri Çizelge 4.8 ve 4.9'de verilmektedir.

Çizelge 4.8 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
Super learner	0.76	177.05	13.30	10.33	33.40
Stacking	0.77	176.11	10.31	13.27	32.45

Çizelge 4.9 Boosting algoritmalarının performans sonuçları (Chi-square öznitelik seçimi uygulanmamış)

ML ALGORITMASI	R2	MSE	RMSE	MAE	MAPE
Stacking regressor	0.785	164.84	9.98	12.84	33.56
Super learner	0.790	161.32	12.70	10.06	34.91

Topluluk tabanlı makine öğrenme algoritmalardan Stacking tekniği ile tasarlanmış iki modelden en iyi performans, Super Learner (R2: 0.79) model ile alınmıştır. Bu aynı zamanda tek tabanlı ve topluluk algoritmaları içinde elde edilen en iyi skordur.

Mevcut veri setine uygulanan tek tabanlı ve topluluk algoritmaları ile elde edilen metrikler incelendiğinde, tek tabanlı modellerde ANN (R2: 0.785), Topluluk Tabanlı Öğrenme Algoritmalarında ise Stacking modelde Super Learner (R2: 0.79) tekniği en iyi tahmin skorlarını üretmiştir. Genel olarak bakıldığında büyük bir fark olmadığı görülmektedir. İleride bu çalışma geliştirilmek istenirse Neural Network veya Ensmble Super Leraner yöntemleri ile tahmin performansları geliştirilebilir.

5. SONUÇLAR VE TARTIŞMA

Dünyada son zamanlarda öğrenci başarısı tahmin etmede makine öğrenmesi modelleri sıkça tercih edilmeye başlanmış durumdadır. Bu konuda hem akademik hem de ticari pek çok çalışma yapılmaktadır. Bu çalışmaların odağında başarıya etki eden olumlu ve olumsuz tüm faktörlerin önceden tespiti edilmesi ve bu sayede erken uyarı ve müdahale sistemlerinin geliştirilmesi amaçlanmaktadır. Böylece başarının kalıcı biçimde arttırılması hedeflenmektedir. Bunu da makine öğrenmesi algoritmaları sayesinde kişiselleştirilmiş yaklaşımlar ve çözümler içeren öğrenci merkezli yaklaşım anlayışı ile gerçekleşmesi beklenmektedir.

Ülkemizde benzer çalışmalar üniversite düzeyinde ve bazı özel okullarda çeşitli kademelerde yürütülmekte olup daha çok öğrencilerin üniversite sınavlarındaki başarısının modellenmesi üzerine örnekler bulunmaktadır.

Lise Giriş Sınavları (LGS) üzerine gerçek öğrenci verileri ile yapılmış birkaç çalışmadan biri olma özelliği ile bu alanda referans niteliği taşımaktadır Sayısal olarak seçilen Feature sayısı arttırılarak daha iyi sonuçlar almanın mümkün olduğu görülmüştür. Sağlık, finans ya da bahis gibi alanlarda Feature Selection ile veri setinin hacmi küçültülüp verim artışı sağlanabilirken eğitsel Veri Madenciliğinde ise öğrenci hakkında daha fazla bilgiye sahip olmanın tahmin sonuçlarını iyileştirdiği görülmüştür. Seçilen öz nitelikler daha da çeşitlendirilerek daha yüksek başarımlar elde edilebilir. Veri setinde öğrencilerin sadece ortaokuldaki sınavda çıkan derslere ait notlar kullanılmış olup okul öncesi dönemden itibaren tüm ders notları ile öğrencilerin ilgi, beceri ve yeteneklerinin yanında fiziksel ve psikolojik gelişimleri, sağlık verileri, ailevi, çevresel, sosyal etkenleri de kullanılarak çok daha yüksek tahmin skorları elde etmek mümkün olabilir. Bu sonuçları sadece öğrenim hayatında sınav tahminleri yapmaktan öte meslek yönelimlerini tespit etmekte de kullanmak mümkündür. Ayrıca aile iş birliği ile öğrencilerin karşılaşılabileceği olumsuz eğilimlerin de önceden tespit edilerek koruyucu önleyici bir eğitim modeli tasa başarının artırılması mümkün hale gelebilecektir.

Öte yandan Eğitsel Veri Madenciliği alanında kullanılan makine öğrenmesi teknikleri bakımından bu çalışmadaki veri seti ve uygulanan modeller göz önüne alındığında; genel olarak ANN R2: 0,785 ve Stacking Model olarak Super Learner mimari R2: 0,79 metrik değerleri ile diğer model ve algoritmalar içinde en iyi performansı göstermişlerdir. Bu çalışmada regresyon tekniği uygulanmış olup sınıflandırma yapılarak daha yüksek accuraecy performansı elde edilebilirdi. Çünkü regresyon yapılmak suretiyle öğrencilerinin doğrudan alabileceği yüzdelik puan tahmin edilmesi amaçlanmıştır. Sınıflandırma ile tahmin yapılmış olsaydı tahminler belirli yüzdelik aralık dilimlerine karşılık gelen sınıf için yapılacağından (% 0-5 arası, %5-10 arası, %10-25 arası gibi.) modelin başarı skorlarının çok daha yüksek çıkması öngörülebilir. Bu alanda daha fazla sayıda öznitelik kullanılarak sınıflandırma algoritmaları ile yeni çalışmalar yapılması düşünülebilir. Bu şekilde yapılacak bir çalışma karar vericiler için odak bir skor yerine belirli bir skala içindeki başarıyı analiz etmeleri süreç yönetimi bakımından daha etkili olabilir.

KAYNAKLAR

Abbasoğlu, B. 2020. Ortaokul öğrencilerinin akademik başarılarının eğitsel veri madenciliği yöntemleri ile tahmini, Asos, 3:1.

Aghalarova, S. and Keser, S. B. 2022. Application of auto ML technique for predicting academic performance of students. El-Cezeri Journal of Science and Engineering, 9(2): 394–412.

Albreiki, B., Zaki, N. and Alashwal, H. 2021. A systematic literature review of student' performance prediction using machine learning techniques. Education Sciences, 11(9).

Alshabandar, R., Hussain, A., Keight, R. and Khan, W. 2020. Students performance prediction in online courses using machine learning algorithms.

Arslan, K. 2020. Eğitimde yapay zeka ve uygulamaları artificial intelligence and applicatıons in education, In: Derleme Makale Western Anatolia Journal of Educational Sciences 11:1.

Aydoğan, M. ve Karci, A. 2019. Kelime Temsil Yöntemleri ile Kelime Benzerliklerinin İncelenmesi. Çukurova Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 34(2): 181–196.

Baker, M. J. (1998). Marketing of the future. Australasian Marketing Journal (AMJ), 6(1) : 5–11.

Balaji, P., Alelyani, S., Qahmash, A. and Mohana, M. 2021. Contributions of machine learning models towards student academic performance prediction: A systematic review. Applied Sciences Switzerland, 11(21).

Başer, S., Hökelekli, O. ve Adem, K. 2020. Ortaöğretimde öğrenim gören öğrenci performanslarının veri madenciliği yöntemleri ile tahmin edilmesi. Bilgisayar Bilimleri ve Teknolojileri Dergisi, 1(1): 22-27

Baykal, A. 2014. Eğitim Sisteminde Kademeler Arası Geçi ve Sınavlar, Ege Üniversitesi, 19 sayfa, İzmir.

Breiman, L. 1996. Stacked regressions. Machine Learning, 24: 49-64.

Brereton, R. G. and Lloyd, G. R. 2010. Support vector machines for classification and regression. In Analyst Royal Society of Chemistry, 135(2).

Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H. and Ghani, N. A. M. 2021. Multiclass prediction model for student grade prediction using machine learning. IEEE Access, 9: 95608–95621.

Can, Ş. and Gerşil, M. 2017. A literature review on the use of genetic algorithms in data mining, International Journal of Computer Engineering and Technology (IJCET), 8(6): 67–76.

CHAUHAN, 2021. Web site: http://meb.ai/UT506of. Date of access: 21.01.2021.

CHUGH, 2020. Web site: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e. Date of access: 08.12.2020.

Çağıltay, K. 2010. Eğitim ve teknoloji: Bardak dolu mu boş mu ya da nasıl doldururuz. ÖDTÜ'lüler Bülteni.

Çetın, A. and Ünsal, S. 2019. Social, psychological effects of central examinations on teachers and their reflections on teachers' curriculum implementations. Hacettepe Egitim Dergisi, 34(2): 304–323.

Çetin, D. ve Bülbül, T. 2017. Okul yöneticilerinin teknostres algıları ile bireysel yenilikçilik özellikleri arasındaki ilişkinin incelenmesi. In Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17(3).

Dhilipan, J., Vijayalakshmi, N., Suriya, S. and Christopher, A. 2021. Prediction of students performance using machine learning. IOP Conference Series: Materials Science and Engineering, 1055(1): 012122.

Düzgün, O. 2017. Veri madenciliği yöntemleri ve işkur için uygulamaya yönelik model önerisi, Uzmanlık Tezi, 105 sayfa, Ankara.

Emmert, S. F. and Dehmer, M. 2019. High-dimensional lasso-based computational regression models: regularization, shrinkage, and selection. In Machine Learning and Knowledge Extraction 1(1): 359–383.

Fisher, R. A. 1936. Statistical methods for research workers, Hafner Publishing Company, 370 sayfa, America.

Ghorbani, R. and Ghousi, R. 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access, 8: 67899–67911.

Gök, M. 2017. Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi, Fen Bilimleri Dergisi 5(3).

Gürsakal, N. 2017. Makine öğrenmesi ve derin öğrenme. Dora Basım Yayın Dağıtım, Bursa, Sayfa 318, Ankara

Howell, D. C. 2014. Chi-square test: Analysis of contingency tables.

Irsoy, O., Yildiz, O. T. and Alpaydin, E. 2012. Design and analysis of classifier learning experiments in bioinformatics: Survey and case studies. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(6): 1663–1675.

Jolliffe, I. T. 2002. Principal component analysis for special types of data, pp. 338-372. Springer, New York.

Jović, A., Brkić, K. and Bogunović, N. 2015. A review of feature selection methods with applications. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings, 1200–1205.

Kabathova, J. and Drlik, M. 2021. Towards predicting student's dropout in university courses using different machine learning techniques. Applied Sciences Switzerland, 11(7).

Kalipe, G., Gautham, V. and Behera, R. K. 2018. Predicting malarial outbreak using machine learning and deep learning approach: A review and analysis. Proceedings - 2018 International Conference on Information Technology, ICIT 2018: 33–38.

Kamber, E. ve Sönmez, T. ve Bolatan, G. İ. 2019. Endüstri 4.0 Türkiye farkındalığı. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi.

Kaushik, B., Sharma, R., Dhama, K., Chadha, A. and Sharma, S. 2023. Performance evaluation of learning models for intrusion detection system using feature selection. Journal of Computer Virology and Hacking Techniques.

Kayapınar, E. 2006. Ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavı (OKS)'na hazırlanan ilköğretim 8. sınıf öğrencilerinin kaygı düzeylerinin incelenmesi (Afyonkarahisar ili örneği), Yüksek Lisans Tezi, Kocatepe Üniversitesi, 117 sayfa, Afyonkarahisar.

Kayhan, H. O. 2019. Uzaktan eğitim öğrencilerinin mezuniyet durumlarının veri madenciliği yöntemleri ile tahminiç, Yüksek Lisans Tezi, Amasya Üniversitesi, 95 sayfa, Amasya.

Keenan, A., Pituch, James P. and Stevens. 2012. Applied multivariate statistics for the social sciences, Routledge. Sayfa 814, United Kingdom.

Kılıç, D. 2014. Sınıf öğretmenlerinin 4+4+4 eğitim sistemi hakkındaki tutumlarının incelenmesi, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 18 (3): 11-21.

Koç, A. 2023. Ortaokul mezunlarının ortaöğretimi tamamlama durumlarının veri madenciliği ile tahmini: trabzon ili örneği. Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, 70 sayfa, Afyonkarahisar.

KUMAR, 2020. Web site: https://towardsdatascience.com/mad-over-mape a86a8d831447. Date of access: 29.04.2020.

Lee, S. J. and Siau, K. 2001. A review of data mining techniques. In Industrial Management and Data Systems 101(1): 41–46.

Maulud, D. and Abdulazeez, A. M. 2020. A Review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4): 140–147.

Mayr, A., Binder, H., Gefeller, O. and Schmid, M. 2014. The evolution of boosting algorithms: From machine learning to statistical modelling. Methods of Information in Medicine, 53(6): 419–427.

MOODY, 2019. Web site: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e. Date of access: 05.09.2019.

Naicker, N., Adeliyi, T. and Wing, J. 2020. Linear support vector machines for prediction of student performance in school-based education. Mathematical Problems in Engineering, 2020.

Nasteski, V. 2017. An overview of the supervised machine learning methods. Horizons, 4: 51–62.

NGUYEN 2020. Web site: https://medium.com/nothingaholic/understanding-the-mean-squared-error-df41e2c87958. Date of access: 15.06.2020.

Obenshain, M. K. 2004. Application of data mining techniques to healthcare data in infection control and hospital epidemiology, 25(8):690-5.

Özsoylu, A. F., Bilim, A. 2017. Endüstri 4.0, Çukurova Üniversitesi, sayfa 41-64, İİBF Dergisi.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K. 2021. Classification and prediction of student performance data using various machine learning algorithms. Materials Today: Proceedings.

Rastogi, S., Srivastava, G. K. and Vishwakarma, S. K. 2023. A review on predicting depression, 020141.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A. and Durán-Domínguez, A. 2020. Analyzing and predicting students' performance by means of machine learning: A review. In Applied Sciences Switzerland 10(3).

Ray, S. 2019. Big data, cloud and parallel computing : trends, prespectives and prospects: Comitcon-2019, In: 14th-16th Proceedings of the International Conference on Machine Learning.

Sikora, R. and Al-laymoun, O. H. 2014. A modified stacking ensemble machine learning algorithm using genetic algorithms. Journal of International Technology and Information Management, 23(1).

Singh, R. S. P. 2020. Machine learning algorithms and ensemble technique to improve prediction of students performance. International Journal of Advanced Trends in Computer Science and Engineering, 9(3): 3970–3976.

Sutha, K. and Tamilselvi, J. J. 2015. A review of feature selection algorithms for data mining techniques. Computer Science.

Şengür, D., Karabatak, S. 2018. Veri madenciliği teknikleri ile öğrencilerin başarım analizi, Türk Bilim ve Teknoloji Dergisi, 13(2): 53-59.

Tosunoğlu, E. , Yılmaz, R. , Özeren, E. ve Sağlam, Z. 2021. Eğitimde makine öğrenmesi: araştırmalardaki güncel eğilimler üzerine inceleme . Ahmet Keleşoğlu Eğitim Fakültesi Dergisi, 3(2): 178-199.

Uçar, E. 2013. Ortaöğretime geçiş sistemi (OGES) yerleştirme puanlarının uzman sistemler ile tahmini. Doktora Tezi. Karabük Üniversitesi, 114 Sayfa, Karabük.

Yağcı, M. 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1).

Yan, K. and Zhang, D. 2015. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators, B: Chemical, 212, 353–363.

Yan, L. ve Liu, Y. 2020. An ensemble prediction model for potential student recommendation using machine learning. Symmetry, 12(5).

Yelegin, A. 2012. Mesleki eğitimde öğrenci altyapısının öğrenci eğitim başarısına etkisinin veri madenciliği yöntemleriyle ortaya çıkarılması. Msc. Thesis, Beykent Üniversitesi, 88 Sayfa, Istanbul.

Yurdakul, S. 2015. Veri madenciliği ile lise öğrenci performanslarının değerlendirilmesi, Yüksek Lisans Tezi, Kırıkkale Üniversitesi, 119 sayfa, Kırıkkale.

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı ve Soyadı : Mehmet ŞENLİGİL

Eğitim

Yüksek Lisans	Çankırı Karatekin Üniversitesi Fen Bilimleri Enstitüsü Elektronik-Bilgisayar Mühendisliği Anabilim Dalı	2021-2023
Lisans	Kocaeli Üniversitesi Teknik Eğitim Fakültesi Elektronik-Bilgisayar Öğretmenliği Bölümü	1995-1999
Lisans	Karabük Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü	2018-2020

İş Deneyimi

Yıl	**Kurum**	**Görev**
1999-Halen	Milli Eğitim Bakanlığı Bilgi İşlem Genel Müdürlüğü	Uygulama Geliştirme Uzmanı

Akademik Aktiviteler

1. Çopur, A., Olgun, D., Coşkun, N., Akgül, K., Doğan, H. H., Üzkurt, İ., Şenligil, M., Kaya, H. ve Ozcan, A. 2009. E-performance in inspection, Elsevier, 1(1): 2642-2647
2. Şenligil, M., Buyrukoğlu, S. 2021. Investigation of the effects of education methods in high school entrance exam (LGS) and comparison of machine learning models in the prediction of the students' success, In: 1st International Conference on Educational Technology and Online Learning.