T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİ VE İSTATİSTİKSEL YÖNTEMLER KULLANARAK ÖĞRENCİLERİN PROGRAMLAMA DERSİNDEKİ KAZANIM BİLGİLERİ İLE BAŞARI TAHMİNİ

YÜKSEK LİSANS TEZİ

Ömer DURALİOĞLU

Bilgisayar ve Bilişim Mühendisliği Anabilim Dalı

EKİM 2022

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENİMİ VE İSTATİSTİKSEL YÖNTEMLER KULLANARAK ÖĞRENCİLERİN PROGRAMLAMA DERSİNDEKİ KAZANIM BİLGİLERİ İLE BAŞARI TAHMİNİ

YÜKSEK LİSANS TEZİ

Ömer DURALİOĞLU

Bilgisayar ve Bilişim Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi M. Fatih ADAK

EKİM 2022

Ömer DURALİOĞLU tarafından hazırlanan "Makine Öğrenimi ve İstatistiksel Yöntemler Kullanarak Öğrencilerin Programlama Dersindeki Kazanım Bilgileri ile Başarı Tahmini" adlı tez çalışması 14/10/2022 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar ve Bilişim Mühendisliği Anabilim Dalı'nda Yüksek Lisans tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :	**Dr. Öğr. Üyesi** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Dr. Öğr. Üyesi** Sakarya Üniversitesi	.....
Jüri Üyesi :	**Dr. Öğr. Üyesi** Dumlupınar Üniversitesi	.....

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "Makine Öğrenimi ve İstatistiksel Yöntemler Kullanarak Öğrencilerin Programlama Dersindeki Kazanım Bilgileri ile Başarı Tahmini" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim. (14/10/2022).

Ömer DURALİOĞLU

TEŞEKKÜR

Yüksek lisans eğitimim boyunca ve tezimin hazırlanması sürecinde her konuda destek olan değerli danışman hocam Dr. Öğr. Üyesi M. Fatih ADAK'a teşekkür ederim.

Tez hazırlama sürecimde birlikte geçirebileceğimiz değerli vakitlerini feda ederek bana her zaman destek olan sevgili eşime ve biricik kızıma ayrıca teşekkür ederim.

Ömer DURALİOĞLU

İÇİNDEKİLER

Sayfa

KISALTMALAR

AutoML	: Otomatik Makine Öğrenmesi
EMT	: Ensemble Meta-Based Tree Modeli
KNN	: K-En Yakın Komşu Algoritması
LMS	: Öğrenim Yönetim Sistemi
LR	: Doğrusal Regresyon
MAE	: Ortalama Mutlak Hata
MEB	: Milli Eğitim Bakanlığı
MSE	: Ortalama Kare Hatası
NTP	: Nesne Tabanlı Programlama
R^2	: Açıklayıcılık Katsayısı
RF	: Rastgele Orman Algoritması
SVM	: Destek Vektör Makineleri
YSA	: Yapay Sinir Ağı
QDA	: Quadratic Discriminant Analysis

SİMGELER

ϵ_k	: Doğrulama hatası
β_0	: Sabit katsayı
β_{p-1}	: Eğim katsayısı
k	: Bağımsız değişken sayısı
n	: Veri seti sayısı
O	: Ağda hesaplanan değer
T	: Gerçek çıktı değeri
T_{ort}	: Gerçek çıktıların ortalaması
Y_i	: Bağımsız değişkenin i numaralı değeri

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

MAKİNE ÖĞRENİMİ VE İSTATİSTİKSEL YÖNTEMLER KULLANARAK ÖĞRENCİLERİN PROGRAMLAMA DERSİNDEKİ KAZANIM BİLGİLERİ İLE BAŞARI TAHMİNİ

ÖZET

Anahtar kelimeler: Eğitsel veri madenciliği, öğrenci performans tahmini, çok çıktılı regresyon, sentetik veri

Mevcut eğitim sisteminde öğrencilerin başarı durumları sınavlarla tespit edilmekte ve sınavdan aldığı puanlara göre kazanımlardaki eksiklikleri belirlenerek bu eksiklikler telafi edilmeye çalışılmaktadır. Ancak tespit edilen eksikliklerin daha sonraki sınavlarda ölçülecek kazanım hedeflerindeki performansına nasıl etki edeceği üzerinde durulmamaktadır. Mevcut duruma göre çözüm aranmakta, gelecek hakkında ise öğrencinin performansını arttıracak planlamalar yapılmamaktadır.

Yapılan bu çalışmada veri seti olarak, 2021-2022 eğitim öğretim yılı 1. döneminde, İstanbul ili Ataşehir ilçesinde bulunan Dr. Nureddin Erk-Perihan Erk Mesleki ve Teknik Anadolu Lisesi, Bilişim Teknolojileri alanındaki 10 ve 11'inci sınıfta okuyan 87 öğrencinin Nesne Tabanlı Programlama dersinde uygulanan 3 sınavdaki puan dağılımları kullanılmıştır. Sınavda sorulan sorular ders bilgi formundaki kazanım başlıklarıyla eşleştirilmiş, her öğrencinin kazanım başlıklarına göre performans oranları tablo haline getirilmiştir. Ancak eldeki veriler kısıtlı olduğundan ve daha anlamlı sonuçlar elde edebilmek için toplanan gerçek veriler kullanılarak sentetik veriler üretilmiştir. Sentetik verinin gerçeğe yakınlık derecesi detaylı sonuç raporu ile teyit edilmiştir.

Öğrencilerin mevcut verilerinden, sonraki sınavlardaki birden çok sayıda kazanım performansı tahmin edileceğinden, çok çıkışlı regresyonu destekleyen doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmıştır. Kullanılan algoritmaların başarı değerlendirmesi için k katmanlı çapraz doğrulama uygulanmıştır. Performans ölçümleri için MAE, MSE, R^2 ve standart sapma hesaplanmıştır. Aşırı uyum sorunu çözümü için KNN ve karar ağacı algoritmalarında en iyi parametre değerleri bulunarak performans iyileştirilmiştir. Sonuçlara göre en iyi performans değerleri KNN algoritması ile elde edilmiştir. Gerçek veri sayısı arttırılarak ve sınavlarda ölçülen kazanımlardaki puan dağılımları birbirine yaklaştırılarak daha iyi sonuçlar elde edilebilir.

Bu çalışmanın devamı olarak öğrencinin sonraki sınavlardaki başarısızlıklarını önlemek için internet ortamında bir sistem tasarlanabilir. Üretilen performans tahminlerine göre geliştirilen bu sistem öğrenciye yönlendirmelerde ve tavsiyelerde bulunulabilir. Oyunlaştırma yöntemleri kullanılarak öğrencinin ilgisini çekerek eğlenerek öğrenmesi sağlanabilir.

PREDICTION OF SUCCESS WITH STUDENTS' GAIN INFORMATION IN THE PROGRAMMING COURSE USING MACHINE LEARNING AND STATISTICAL METHODS

SUMMARY

Keywords: Educational data mining, student performance prediction, multiple output regression, synthetic data

In the current education system, the success of the students is determined by exams and the deficiencies in the achievements are determined according to the scores they get from the exam and these deficiencies are tried to be compensated. However, it is not focused on how the identified deficiencies will affect the performance of the achievement targets to be measured in the next exams. Solution is sought according to the current situation, and plans are not made for the future to increase the student's performance.

In this thesis, as the data set, in the 1st semester of the 2021-2022 academic year, The score distributions of 87 students studying in the 10th and 11th grades of Dr. Nureddin Erk-Perihan Erk Vocational and Technical Anatolian High School in the field of Information Technologies in the 3 exams applied in the Object Oriented Programming course were used. The questions asked in the exam were matched with the achievement titles in the course information form, and the performance rates of each student were tabulated according to the achievement titles. However, since the available data are scarce and in order to obtain more meaningful results, synthetic data has been produced by using the collected real data. The degree of closeness of the synthetic data to real data was confirmed by the detailed result report.

Linear regression, k-nearest neighbor and decision tree algorithms supporting multi-output regression were used, since multiple achievement performances in the next exams would be estimated from the students' existing data. K-layer cross validation was applied to evaluate the success of the algorithms used. MAE, MSE, R2 and standard deviation were used for performance measurements. For the solution of the overfitting problem, the performance was improved by finding the best parameter values in the KNN and decision tree algorithms. According to the results, the best performance values were obtained with the KNN algorithm. Better results can be obtained by increasing the number of actual data and approximating the score distributions in the achievements measured in the exams.

As a future study, an online system can be designed to prevent the student's failures in the next exams. This system, which is developed according to the performance estimates produced, can provide guidance and advice to the student. By using gamification, it can be ensured that the student learns while having fun by attracting attention.

1. GİRİŞ

Gelişen teknoloji sayesinde saklanan ve kullanılan veri boyutları her geçen gün artmaktadır. Verilerin ne şekilde kullanılarak anlamlı hale getirilebileceği araştırıldıkça veri madenciliği ve makine öğrenmesi yöntemleri önem kazanmıştır. Veri madenciliği, çok büyük miktardaki veri içerisinden işe yarayacak verileri yöntemler kullanarak kullanılabilir hale getirmektir. Makine öğrenmesi, sisteme girilen veriler ve kullanılan algoritmalar sayesinde bilgisayarın öğrenmesini sağlayan tekniktir. Bu alanlarda yapılan araştırmalar gün geçtikçe daha çok önem kazanmaktadır.

Veri madenciliği ve makine öğrenmesi yöntemleri ile birçok alanda olduğu gibi eğitim alanında da yapılan çalışmalar artarak devam etmektedir. Öğrencilerin mevcut verileri kullanılarak gelecekteki durumları hakkında tahminler üretilmekte, bu tahminlere göre eğitim sürecinin şekillenmesi istenmektedir. Elbette ki çalışmalardaki temel amaç, bu yöntemler sayesinde öğrencilerin performansını arttırarak daha iyi eğitim almalarını sağlamaktır.

Bu tez çalışmasında öğrencilerin girdiği sınavdaki ders kazanımlarındaki performanslarına göre sonraki girecekleri sınavlardaki kazanım performansları tahmin edilmiştir. Bunun için çok çıkışlı regresyonu destekleyen doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmış ve elde edilen analiz sonuçları karşılaştırılmıştır.

Veri olarak 2021-2022 eğitim öğretim yılı 1. döneminde, İstanbul ilindeki Ataşehir Dr. Nureddin Erk - Perihan Erk Mesleki ve Teknik Anadolu Lisesi, Bilişim Teknolojileri alanındaki 10 ve 11'inci sınıfta okuyan 87 öğrencinin Nesne Tabanlı Programlama dersinde uygulanan 3 sınavın not dağılımları kullanılmıştır. Öncelikle sınav soruları ile ders kazanımları eşleştirilmiş, öğrencilerin her sorudan aldığı puanlar eşleştiği ilgili kazanımlara göre oranlanarak 0 ile 100 arasındaki performanslarını gösterecek şekilde hesaplanarak tablo haline getirilmiştir. Mevcut veriler az olduğu için 5000 sentetik veri üretilerek kullanılmıştır. Böylece elde edilecek sonuçların daha anlamlı olması sağlanmıştır.

1.1. Literatür Taraması

Aydemir (2019), çalışmasında Türkiye'de bir üniversitedeki öğrencilerin yabancı dil dersinden geçme notlarını veri madenciliği ile tahmin edilmiştir. Dersi alan 3794 öğrencinin verileri, 12 girdi ve 1 çıktı olacak şekilde tasarlanmıştır. Weka programında bulunan veri madenciliği yöntemleriyle uygulama yapılmış ve uygulanan modeller arasında ortalama mutlak hatası 1.22, korelasyon katsayısı 0.80 ile Bagging yöntemi en iyi sonucu vermiştir [1].

Gök (2017), Türkçe ve Matematik derslerindeki başarıyı etkileyen etmenleri tespit edebilmek amacıyla, 6, 7 ve 8. sınıfta okuyan öğrencilere 24 soru içeren bir anket uygulamıştır. Anket ile sağlanan veriler, regresyon / çok sınıflı makine öğrenmesi modelleri ile ilgili ders ve dönem sonundaki başarı durumları tahmin edilmiştir. Puan tahmininde en iyi sonucu rastgele orman yöntemi, not tahmininde ise korelasyon tabanlı öznitelik alt kümesi yöntemi ile lojistik sınıflandırma algoritması birlikte kullanılınca en iyi sonucu vermiştir [2].

Güner ve Çomak (2010), Pamukkale Üniversitesi mühendislik fakültesindeki 434 öğrencinin üniversiteye giriş sonuçlarındaki testlerdeki başarı durumları ve lise mezuniyet puan verileriyle üniversitedeki Matematik 1 dersindeki başarılarının tahmininde bulunmuştur. Veri kümesinde 20 özellik kullanılmış, bu küme içindeki 3 özellik ise sadece test amacıyla kullanılmıştır. Karar destek sistemi ile %86,36 oranında doğru tahmin edilmiştir [3].

Abbasoğlu (2020), ortaokul öğrencilerinin yılsonu başarı ortalamasına etki eden demografik özelliklerin ve sosyoekonomik durumların eğitsel veri madenciliği ile tahminde bulunmuştur. 2019-2020 eğitim öğretim yılı 2.döneminde, Yalova'da demografik olarak faklı 4 resmi ortaokuldaki 5, 6, 7 ve 8. sınıfta okuyan 1395 öğrencinin, 27 bağımsız değişken verisi kullanılmıştır. Weka programındaki tüm algoritmalar kullanılmıştır. Sınıflandırıcı algoritmalarının kullanılması ile yılsonu genel başarı ortalamasında lojistik algoritması en iyi tahmini gerçekleştirmiştir [4].

Aghalarova ve Bozkurt Keser (2021), öğrencilerin akademik performansını tahmin etmek için Otomatik Makine Öğrenmesi (AutoML) ile veri seti için en iyi modeli araştırılmıştır. Veri olarak Kalboard 360 e-öğrenme sistemindeki 480 öğrencinin 17 özniteliği kullanılmıştır. AutoML yöntemi ile KNN ve SVM algoritmaları

karşılaştırılmaktadır. Önerilen AutoML yönteminin daha iyi sonuçlar verdiği görülmüştür [5].

Başer ve ark. (2020), ortaöğretimdeki öğrencilerin başarı ölçütü tespitinde bireysel ve demografik sınıflandırma için yöntemler uygulanmıştır. Veriler UC Irvine Machine Learning Repository veri tabanından 395 öğrencinin 31 özniteliği şeklinde elde edilmiştir. Weka uygulaması ile Iterative Classifier, OneR, LogitBoost ve Yapay sinir ağları yöntemleri kullanılmış ve en iyi sonuçlar OneR yöntemi ile elde edilmiştir [6].

Şengür ve Tekin (2013), yapay sinir ağları ve karar ağacı ile öğrencilerin mezuniyet notları tahmin edilmiştir. Veri olarak 2011 yılındaki Fırat Üniv., Eğitim Fak., Bilgisayar ve Öğretim Teknolojileri Eğitimi bölümünden mezun olmuş 127 öğrencinin 4 yılda aldığı 49 dersin yılsonu notlarından faydalanılmıştır. Yapılan çalışmalar sonucunda, Yapay sinir ağlarının karar ağaçlarına göre daha iyi tahmin sonucu verdiği görülmüştür [7].

Altun ve ark. (2019), Akdeniz Üniv. Eğitim Fak. Sınıf Öğretmenliği bölümünden mezun olan 578 öğrencinin verileri kullanılarak mezuniyet notlarının tahmini için çalışma yapılmıştır. Çoklu doğrusal regresyon analizi ile %94.30 ve yapay sinir ağları ile %94.43 başarı oranı bulunmuş, yapılan değerlendirmede birbirine yakın sonuçlar çıktığı görülmüştür [8].

Aydemir ve ark. (2019), çalışmasında öğrencilerin Türk Dili dersindeki başarıları veri madenciliği ile tahmin etmiştir. Türkiye'deki bir devlet üniversitesindeki bu dersi alan 160 öğrencinin verileri kullanılmıştır. Decision stump, random tree, random forest, REP tree ve M5P yöntemleri kullanılmıştır. Weka uygulamasında kullanılan algoritmalarda en yüksek başarı oranı randon forest olduğu görülmüştür [9].

Kanchana ve ark. (2021), çalışmasında öğrenci başarısı tahmini için SVM, naive bayes ve karar ağacı yöntemleri kullanılmıştır. Veri olarak Sri Lanka'da bir üniversitede mühendislik alanında 2015-2016 akademik yılındaki 126 öğrencinin 3 dönemdeki bilgileri kullanılmıştır. Üç model ile %77.0 ile %80.5 arasında doğruluk bulunmuştur [10].

Qazdar ve ark. (2019), çalışmasında öğrencilerin başarı durumlarını tahmin etmek için Python ile çok değişkenli regresyon algoritması kullanılarak iki model tasarlanmıştır. Veriler 2016 ile 2018 yılları arasında Fas'taki H.E.K. Lisesi'nde 478 öğrencinin verisi kullanılmıştır. Sistemin sonuçlarının kontrolü için ilk aşamada MAE, RMSE, RAE, RSE gibi ölçütler kullanılmıştır. İkinci aşamadaki değerlendirme okul müdürünün başkanlığındaki kurul ile yapılmıştır. Sonuçta üzerinde çalışılan sistemin öğrenci performansı hakkında daha iyi tahminler yaptığı gözlenmiştir [11].

Hasib ve ark. (2022), çalışmasında Lojistik Regresyon, KNN, SVM, XGBoost ve naive bayes algoritmalarını kullanarak ortaöğretimdeki öğrenciler hakkında başarı tahmini yapmaktadır. Veriler Portekiz okul raporları ve anketlerinden elde edilmiştir. Test sonuçlarına göre en iyi doğruluğu %96.89 oranında SVM ile elde edilmiştir [12].

Bujang ve ark. (2021), çalışmasında öğrencilerin önceki notlarına göre final notlarını tahmin etmek için karar ağacı, rastgele orman, destek vektör makinesi, lojistik regresyon algoritmaları kullanılmıştır. Veri olarak Malezya Politeknik'teki 489 Bilgi ve İletişim Teknolojisi bölüm öğrencileri kullanılmıştır. Elde edilen sonuçlarda, %99.8 oranındaki en yüksek doğruluğun karar ağacı algoritması (J48) olduğu görülmüştür [13].

Ezz ve Elshenawy (2020), çalışmasında AL-Azhar Üniversitesi'inde mühendislik bölümü hazırlık sınıfında okuyan öğrencilerin tahmin edilen akademik başarı durumuna göre yedi mühendislik bölümünden biri önerilmektedir. Veriler AL-Azhar Üniversitesi'nde 2012-2018 yılları arasında kayıtlı 1841 öğrenciden elde edilmiştir. SVM, QDA, KNN, LR, RF algoritmaları kullanılmış ve her mühendislik bölümü için farklı algoritmanın doğruluk oranının yüksek çıktığı görülmüştür [14].

Tomasevic ve ark. (2019), çalışmasında öğrencilerin sınav sonuçlarını tahmin etmiştir. Öğrencinin önceki notlarının sonuca etki ettiği ancak demografik verilerin tahminlerin kesinliği üzerinde önemli bir etkisinin olmadığı gözlenmiştir. Open University Learning Analytics Dataset (OULAD) ile 32593 öğrencinin verisi kullanılmıştır. En iyi sonuçları YSA modellerinin verdiği görülmüştür. Beş ve altıncı değerlendirmeden sonra ise SVM biraz daha iyi performans göstermiştir [15].

Trakunphutthirak ve Lee (2021), çalışmasında çevrimiçi öğrenme yönetim sistemindeki (LMS) verileri kullanarak öğrencinin akademik performansını etkileyen faktörler hakkında tahminde bulunmuştur. Bunun için bir Tayland üniversitesinden alınan 1708 öğrencinin LMS verileri, internet kullanım günlük dosyaları, zamansal ve demografik veriler kullanılmıştır. Verileri farklı ağırlıklarda kullanarak algoritma verimliliklerinin farklı sonuçlar verdiği gözlenmiştir [16].

Buraimoh ve ark. (2021), çalışmasında altı makine öğrenimi modeliyle öğrenci tahmini yapmıştır. Veriler Kalboard 360 LMS'den alınarak 480 öğrenci verisi kullanılmıştır. Dengesiz veriler performans düşüklüğüne sebep olduğu için Smote yöntemi uygulanmıştır. Regresyon ağacı ve lojistik regresyon 0.86 doğruluk oranı ile en iyi performansı göstermiştir [17].

Auwal ve ark. (2020), çalışmasında doğru performans tahmini için öğrenci özelliklerini kullanarak bir sınıflandırıcı geliştirilmiştir. Veriler Nijerya'daki bir yükseköğretim kurumundan elde edilmiştir. Toplam 250 anket uygulanmış ancak 149'u doğru şekilde doldurularak geri gönderilmiştir. Weka yazılımı ile demografik, bilişsel, bilişsel olmayan veriler ile analiz yapılmıştır. En doğru sınıflandırıcı olarak naive bayes bulunmuştur [18].

Balaji ve ark. (2021), çalışmasında arama kriterlerine göre 2700 makale değerlendirerek, oluşturulan kalite puanlarına göre 56 makaleye kadar azaltılmıştır. Bu çalışmalardaki makine öğrenmesinin öğrenci akademik performans tahmine katkıları incelenmiştir [19].

Miguéis ve ark. (2018), çalışmasında üniversitedeki ilk yılın sonunda elde edilen verileri, akademik performansını tahmin etmek için kullanmıştır. Bunu iki aşamalı bir model ile gerçekleştirmiş ve model tarafından öngörülen öğrenci performans seviyelerine göre öğrencileri bölümlere ayırmayı önermektedir. Bir kamu araştırma üniversitesinin Avrupa Mühendislik Okulu'ndan 2003 ile 2015 yıllarını kapsayan 2459 öğrenciden oluşan veri seti kullanılmıştır. En iyi sonucu rastgele orman algoritması, en kötü sonucu ise Naive Bayes algoritmasının verdiği görülmüştür [20].

Yağcı (2022), çalışmasında sinir ağları, rastgele orman, destek vektör makineleri, lojistik regresyon, naive bayes ve k-en yakın komşu algoritmaları kullanılarak öğrencilerin yarıyıl sonu notlarını tahmin etmiştir. Algoritmaların performansları

karşılaştırılmıştır. Veriler 2019-2020 güz döneminde Türkiye'deki bir devlet üniversitesindeki Türkçe-1 dersini alan 1854 öğrencinin notlarının öğrenci bilgi sisteminden alınmasıyla elde edilmiştir. Rastgele orman algoritması ile %74,6 doğruluk elde edilmiştir [21].

Almasri ve ark. (2019), çalışmasında Ensemble Meta-Based Tree modeli (EMT) ile öğrenci performansını tahmin etmiştir. EMT modeli, 47 öğrenme tekniğinin birleştirilmesiyle oluşturulmuştur. Veri olarak 13 özellik içeren 400 öğrenci kaydı kullanılmıştır. J48 algoritmasının %94,3 doğruluk değerlerine sahip olduğu, boosting yöntemini uygulayarak %98,3 doğruluk değerine ulaştığı gözlenmiştir [22].

Koç ve Akın (2022), çalışmasında belirli değişkenleri kullanarak öğrencilerin başarı durumlarını modelleyip tahmin etmiştir. 2019 yılında Türkiye'deki 81 ilde lise giriş sınavına (LGS) giren öğrencilerin başarı oranları makine öğrenmesi regresyon ve beta regresyon modeli ile hesaplanmıştır. Destek vektör regresyonu, rastgele orman, karar ağacı ve beta regresyon modeli kullanılmıştır. Beta regresyon ve rastgele orman modelleri en iyi sonucu vermiştir [23].

Yavuzarslan ve Erol (2022), çalışmasında 2020 Bahar yarıyılındaki Temel Bilgisayar Uygulamaları dersindeki 93 öğrencinin 10 hafta süreyle kullandıkları Moodle tabanlı öğrenme yönetim sistemindeki log verileri kullanılarak başarı tahmini yapılmıştır. Bu amaçla KNN, Naive Bayes, SVM, CART ve C5.0 algoritmaları kullanılmıştır. Veri setleri dengesiz olduğundan dolayı SMOTE yöntemi kullanılmıştır. En iyi sonuçların %97 başarı oranıyla CART ve SVM algoritmalarıyla elde edildiği görülmüştür [24].

Ram ve ark. (2021), çalışmasında öğrencilerin başarı durumunu tahmin etmek için doğrusal regresyon ve rastgele orman algoritmaları karşılaştırılmıştır. archive.ics.uci.edu/ml/datasets/student+performance web adresindeki verilerden faydalanılmıştır. %92 başarı oranıyla rastgele orman algoritması ile elde edilen sonuçların daha iyi olduğu görülmüştür [25].

Wakelam ve ark. (2020), çalışmasında az öğrencinin olduğu durumlarda başarı tahmini yapmıştır. Üniversite son sınıfta okuyan 23 kişilik bir grubun verileri kullanılmıştır. Karar ağacı, k-en yakın komşu ve rastgele orman algoritmaları kullanılmıştır. En doğru tahminleri rastgele orman algoritması gerçekleştirmiştir [26].

Haridas ve ark. (2019), çalışmasında akıllı öğretim sistemi olan AmritaITS kullanan 2123 Hintli öğrencinin 3 yıllık verilerini kullanarak öğrencilerin İngilizce ve Matematik derslerindeki performanslarını tahmin etmiştir. Karar ağacı, lojistik regresyon, sinir ağları, SVM, KNN ve rastgele orman modelleri kullanılmıştır. KNN ve rastgele ormanın en iyi oranları verdiği görülmüştür [27].

Sixhaxa ve ark. (2022), çalışmasında demografik, akademik ve davranışsal özelliklerinin öğrenci performansı üzerindeki etkileri tahmin edilmiştir. Kalboard 360 LMS sisteminden 16 özelliğe sahip 480 veri örneği elde edilerek kullanılmıştır. Çalışmada gaussian naive bayes, destek vektör makinesi, random forest, k-en yakın komşu ve lojistik regresyon algoritmaları kullanılmıştır. %81.67 doğruluk oranıyla en iyi performans gösteren model destek vektör makinesi olmuştur [28].

Hashim ve ark. (2020), çalışmasında öğrencilerin final notlarını tahmin etmek için karar ağacı, naive bayes, lojistik regresyon, destek vektör makinesi, K-en yakın komşu, sıralı minimal optimizasyon ve sinir ağları algoritmalarının performansları karşılaştırılmıştır. Çalışmada 2017-2018 ve 2018-2019 akademik yıllarında Basra Üniversitesi, Bilgisayar Bilimleri ve Bilişim Teknolojileri Fakültesi lisans eğitim programındaki derslerden elde edilen 499 öğrenci verisi kullanılmıştır. Lojistik regresyonun en iyi performansı gösterdiği sonucuna varılmıştır [29].

Öztürk (2022), tez çalışmasında Anadolu Üniv., Anadolum eKampüs Öğrenme Yönetim Sistemini kullanan öğrencilerin özelliklerinin belirlenmesi, akademik performanslarının tahmin edilmesi, sistemle uyumlu bir performans değerlendirme yapısının oluşturulması ve bu yapının etkilerinin incelenmesini yapmıştır. Tahmin modeli geliştirilirken makine öğrenmesi ve derin öğrenme algoritmalarından faydalanılmıştır. Öğrencilerin notları GBT algoritmasıyla %72,16 doğrulukta tahmin edilmiştir [30].

Selvi (2020), tez çalışmasında öğrencilerin akademik performansını tahmin etmek amaçlamıştır. Akademik performansı etkilediği düşünülen değişkenleri elde etmek için, 2019 yılındaki sınavla öğrenci almakta olan 4 değişik lise grubunda öğrenim gören 9. sınıftaki öğrenciler ile ortaokulda öğrenim görmekte olan 8. sınıfta okuyan öğrencilere 32 soruluk anket uygulanmıştır. Yapay sinir ağları, J48, rastgele orman,

karar ağacı, RapTree ve HoeffdingTree algoritmaları kullanılmıştır. Rastgele orman algoritmasının en iyi sonucu verdiği görülmüştür [31].

Sulak (2021), tez çalışmasında açık öğretim lisesine kayıtlı öğrencilerin normal sürelerinde mezun olabilme durumlarını tahmin etmiştir. Veri olarak 2010-2012 arasındaki açık öğretim lisesine kayıt olan 142714 öğrenci bilgisi kullanılmıştır. Karar ağacı, k-en yakın komşu, destek vektör makinesi ve yapay sinir ağları algoritmaları kullanılmıştır. Karar ağacı algoritması %99.99 doğru tahmin oranı ile en iyi sonucu vermiştir [32].

Özdemir (2016), tez çalışması belirlenen faktörlerin etkisiyle öğrencilerin başarılarının nasıl etkilendiğinin öngörülmesine dayanmaktadır. Veriler İstanbul'un farklı ilçelerindeki liselerden 2371 öğrenciden elde edilmiş, 1706 öğrencinin verisi çalışmada kullanılmıştır. k-en yakın komşu, naive bayes, C4.5 karar ağacı, logistik regresyon ve destek vektör makinesi algoritmaları kullanılmıştır. C4.5 karar ağacı algoritmasının en iyi sonuçları verdiği görülmüştür [33].

Can (2021), tez çalışmasında üniversite sınavına giren öğrenci verileri ile veri madenciliği ve makine öğrenmesiyle sınavdaki başarı oranı tahmini yapılmıştır. Veriler 1979-2020 arasındaki üniversite sınavına giren 677 adaya uygulanan anketlerden elde edilmiştir. Doğrusal regresyon, rastgele orman, destek vektör makinesi, k-en yakın komşu ve gaussian NB algoritmaları kullanılmış, en iyi sonucu %73.77 doğruluk oranı ile gaussian NB modeli vermiştir [34].

Can (2021), tez çalışmasında eğitim alanında kullanılan makine öğrenmesi algoritmalarından hangisinin kullanılabileceğini tespit etmeye çalışmıştır. Veriler Kaliforniya Üniv. Makine Öğrenmesi Veri Havuzu'ndan alınmıştır. Bu veri seti, üç öğretim elemanının Gazi Üniversitesi'nde 5820 öğrenci tarafından likert tipi ölçek ile değerlendirilmesiyle oluşturulmuştur. Karar ağacı, rastgele orman ve yapay sinir ağları algoritmaları incelenmiştir. En iyi sonuçlar %98.57 doğru sınıflama oranı ile karar ağacı algoritmasıyla elde edilmiştir [35].

Olgun (2021), tez çalışmasında veri madenciliği yöntemlerinden sınıflandırma yöntemleri ile, ters yüz sınıflardaki video izlenme durumlarının verileri kullanılarak öğrencilerin başarı durumları tahmin edilmiştir. Veri olarak Temel Bilgisayar Uygulamaları dersindeki 404 üniversite öğrencisinin 4 dönem boyunca video izleme

verileri kullanılmıştır. Toplamda 17 nitelik kullanılmıştır. SMOTE tekniği ile veri seti dengelenmiş, sınıflandırma algoritması olarak rastgele orman, destek vektör makineleri ve naive bayes kullanılmıştır. Karşılaştırma sonucunda, %83.11 doğruluk oranıyla en iyi algoritma rastgele orman olmuştur [36].

Şahin (2021), tez çalışmasında ortaokul öğrencilerinin sosyoekonomik, demografik özellikleri ve ders notları kullanılarak yılsonu başarı durumlarının makine öğrenmesi yöntemleri ile tahmini ve öğrencilerin başarısına etki eden nitelikler belirlenmiştir. İstanbul, Tuzla'da bir ortaokulda 2019-2020 eğitim yılındaki 328 öğrencinin verileri kullanılmıştır. Sınıflandırma için k-en yakın komşu, karar ağacı, rastgele orman, destek vektör makinesi, yapay sinir ağları, lojistik regresyon ve naive bayes algoritmaları kullanılmıştır. Çalışma sonucunda en iyi sonucu verdiği için rastgele orman modeli ile akademik başarı tahmini için bir sistem geliştirilmiştir [37].

Özarslan (2014), tez çalışmasında Kırıkkale Üniversitesi birinci sınıfta okuyan ve Temel Bilgi Teknolojileri Kullanımı dersini yüz yüze ve uzaktan alan 672 öğrenciye ait veriler sınıflandırma algoritmaları ile incelenerek başarıya etki eden faktörler belirlenmiştir. Çalışmada WEKA yazılımı kullanılmış ve en iyi sonucu %82.2222 doğruluk oranı ile J48 karar ağacı algoritmasının verdiği görülmüştür [38] .

Kayhan (2019), tez çalışmasında 2016-2017 güz döneminde Amasya Üniv., Uzaktan Eğitim Merkezi'ndeki 4 ön lisans programına kayıt yaptıran öğrencilerinin normal süresinde mezun olma durum tahmini yapılmaya çalışılmıştır. Çalışmada karar ağacı, naive bayes, destek vektör makinesi, rastgele orman ve yapay sinir ağları algoritmaları kullanılmıştır. Çocuk gelişimi programında en iyi sonucu destek vektör makinesi verirken, diğer bölümlerde ise en iyi sonucu karar ağaçları algoritması vermiştir [39].

Altınsoy (2019), tez çalışmasında yapay zekâ yöntemleri kullanılarak üniversitedeki eğitim yöntemlerinin başarı durumuna etkisi hakkında değerlendirmeler yapılmıştır. Veri olarak erişime açık olan HarvardX-MITX çevrimiçi derslerindeki (2012 güz, 2013 bahar ve 2013 yaz) veriler kullanılmıştır. Destek vektör makinesi, lojistik regresyon, sade bayes, bayes net, rastgele orman, rastgele ağaç ve karar ağacı algoritmaları kullanılmıştır. En iyi sonucu % 98.4058 başarı oranı ile karar ağacı algoritmasının verdiği görülmüştür [40].

Göker (2012), tez çalışmasında üniversite sınavına giren öğrencilerin veri madenciliği ile başarı tahminlerinin yapılarak bir anlamda başarı konusunda erken uyarı sistemi geliştirmiştir. Gölbaşı Ahmet-Alper Dinçer Lisesi mezun öğrencilerinden elde edilen 220 adet veri kullanılmıştır. Bu amaçla algoritmalar karşılaştırılmış ve en iyi sonucu veren %87.27 başarı oranı ile naive bayes algoritması ile Microsoft Visual Studio 2008 C#.Net kullanılarak bir yazılım geliştirilmiştir [41].

Yurdakul (2015), tez çalışmasında öğrenci performansına etki eden faktörlerin belirlenip, bu faktörlerin arasındaki ilişkisi araştırılmıştır. Kırıkkale ilindeki Anadolu Liselerinde okuyan 11. sınıf öğrencilerine uygulanan anket sonucunda elde edilen 231 adet veri kullanılmıştır. Weka 3.7 programı kullanılmıştır. En iyi sonucun %88.73 ile çok katmanlı algılayıcı algoritmasının verdiği görülmüştür [42].

Aydemir (2017), tez çalışmasında veri madenciliği sınıflama algoritmaları içinden en iyisini seçerek meslek yüksek okulundaki öğrencilerinin başarıları durumlarını tahmin etmektir. Pamukkale Üniv. meslek yüksek okullarına 2009- 2010 arasında kayıt olan 1.387 öğrencinin bilgilerinden faydalanılmıştır. Kullanılan veriler içindeki akademik not ortalamasına göre yapılan tahminde en iyi performansı sıralı minimum optimizasyon algoritması göstermiştir. Mezuniyet yılına göre yapılan tahminde ise en iyi performansı J4.8 ve naive bayes göstermiştir [43].

Deshmukh (2015), tez çalışmasında öğrencilerin okulu bırakmasını önlemek amacıyla risk altındaki öğrencilerin tespit edilmesi amaçlanmıştır. İleri beslemeli sinir ağları, lojistik regresyon ve destek vektör makinesi sınıflandırma algoritmaları uygulanmıştır. Veriler 2012-2017 arasındaki Nevada Las Vegas Üniv. bilgisayar bilimlerindeki 508 lisans öğrencisinden toplanmıştır. Yetersiz veri içeren sınıftaki veri noktalarını artırmak için rastgele aşırı örnekleme tekniği kullanılmıştır. Birinci yıl için risk altındaki öğrencilerin tahmin edilmesinde en iyi performansı lojistik regresyon, ikinci yıl için tahmin edilmesinde ise ileri beslemeli sinir ağlarının en iyi performansı gösterdiği görülmüştür [44].

Bastem (2021), tez çalışmasında öğrencilerin akademik başarı makine öğrenmesi yöntemleriyle tahmin edilmiştir. Veriler Kaggle web sitesi üzerinden Portekiz'de bulunan iki okulun 649 öğrencinin bilgilerinden elde edilmiştir. Karar ağacı, rastgele

orman ve lojistik regresyon algoritmaları kullanılmıştır. En iyi doğruluk oranını karar ağacı algoritması vermiştir [45].

Olga (2016), tez çalışmasında öğrencilerin performansını ve yılsonu sınavlarının sonuçlarını tahmin etmiştir. 247 örnek ve 5 özellikten oluşan veri seti kullanılmıştır. SPSS ve WEKA yazılımlarından faydalanılmıştır. Bayes net, naive bayes, çok katmanlı algılayıcı, J48 ve destek vektör makinesi algoritmaları kullanılmıştır. En düşük maliyetin destek vektör makine algoritması ile elde edilmiştir. Bu sebeple tahmin için bu modelin kullanılması önerilmiştir [46].

Bah (2018), tez çalışmasında Atılım Üniv. bilgi sistemlerindeki öğrenci bilgilerini kullanarak naive bayes, lojistik regresyon, çok katmanlı algılayıcı, destek vektör makinesi, k-en yakın komşu ve karar ağacı algoritmalarının öğrenci performansını tahmin etmeleri incelenmiştir. Veri olarak Atılım Üniv. Yazılım Mühendisliği'ndeki 185 lisans öğrencisinin bilgileri kullanılmıştır. Naive bayes ve k-en yakın komşu algoritmaları en iyi sonucu vermiştir [47].

2. METOT

2.1. Veri Madenciliği

Veri madenciliği, çok büyük miktardaki bilginin saklandığı veri tabanlarından, istediğimiz amaca uygun şekilde gelecek ile ilgili tahminler yapmamıza yarayacak anlamlı veriye ulaşma ve veriyi kullanma işidir [48].

Veri madenciliğinde izlenen adımlar genellikle aşağıdaki gibidir [49]:

1. Problemin tanımlanması,

2. Verilerin tanımlanması ve toplanması,

3. Verilerin hazırlanması,

4. Modelleme,

5. Modelin değerlendirmesi,

6. Modelin kullanımı.

Problemin tanımlanması: Bir veri madenciliği projesindeki en önemli aşamadır. Bu aşamada hedefler iyi belirlenmeli, verilerin nasıl kullanılacağı ve nasıl analiz edileceği çok iyi planlanmalıdır. Projenin işleyiş için gerekli olan personelden yazılımlara kadar her şey iyi planlanarak süreç hakkında bir zaman çizelgesi hazırlanmalıdır.

Verilerin tanımlanması ve toplanması: İlk verilerin toplanması, verilerin tanımlanması, verilerin araştırılması ve veri kalitesinin analizi adımlarından oluşur. Burada elde edilen veriler daha sonraki adımların geliştirilmesine etki edeceği için verilerin hedeflere uygunluğu ve veri kalitesi iyi analiz edilmelidir.

Verilerin hazırlanması: Verilerin seçilmesi, verilerin temizlenmesi, verilerin oluşturulması, verilerin bütünleştirilmesi ve verilerin biçimlendirilmesi adımlarından oluşur. Analizde kullanılacak veriler seçilir ve bu veriler temizlenir. Kullanılacak modelde gerekliyse mevcut verilerden yeni veriler türetilebilir. Verileri farklı

kaynaklardan elde etmek gerekiyorsa hepsi birbirine uyum sağlayacak şekilde birleştirilir.

Modelleme: Modelleme tekniğinin seçimi, test tasarımının oluşturulması, modellerin oluşturulması ve modellerin değerlendirilmesi adımlarından oluşur. Veri madenciliği problemleri için mevcut olan birçok teknik içinden seçilen tekniğe uygun verilerin formatının değiştirilmesi gerekiyorsa, önceki adımlara dönülerek düzenleme yapılması gerekebilir. Eldeki verilerin bir kısmı seçilen modeli eğitilmesi için kullanılır. Kalan kısım ile de eğitilerek oluşturulan modelin doğruluğu test edilerek, hata oranları kontrol edilir.

Modelin değerlendirmesi: Sonuçların değerlendirilmesi, sürecin kontrol edilmesi ve sonraki adımların belirlenmesi adımlarından oluşur. Oluşturulan modelin beklenen hedefi ne kadar karşıladığı ve eksik yanları kontrol edilir. Gerçek hayat koşullarına göre geliştirilen uygulamalar üzerinde oluşturulan model test edilir ve detaylı inceleme yapılarak projenin kullanılması ya da yeni projeye geçilmesi için karar verilir.

Modelin kullanımı: Model kullanımının planlaması, izleme ve bakım planlaması, sonuç raporunun üretilmesi ve projenin gözden geçirilmesi adımlarından oluşur. Geliştirilen proje, kullanıcıların yararlanabileceği şekilde tasarlanır. Oluşturulan modelden elde edilecek sonuçlar günlük hayatta kullanılmaya devam edilecekse, modelin izlenmesi gerekir. Modelin işleyişinin izlenmesi, hatalı sonuçların alınmasını önlemek açısından önemlidir. Projenin sonunda tüm çıktıları içeren ve süreci özetleyen bir rapor hazırlanması, daha sonraki süreçte yapılacak olan başka projelerin iyileştirilmesine yardımcı olacaktır.

Veri madenciliği modelleri işlevlerine göre sınıflama ve regresyon, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntüler olarak 3 grupta toplanır. Sınıflama ve regresyon modelleri tahmin edicidir ve mevcut verilerden hareket edilerek gelecek ile ilgili tahmin yürütülmesi amaçlanır. Sınıflama ve regresyon modellerinde karar ağacı, yapay sinir ağları, genetik algoritmalar, k-en yakın komşu, naive-bayes, lojistik regresyon gibi bazı teknikler kullanılır.

Kümeleme ile birliktelik kuralları ve ardışık zamanlı örüntü modelleri ise tanımlayıcıdır. Küme üyelerinin birbirine çok benzediği halde özellikleri farklı olan

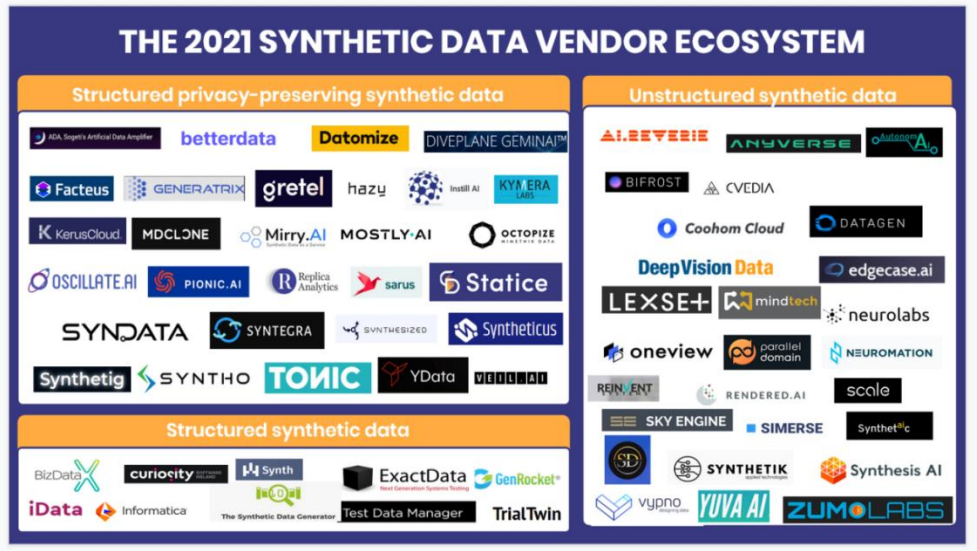
kümelerin bulunmasından sonra veri tabanında bulunan kayıtların buradaki farklı kümelere bölünmesidir [50].

Veri madenciliği bankacılık, sigortacılık, pazarlama, tıp ve biyoloji, e-ticaret gibi birçok alanda kullanılmaktadır. Son yıllarda eğitim alanında da kullanıldığı görülmektedir. Bu durum eğitsel veri madenciliği ve öğrenme analitiği kavramlarının ortaya çıkmasına sebep olmuştur. Eğitsel veri madenciliği disiplinler arası bir araştırma alanıdır ve eğitim sürecinde yer alan bütün sistemlerden gelen verileri inceleyerek faydalı bilgiler üretmeyi amaçlamaktadır. Öğrenme analitikleri ise keşfedilen bilgilerin öğrenme sürecinin iyileştirilmesi ve öğretim tasarımında kullanılması ile ilgilenir. Eğitsel veri madenciliğinde mevcut veri madenciliği teknikleri kullanılırken, öğrenme analitiklerinde ise bu tekniklerin yanında istatistiksel ve görselleştirme araçları ile sosyal ağ analizi teknikleri gibi tekniklerden de faydalanılmaktadır [51].

2.2. Sentetik Veri Üretimi

Sentetik veri üretimi ile elde edilmiş mevcut verilerin benzerlerini sağlanır. Sentetik veriler gerçek verilerle ilişkili yapay verilerdir. Verilerin yetersiz olduğu durumlarda, veri elde etmenin zor ve maliyetli olduğu durumlarda ya da eksik verilerin tamamlanması gerektiği durumlarda kullanımı oldukça faydalıdır [52].

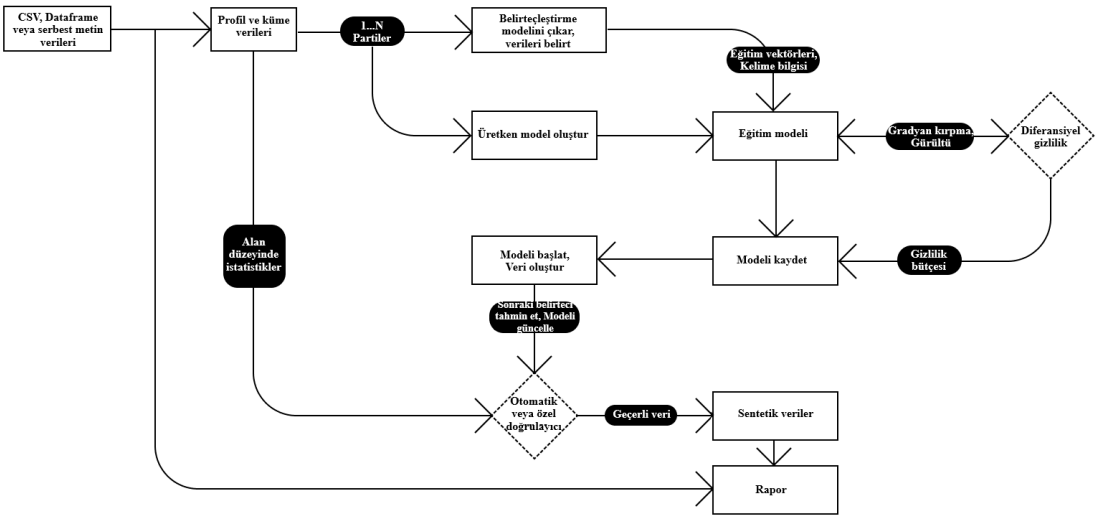
Otomotiv, robotik, finansal hizmetler, siber güvenlik, sosyal medya, oyun, sağlık, üretim, perakende gibi birçok alanda sentetik veri ile geliştirilen uygulamalar kullanılmaktadır. Ek olarak veriler pazarlama, makine öğrenmesi, operasyonlar, insan kaynakları gibi iş alanlarını hızlandırmak amacıyla da sentetik veriler kullanılmaktadır [53]. Şekil 2.1’de ekim 2021’deki sentetik veri üreten şirketler görülmektedir.



Şekil 2.1. Ekim 2021'deki sentetik veri şirketi ekosistemi [54].

Bu çalışmada, sentetik veri üretiminde Gretel'den faydalanılmıştır. Gretel, her türlü metin veya yapılandırılmış veriden yeni sentetik örnekler öğrenmek ve oluşturmak için Uzun Kısa Süreli Bellek (LSTM) yapay sinir ağını kullanmaktadır [55]. Gretel ile sentetik veri üretiminden sonra, üretilen verilerle ilgili detaylı bir rapor sunulmaktadır. Bu rapor ile üretilen verilerin kalitesi ve kullanılabilirliği hakkında bilgiler vermektedir. Bu tez çalışmasının gerçekleştirildiği süreçte Gretel, Amplify modelini uygulamaya sunduğunu duyurmuştur. Amplify modeli, geliştirerek büyük hacimdeki ve tablo halindeki sentetik verileri daha hızlı oluşturmaktadır [56].

Şekil 2.2'de Gretel'in sentetik veri üretiminde kullandığı diyagram görülmektedir.



Şekil 2.2. Gretel'in sentetik veri üretmede kullandığı diyagram [57].

2.3. Makine Öğrenmesi

Makine öğrenmesi ile mevcut veriler kullanılarak gelecek hakkında tahminler üretilir. Bu tahminler kategorik yada sayısal değerler olabilir. Denetimli, yarı denetimli, denetimsiz ve takviyeli olmak üzere dört çeşit makine öğrenimi yöntemi vardır. Denetimli makine öğrenmesi ile bir model seçilerek sistem eğitim verileriyle eğitilir ve test verileriyle eğitilen sistem performansı kontrol edilir. Yarı denetimli makine öğrenmesi, etiketli ve etiketsiz verilerin olduğu durumlarda kullanılır. Denetimsiz makine öğrenesi ile veriler arasındaki benzerliklere göre kümeleme yapılır. Takviyeli makine öğrenimi, kendi başına karar alabilen bir sistemin doğru karar almasını nasıl öğrenebileceği ile ilgilenir [58].

Bu çalışmada gelecek hakkında sayısal tahmin yapılacağı için denetimli makine öğrenmesindeki regresyon yöntemleri kullanılmıştır.

2.4. Kullanılan Regresyon Yöntemleri

Makine öğrenmesi ile mevcut verilere uygulanan matematiksel ve istatistiksel işlemler ile çıkarım elde edilir ve tahminler üreten sistemler oluşturulur. Makine öğrenmesi ile üretilmek istenen çıktı kategorik ise sınıflandırma kullanılır. Üretilmek

istenen çıktı sayısal ise regresyon kullanılır. Gözlemlenen benzerliklere göre aynı kümelere atama isteniyorsa kümeleme kullanılır [59].

Regresyon analizi, iki veya daha çok sayısal değişkenin arasındaki ilişkiyi ölçmek için kullanılır. Bir değişken ile analiz yapılıyorsa tek değişkenli regresyon olur. Birden fazla değişken kullanılarak analiz yapıldığında ise çok değişkenli regresyon olur. Regresyon analizi sayesinde değişkenler arasındaki ilişkinin varlığı ve bu ilişikinin gücü hakkında bilgi edinilir. Regresyon sayesinde değişkenlerden birinin değeri bilindiğinde diğer değişken hakkında tahmin yapılması sağlanır [60].

Çok çıktılı regresyonda çıktı değerleri, girdi değerleriyle birlikte aynı zamanda birbirlerine de bağlıdır. Doğrusal regresyon, k-en yakın komşu, karar ağacı, rastgele orman algoritmaları çok çıkışlı regresyon problemlerini doğrudan çözmek için kullanılabilir [61][62]. Şekil 2.3'te çok çıktılı regresyon örneği görülmektedir.

Şekil 2.3. Çok çıktılı regresyon örneği [63].

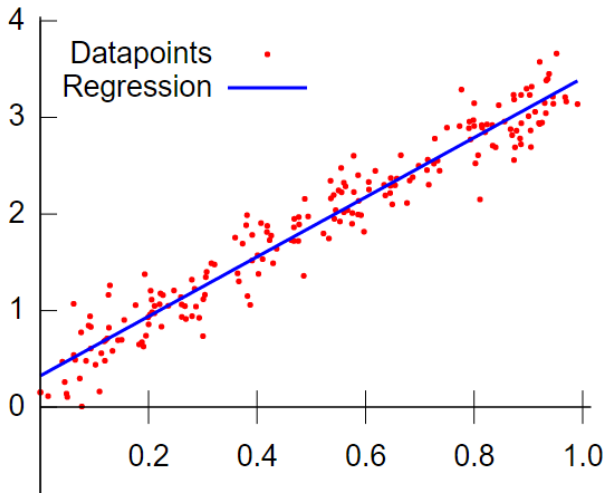
Bu tez çalışmasında öğrencilerin girdiği sınavlar arasındaki ilişkinin varlığı ve gücü hakkında bilgi edinip gelecek hakkında tahminlerde bulunabilmek için çok çıktılı regresyon modellerinden doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmıştır.

2.4.1. Doğrusal Regresyon (Linear Regression)

Doğrusal regresyon, bağımsız bir değişken ile bağımlı bir değişkenin arasındaki ilişki durumunu tahmin etmek için kullanılır. Genellikle regresyon modelleri değişkenler ile tahmin arasındaki bağlantıyı çözmek için kullanılır. Bu modeller bağımlı ve

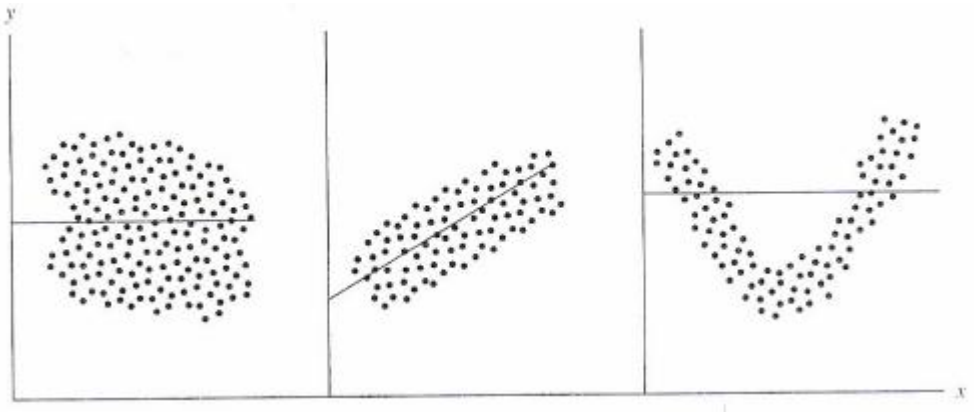
bağımsız değişkenler arasındaki varsayılan ilişki biçimine ve kullanılan bağımsız değişken sayısına göre farklılık gösterir. [62]

Sadece bir değişken değerini kullanarak başka bir değişken değerinin değişmesini açıklamak istediğimizde yeterli olmayacaktır. Birden çok değişkenin etkisi söz konusu olacağından diğer değişkenleri de göz önünde bulundurmak gerekecektir. Dolayısıyla her ne kadar değişkenler arasındaki ilişki tek denklemle açıklansa bile çok sayıda bağımsız değişkene ihtiyaç duyulacaktır. Bu sebeple değişkenler arasında doğrusal ilişki olduğu varsayılarak çoklu doğrusal regresyon modelleri oluşturulur [64]. Şekil 2.4'te bir veri serisi için doğrusal regresyon örneği görülmektedir.



Şekil 2.4. Bir veri serisi için doğrusal regresyon örneği [60].

Şekil 2.5'teki örnek serpilme diyagramlarında, birinci örnekte iki değişken arasında bir ilişkinin olmadığı, ikinci örnekte iki değişken arasında doğrusal ilişki olduğu, üçüncü örnekte ise iki değişken arasında doğrusal olmayan bir ilişki olduğu görülmektedir.



Şekil 2.5. Değişkenler arasındaki doğrusal ilişkiyle ilgili örnek diyagramlar [64].

Doğrusala yakın ilişki gözlendiğinde, basit doğrusal regresyon modelleri denklem 2.1 ile ifade edilir.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.1}$$

Çoklu doğrusal regresyon modelleri için ise denklem 2.2 ile ifade edilir.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} \varepsilon_i \tag{2.2}$$

Denklemlerde Y_i bağımsız değişkenin i numaralı değeridir. $\beta_0, \beta_1, \ldots, \beta_{p-1}$ regresyon katsayılarıdır. β_0 sabit katsayıdır ve basit doğrusal modelde regresyon doğrusunun Y eksenini kestiği noktadır. $\beta_1, \ldots, \beta_{p-1}$ eğim katsayılarıdır ve değerleri sıfır ise bağımsız değişkenlerin bağımlı değişken üzerinde hiç etkisinin olmadığını gösterir. Bu katsayı değerlerinin sıfıra göre farklarının büyümesi ise bu etkinin güçlendiğini gösterir [64].

Python programlama dilindeki doğrusal regresyonda kullanılan parametreler şunlardır (sklearn.linear_model.LinearRegression) [65]:

fit_intercept: Kesmenin hesaplanıp hesaplanmayacağı ayarlanır. Varsayılan değeri True olur.

copy_X: True değeri alırsa X kopyalanır, aksi takdirde üzerine yazılır. Varsayılan değeri True olur.

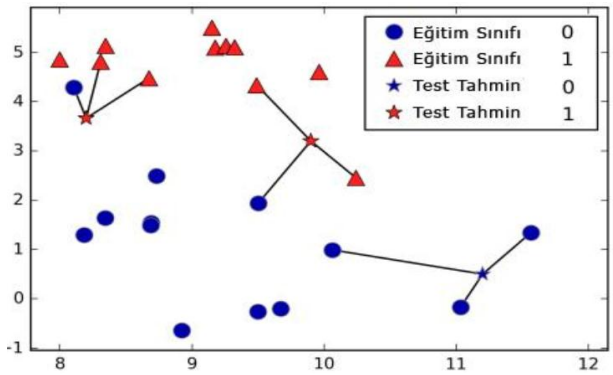
n_jobs: Çok fazla veri olduğunda hesaplamada kullanılacak işlemci sayısı. -1 olursa tüm işlemciyi kullanır. Varsayılan değeri None olur.

positive: True olursa katsayıları pozitif olmaya zorlar. Yoğun dizilerde kullanılır. Varsayılan değeri False olur.

2.4.2. K-En Yakın Komşu (K-Nearest Neighbors)

KNN algoritması, mevcut veriler arasında bir korelasyon olduğunu varsayar ve yeni veriyi mevcut verilerle daha benzer olan kategoriye dahil eder. KNN algoritması, mevcut tüm bilgileri depolar ve benzerliğe bağlı olarak yeni bir veri noktasını sınıflandırır. Bu da yeni veriler ortaya çıktıkça uygun şekilde gruplandırılmasını sağlar. [62]

KNN regresyon mantığı ise Şekil 2.6'daki gösterilen yapıya çok benzemektedir. Aradaki tek fark sayılarla işlem yapılmasıdır. Veri seti için regresyon hesaplanır ve ardından seçilen sayı ile parametre alınarak komşuların sonuçları kontrol edilir. Sonuçların ortalaması alınarak tahmini bir sonuç elde edilir [66].



Şekil 2.6. k=3 için k-en yakın komşu algoritması örneği [37].

Python programlama dilindeki KNN regresyonda kullanılan parametreler şunlardır (sklearn.neighbors.KNeighborsRegressor) [67]:

n_neighbors: Kullanılacak komşu sayısı. Varsayılan değeri 5'tir.

weights: Kullanılacak ağırlık fonksiyonu. Uniform, distance, [callable] seçenekleri vardır. Varsayılan değeri Uniform olur.

algorithm: En yakın komşuları hesaplarken kullanılacak algoritma. Ball_tree, kd_tree, brute, auto seçenekleri vardır. Varsayılan değeri Auto olur.

leaf_size: Ball_tree veya kd_tree'deki yaprak sayısıdır. Varsayılan değeri 30 olur.

p: Minhowski metriği için belirlenen güç parametresidir. Varsayılan değeri 2 olur.

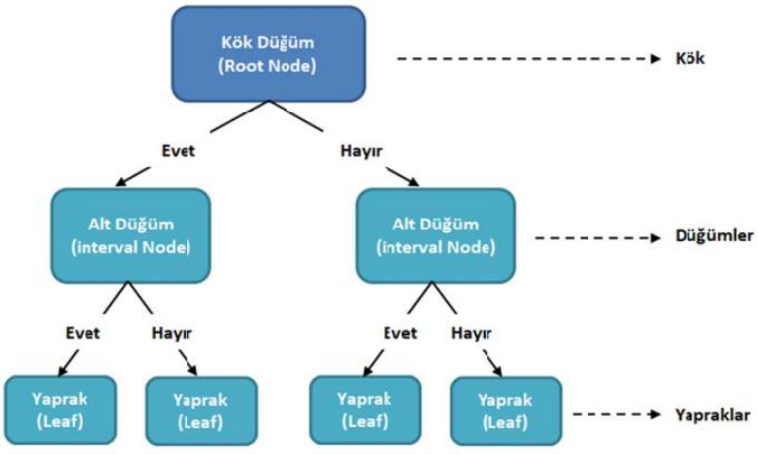
metric: Mesafe hesaplanırken kullanılan metrik değeridir. Varsayılan değeri Minkowski olur.

metric_params: Metrik için ek anahtar sözcük bağımsız değişkenleri değeridir. Varsayılan değeri None olur.

n_jobs: Komşu aramasında kullanılacak işlemci sayısı. -1 olması durumunda tüm işlemciler kullanılır. Varsayılan değeri None olur.

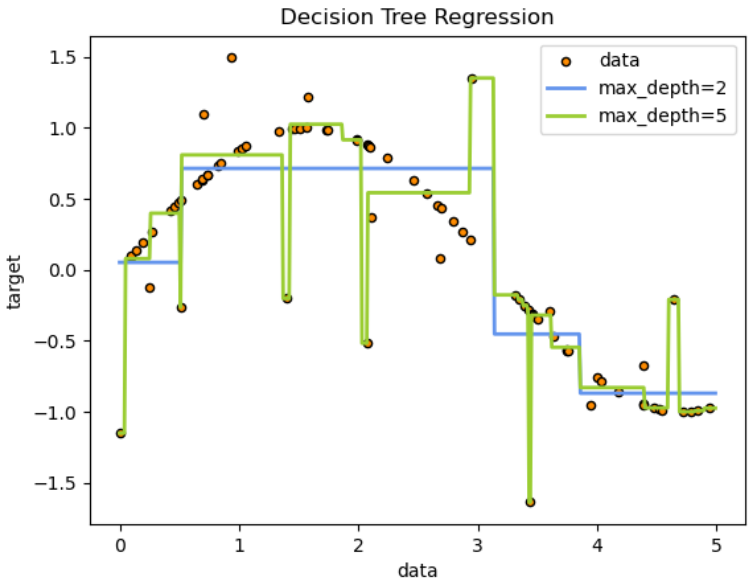
2.4.3. Karar Ağacı (Decision Tree)

Karar ağacı algoritması, bir veri kümesinin daha küçük alt kümelere bölünmesiyle oluşur. Verilerin özelliğine göre ayrılan her biri verileri temsil eden dal bulunan düğümler ve yapraklar bulunan bir ağaç yapısı oluşur. Yaprak, uç noktadaki hesaplanmış sayısal bir kararı temsil etmektedir. En yüksek karar düğümüne kök düğüm denir [62]. Şekil 2.7'de örnek karar ağacı görülmektedir.



Şekil 2.7. Karar ağacı yapısı örneği [32].

Karar ağacı regresyonu, bir nesnenin özelliklerini gözlemleyerek anlamlı ve sürekli sonuç üretmek ve gelecekteki verileri tahmin etmek için bir ağaç yapısında modeli eğitir [68]. Şekil 2.8'de ağacın maksimum derinliği yüksek ayarlandığında, eğitim verilerinin çok ince ayrıntılarını yani gürültüden öğrendiği ve fazla uyduğu görülebilir [70]. Karar ağacında aşırı uyum (overfitting) yada öğrenememe (underfitting) sorunları ile karşılaşılabilir. Bu gibi durumlarda hiperparametre değerleri ile iyileştirme sağlanır. Hiperparametre değerlerine örnek olarak derinlik değeri, düğümün bölünmeden önce gereken örnek sayısı, maksimum yaprak sayısı, bir yapraktaki en az örnek sayısı değerleri girilerek çıktılar kontrol edilir [69].



Şekil 2.8. Karar ağacı, ek gürültülü gözlem ile bir sinüs eğrisine uyma örneği [70].

Python programlama dilindeki karar ağacı regresyonda kullanılan parametreler şunlardır (sklearn.tree.DecisionTreeRegressor) [71]:

splitter: Düğüm bölünürken kullanılacak stratejidir. Best, random seçenekleri vardır. Varsayılan değeri Best olur.

max_depth: Ağacın derinlik sayısıdır. Varsayılan değeri None olur.

min_samples_split: Bir düğümü bölmek için gereken en az örnek sayısıdır. Varsayılan değeri 2 olur.

min_samples_leaf: Bir yaprak düğümünde olması gereken en az örnek sayısıdır. Varsayılan değeri 1 olur.

min_weight_fraction_leaf: Bir yaprak düğümde olması gereken ağırlık toplamının en az ağırlık kesridir. Varsayılan değeri 0.0 olur.

random_state: Tahmin yapılırken rastgele seçim durumu kontrol edilir. Bir değer verildiğinde her çıkan sonuç daha tutarlı olur. Varsayılan değeri None olur.

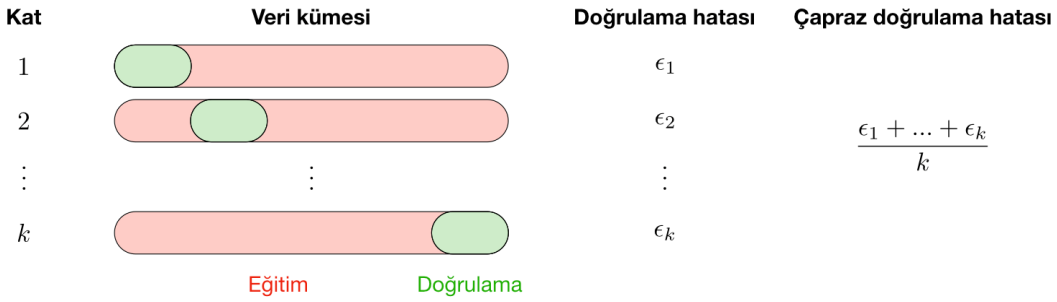
max_leaf_nodes: Kullanılabilecek yaprak düğümü sayısıdır. Varsayılan değeri None olur.

min_impurity_decrease: Düğüm bölünmesi için örneklere değer sınırı konulur. Varsayılan değeri 0.0 olur.

ccp_alpha: Karmaşıklık budaması için kullanılan karmaşıklık parametresidir. Varsayılan değeri 0.0 olur.

2.5. K Katmanlı Çapraz Doğrulama (K-Fold Cross Validation)

Çapraz doğrulama, veri kümesi içerisindeki makine öğrenimi yöntemini değerlendirmek için kullanılır. Yöntemdeki k değeri, seçilen veri kümesinin kaç gruba bölüneceğini belirleyen parametre değeridir [72]. Genellikle k değeri 5 yada 10 olarak kullanılır [73]. Şekil 2.9'da k katmanlı çapraz doğrulamanın çalışma yapısı görülmektedir.



Şekil 2.9. K katmanlı çapraz doğrulama çalışma yapısı [73].

K katmanlı çapraz doğrulama yönteminin çalışması şu adımlardan oluşur [72]:

1. Veriler rastgele karıştırılır,

2. Veriler istenilen k değerine göre gruplara bölünür,

3. Her adımda farklı bir adet doğrulama kümesi seçilir ve k-1 adet küme ise eğitim için kullanılır,

4. Belirlenmiş k adet denemeden sonra ortalama hata oranı bulunur,

5. Çıkan sonuçlara göre çapraz doğrulama uygulanan modelin performansı analiz edilmiş olunur.

Python programlama dilindeki K katmanlı çapraz doğrulamada kullanılan parametreler şunlardır (sklearn.model_selection.RepeatedKFold) [74]:

n_splits: Kat sayısı. En az 2 olmalıdır. Varsayılan değeri 5 olur.

n_repeats: Tekrarlanma sayısı. Varsayılan değeri 10 olur.

random_state: Tekrarlanan çapraz doğrulama örneğinin rastgele seçilme durumunu kontrol eder. Varsayılan değeri None olur.

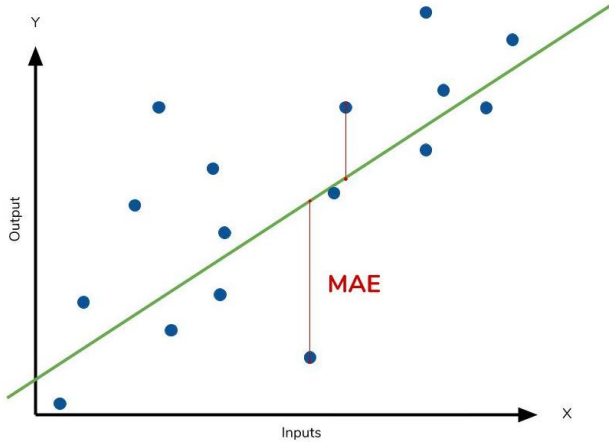
2.6. Kullanılan Performans Ölçümleri

2.6.1. Ortalama Mutlak Hata (Mean Absolute Error - MAE)

Hatanın hedef çıktıya bölümüyle elde edilen değer toplamının veri seti sayısına bölünmesi sonucu elde edilir (denklem 2.3). Denklemdeki T değeri gerçek çıktı, O değeri ağda hesaplanan çıktı, n değeri ise veri seti sayısıdır [75]. Çıkan sonucun sıfıra yakın değer alması regresyon modelinin tahmin ettiği değerler ile gerçek değerlerin birbirine çok yakın olduğunu gösterir [76].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|T-O|}{T} \tag{2.3}$$

Şekil 2.10'daki yeşil çizgi ile oluşturulan modelin tahminleri, mavi noktalar verileri göstermektedir.



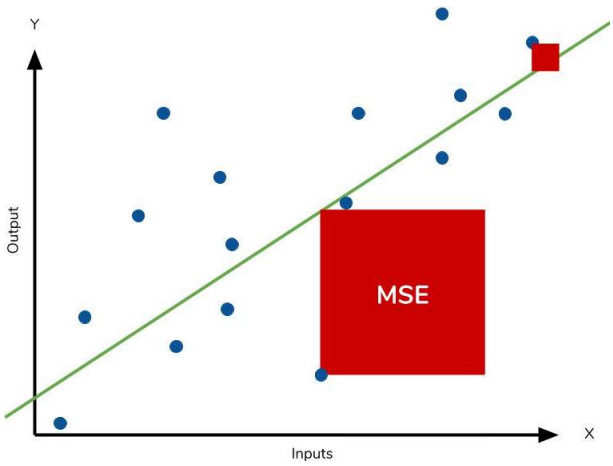
Şekil 2.10. MAE'nin grafiksel açıklaması [77].

2.6.2. Ortalama Kare Hatası (Mean Squared Error - MSE)

Hata karelerinin toplamının eğitim veri seti sayısına bölümü ile elde edilir (denklem 2.4). Denklemdeki T değeri gerçek çıktı, O değeri ağda hesaplanan çıktı, n değeri veri seti sayısıdır [75]. MSE her zaman pozitif değer alır ve çıkan sonucun sıfıra yakın olması regresyon modelinin iyi performans gösterdiğini belirtir [76].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(T - O)^2 \tag{2.4}$$

Şekil 2.11'deki yeşil çizgi oluşturulan modelin tahminleri, mavi noktalar verileri göstermektedir.



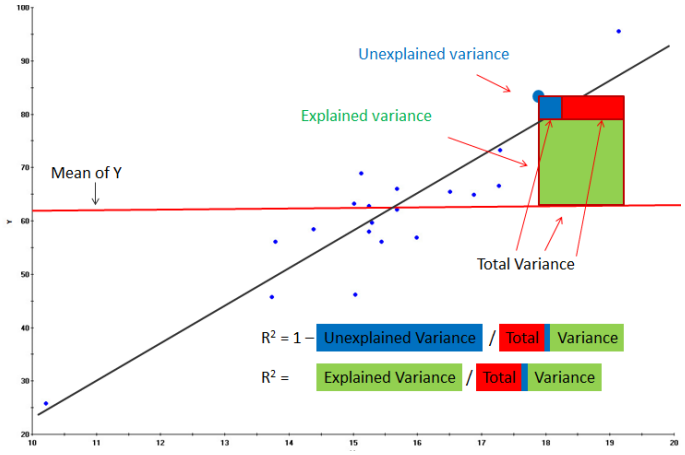
Şekil 2.11. MSE'nin grafiksel açıklaması [77].

2.6.3. Açıklayıcılık Katsayısı (Explanatory Coefficient - R^2)

Bağımsız değişken ve gerçek çıktıların ortalaması kullanılır (denklem 2.5). Denklemdeki T değeri gerçek çıktı, T_{ort} değeri gerçek çıktıların ortalaması, O değeri ağda hesaplanan çıktı, n değeri veri seti sayısı, k değeri bağımsız değişken sayısıdır [75]. R^2 sonucu 0 ile 1 arasında olur. Hesaplanan değerin bire yakın olması, regresyon modelinin performansının iyi olduğunu, sıfıra yakın olması ise modelin performansının kötü olduğunu gösterir. Veri sayısı arttıkça R^2'nin güvenilirliği de artmaktadır [76].

$$R^2 = 1 - \left[1 - \frac{\sum(T-O)^2}{\sum(T-T_{ort})^2}\right]\frac{n-1}{n-k-1} \tag{2.5}$$

Şekil 2.12'de R^2 hesabının grafiksel açıklaması gösterilmiştir.



Şekil 2.12. R^2'nin grafiksel açıklaması [78].

3. LİSE DÜZEYİNDEKİ ÖĞRENCİ VERİLERİ İLE BİR PİLOT ÇALIŞMA

3.1. Veri Seti

Çalışmadaki amaç, öğrencilerin girdiği sınavdaki kazanım başarıları durumuna göre sonraki gireceği sınavlardaki kazanım başarılarını tahmin edip, bu tahmine göre gelecekteki başarısızlıkları önlemek için önlemler alınmasına yardımcı olmaktır. Bu nedenle öğrencilerin girdiği örnek sınavlardaki notlarına ihtiyaç duyulmuştur.

Veri seti olarak 2021-2022 eğitim öğretim yılı 1. döneminde, İstanbul ilindeki Ataşehir Dr. Nureddin Erk - Perihan Erk Mesleki ve Teknik Anadolu Lisesi, Bilişim Teknolojileri alanındaki 10 ve 11'inci sınıfta okuyan 87 öğrencinin Nesne Tabanlı Programlama dersinde uygulanan 3 sınavın not dağılımları kullanılmıştır. Toplanan verilerin kısıtlı olmasından dolayı toplanan bu veriler ile Gretel sistemi üzerinden 5000 sentetik veri üretilerek kullanılmıştır.

3.2. Veri Ön İşleme

10 ve 11'inci sınıftaki Nesne Tabanlı Programlama dersinin kazanım başlıkları, MEB tarafından hazırlanan ders bilgi formlarına bakılarak elde edilmiştir [79]. Tablo 3.1'de Nesne Tabanlı Programlama dersinin kazanım başlıkları numaralandırılmış şekilde görülmektedir.

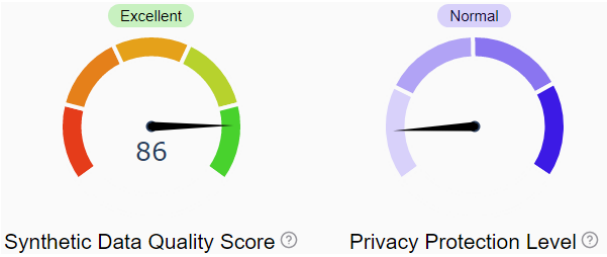
Tablo 3.1. Ders bilgi formlarından elde edilen kazanım başlıkları.

Kazanım Numarası	Kazanım Başlıkları
Kazanım 1	NTP çalışma ortamını kullanır.
Kazanım 2	İsim uzaylarını kullanır.
Kazanım 3	Veri türlerini ve değişkenleri kullanır.
Kazanım 4	Aritmetiksel operatörleri kullanır.
Kazanım 5	Şart ifadelerini kullanır.
Kazanım 6	Mantıksal operatörleri kullanır.
Kazanım 7	Döngü yapılarını kullanır.
Kazanım 8	Hata ayıklaması yapar.

Yapılan sınav soruları kazanım başlıklarına göre gruplanmıştır. Gruplama sonucunda 1. sınavda 1, 3, 4, 5 numaralı kazanımlar, 2. sınavda 1, 3, 4, 5 numaralı kazanımlar ve 3. sınavda 1, 3, 5, 7 numaralı kazanımlarla ilgili sorular olduğu görülmüştür. Excel programı kullanılarak her sorudan alınan puanlar tablo haline getirilmiştir. Daha sonra sorular kazanım başlıkları eşleştirilerek öğrencinin her kazanımdan toplamda kaç puan aldığını gösteren yeni bir tablo oluşturulmuştur. Sınavdaki sorulara bağlı olarak her kazanımdan farklı oranda puan olduğu için kazanımlardan alınan puanlar yüzlük sisteme göre yeniden düzenlenmiştir. Yani öğrencinin sınavda her kazanımın yüzde olarak başarı oranları oluşturulan bir tabloda hesaplanmıştır. Böylece çalışmadaki modellerde kullanılacak veri seti elde edilmiştir.

Veri setinde bulunan toplam 87 öğrenciye ait verinin az olması sebebiyle verimli sonuç elde etmeyi zorlaştıracağından, Gretel sistemi ile toplanan mevcut veriler kullanılarak 5000 sentetik veri üretilmiştir. Gretel sistemi mevcut verileri alarak sentetik veri üretmekte ve üretim sonucunda üretilen verilerle ilgili detaylı bir rapor sunmaktadır. Şekil 3.1'de görüldüğü gibi Gretel sistemi ile epochs=10000 ve batch

size=5000 değerleri verilerek 86 sentetik veri kalite puanı ile 5000 sentetik veri üretilmiştir.



Şekil 3.1. Gretel sentetik veri üretim raporundaki veri kalite puanı.

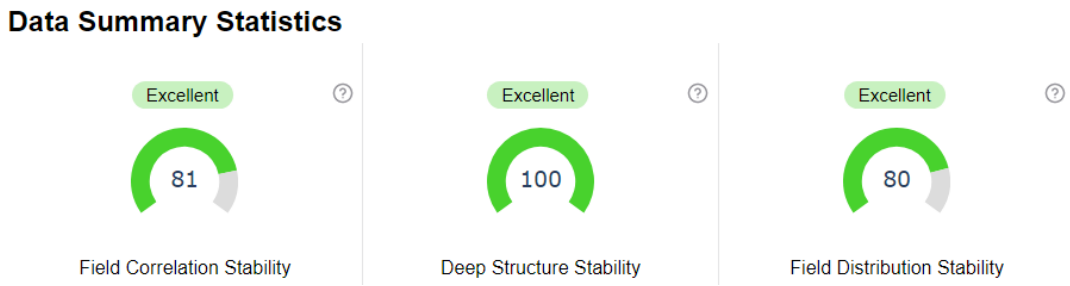
Gretel sistemine göre mükemmel kategorisiyle elde edilen 86 sentetik veri kalite puanı, üretilen sentetik verilerin gerçek veri kümesiyle aynı istatistiksel özellikleri ne kadar iyi koruyabildiğinin bir tahminidir. Yani bu puan, sentetik veri kümesi kullanılmasıyla çıkarılan bilimsel sonuçların gerçek veri kümesi kullanılması durumuyla aynı olup olmayacağına dair bir fayda puanı ya da güven puanı olarak da görülebilir [80]. Gizlilik koruma seviyesi ise sentetik veri üretiminde etkinleştirilen gizlilik mekanizmalarına göre belirlenir.

Sentetik veri üretiminden elde edilen puan mükemmel veya iyi olduğu durumlarda Şekil 3.2'deki listelenen kullanım örneklerinden herhangi biri üretilen sentetik verilerin kullanımı için uygun olmaktadır. Elde edilen puan orta olduğunda ise geçerli kullanım örnekleri daha sınırlı olmaktadır [80]. Öğrenci verilerini kullanılarak üretilen sentetik veriler 86 kalite puanıyla mükemmel kategorisinde görüldüğü için makine öğrenmesinde kullanıma uygun olduğu görülmüştür.

Synthetic Data Use Cases	Excellent	Good	Moderate	Poor	Very Poor
Significant tuning required to improve model	○	○	○	○	⚠
Improve your model using our tips and advice	○	✓	✓	✓	○
Demo environments or mock data	✓	✓	✓	○	○
Pre-production testing environments	✓	✓	✓	○	○
Balance or augment machine learning data sources	✓	✓	○	○	○
Machine learning or statistical analysis	✓	✓	○	○	○

Şekil 3.2. Geretel sentetik veri raporundaki veri kullanım örnekleri.

Şekil 3.3'te Gretel'in sentetik veri üretim raporundaki veri özeti istatistikleri görülmektedir.



Şekil 3.3. Gretel veri özeti istatistikleri.

Alan Korelasyon Kararlılığı: Her alan çiftindeki korelasyonu hesaplamak için önce eğitim verilerinde, daha sonra sentetik verilerde hesaplanır. Bulunan değerler arasındaki mutlak fark hesaplanır ve tüm alanlarda ortalama alınır. Bulunan ortalama değer ne kadar düşükse, alan korelasyon kararlılığı kalite puanı o kadar yüksek olur.

Derin Yapı Stabilitesi: Önce gerçek verilerle, daha sonra sentetik veriler üzerinde hesaplanan bir temel bileşen analizi karşılaştırılır. Bulunan temel bileşenler arasındaki dağılım mesafesi karşılaştırılarak bir kalite puanı oluşturulur. Temel bileşenler ne kadar yakınsa kalite puanı o kadar yüksek olur.

Alan Dağıtım Kararlılığı: Sentetik verilerin alan dağılımlarının gerçek verileri ne kadar iyi temsil ettiğini gösteren bir ölçüdür. Her sayısal veya kategorik alan için Jensen-Shannon mesafesi denilen yöntemi uygulayarak iki dağılımı karşılaştırmak için bir yaklaşım kullanılmaktadır. Bulunan mesafe puanı tüm alanlarda ortalama olarak ne kadar düşükse, alan dağıtım kararlılığı kalite puanı da o kadar yüksek olur.

Üretilen verilerin makine öğrenmesinde kullanılabilmesi için bu puanlamaların yüksek olması önemlidir. Puanlamaların mükemmel kategorisinde olması verilerin makine öğrenmesi çalışmalarında kullanılabilir olduğunu göstermektedir.

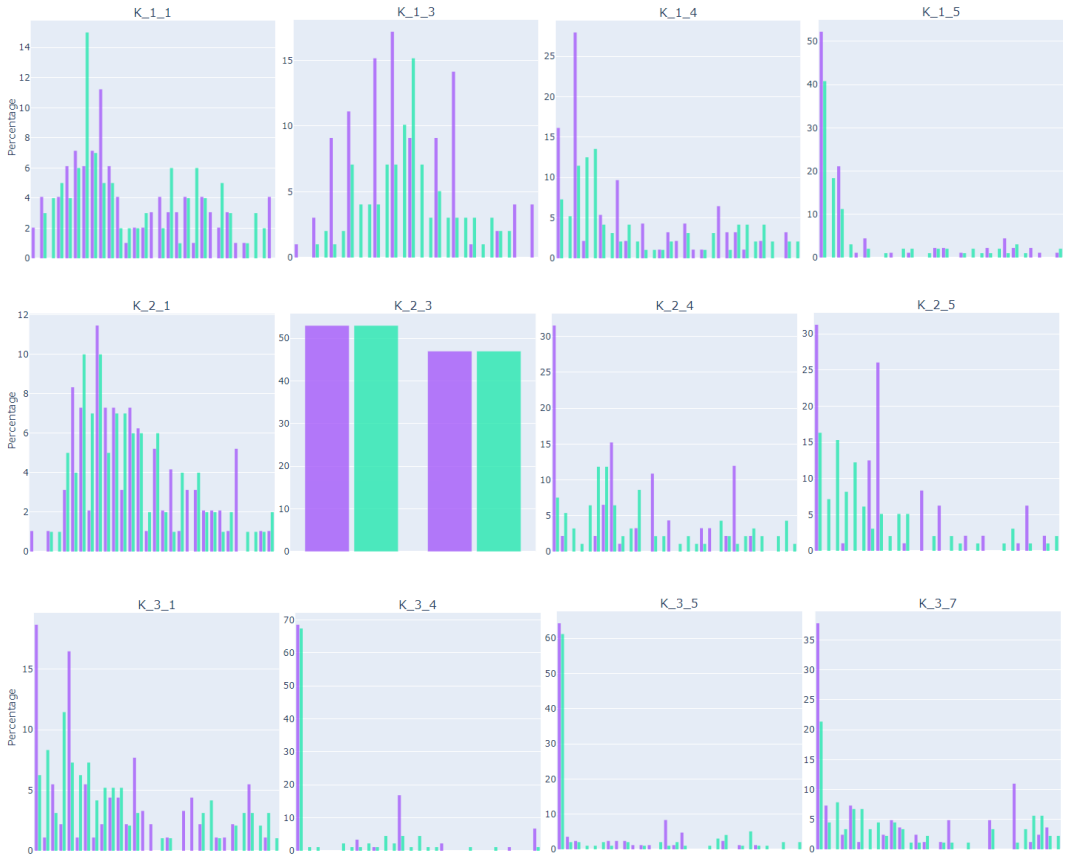
Şekil 3.4'te üretilen sentetik verilerin veri seti başlıklarına göre bilgiler yer almaktadır. Veri seti başlığındaki birinci sayı sınav numarasını, ikinci sayı ise o sınavdaki kazanım numarasını temsil etmektedir. Örneğin K_2_3 başlığı, ikinci sınavdaki üçüncü kazanımı temsil etmektedir. Numaralarına göre kazanım başlıkları Tablo 3.1'de görülmektedir.

Training field overview

Field	Unique	Missing	Ave. Length	Type	Distribution Stability
K_2_3	2	0	1.94	Binary	Excellent
K_3_4	12	0	5.06	Numeric	Moderate
K_1_3	14	0	10.31	Numeric	Good
K_2_5	15	0	7.09	Numeric	Good
K_3_1	29	0	9.48	Numeric	Good
K_3_5	22	0	5.51	Numeric	Good
K_1_5	20	0	6.15	Numeric	Excellent
K_2_1	52	0	10.90	Numeric	Excellent
K_1_4	29	0	8.99	Numeric	Excellent
K_1_1	67	0	10.64	Numeric	Excellent
K_2_4	18	0	7.22	Numeric	Excellent
K_3_7	24	0	5.37	Numeric	Excellent

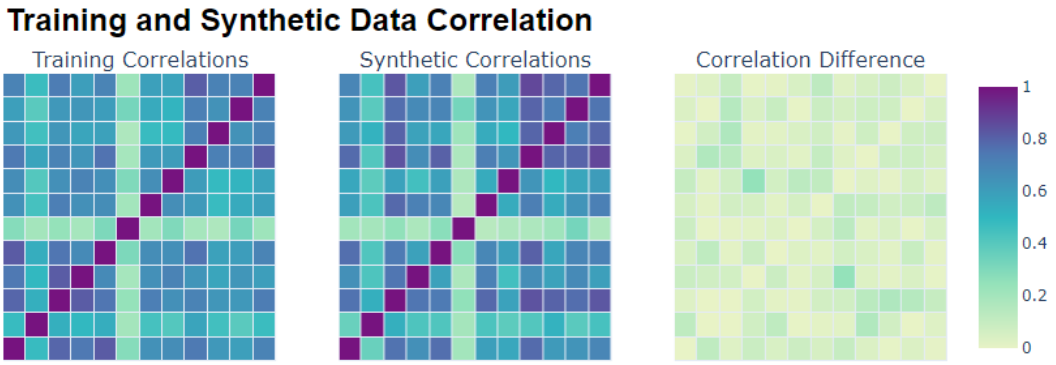
Şekil 3.4. Gretel eğitim alanına genel bakış.

Bir veri kümesi çok sayıda benzersiz alan ya da çok sayıda eksik veri içerdiğinde, bu özellikler modelin verilerin istatistiksel yapısını doğru bir şekilde öğrenmesini engelleyebilmektedir. Şekil 3.5'te her sınavdaki kazanımlara göre gerçek veri ve sentetik veri dağılım yüzdeleri verilmiştir. Şekil 3.5'teki grafiklerde gerçek veriler mor renk ile, sentetik veriler ise yeşil renk ile gösterilmiştir. Grafikler incelendiğinde gerçek veriler ile üretilen sentetik verilerin birbirine yakın dağılım gösterdiği görülmüştür. Burada daha fazla gerçek verinin olması, gerçeğe daha yakın sentetik verilerin elde edilmesini sağlamaktadır.



Şekil 3.5. Gerçek ve sentetik verilere göre Gretel dağılım raporu.

Şekil 3.6'da gerçek verilerin ve sentetik verilerin ayrı ayrı kendi alanları arasındaki korelasyon verilmiştir. Sağda ise gerçek verilerin ve sentetik verilerin alanları arasındaki fark verilmiştir. Her iki veri grubu ayrı değerlendirildiğinde korelasyon oranlarının birbirine benzediği, ikisi arasındaki korelasyon farkının ise az olduğu görülmüştür.



Şekil 3.6. Gerçek ve sentetik veri korelasyonu Gretel rapor grafikleri.

3.3. Uygulanan Yöntem

Çalışmanın uygulama aşamasında makine öğrenmesi ve analizlerindeki kullanışlı yapısından dolayı Python programlama dili kullanılmıştır. Python kodlarını çalıştırmak için ise kullanım kolaylığı sağlayıp herhangi bir kurulum gerektirmediğinden dolayı Google Colaboratory kullanılmıştır.

87 öğrencinin verileri yetersiz olacağı düşünülerek 5000 sentetik veri üretilmiştir. 5000 verinin %80'i yani 4000 veri sistemin eğitimi için, %20'si yani 1000 veri ise test için kullanılmıştır.

Çalışmadaki amaç öğrencinin girdiği sınavda elde ettiği kazanıma göre başarısı sonuçlarına göre sonraki sınavlardaki kazanımlardaki başarısını tahmin etmektir. Öğrencinin girdiği sınavdaki örneğin dört kazanımdaki başarı oranı sisteme girilecek ve sonraki sınav kazanımlarındaki başarı oranları tahmin edilecektir. Dolayısıyla çok çıktılı regresyon kullanılması gerekmektedir. Çok çıktılı regresyonu destekleyen doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmıştır. Modelleri değerlendirmek için k katmanlı çapraz doğrulama uygulanmıştır. Performanslarını ölçmek için ise ortalama mutlak hata (MAE), ortalama kare hatası (MSE), açıklayıcılık katsayısı (R^2) kullanılmıştır.

İlk aşamada 87 veriyle analiz yapılmış ve aynı işlemler sentetik veriyle yapıldığında sentetik verilerle daha iyi sonuçlar alınmıştır. KNN ve karar ağacı algoritmalarındaki parametrelerin iyileştirilmesiyle performansları arttırılmıştır.

4. ARAŞTIRMA BULGULARI

4.1. Doğrusal Regresyon Uygulama Sonuçları

Öğrencilerin girdiği sınavdan, sonraki sınav performanslarını tahmin etmek için gerçek veriler kullanılarak elde edilen performans ölçümleri Tablo 4.1'de gösterilmiştir.

Tablo 4.1. Gerçek veriler ve doğrusal regresyon ile elde edilen değerler.

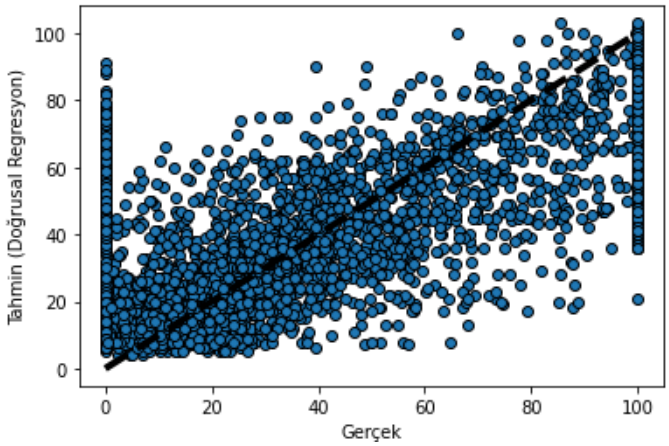
	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	26,477	951	0.471	7,609
1. sınavdan 3. sınav tahmini	21,256	535	0.622	6,324
2. sınavdan 3. sınav tahmini	21,981	653	0.526	8,922

Gerçek verilerden üretilen sentetik veriler kullanılarak elde edilen performans ölçümleri Tablo 4.2'de gösterilmiştir. Sentetik veriler sayesinde elde edilen sonuçların daha iyi olduğu görülmüştür.

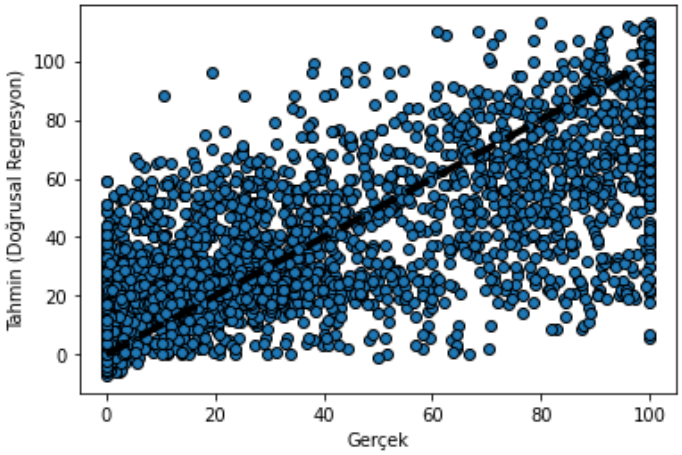
Tablo 4.2. Sentetik veriler ve doğrusal regresyon ile elde edilen değerler.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	20,423	768	0.448	0.578
1. sınavdan 3. sınav tahmini	14,980	451	0.567	1,034
2. sınavdan 3. sınav tahmini	16,320	500	0.525	0.991

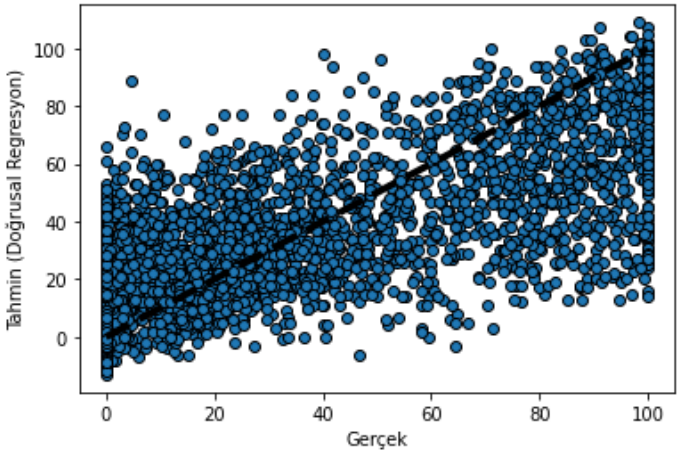
Doğrusal regresyon ile test verileri kullanılarak elde edilen tahminlerle gerçek değerlerin dağılım grafikleri Şekil 4.1, Şekil 4.2, Şekil 4.3'te görülmektedir.



Şekil 4.1. Sentetik veriler, doğrusal regresyon ile 1. sınavdan 2. sınav tahmini



Şekil 4.2. Sentetik veriler, doğrusal regresyon ile 1. sınavdan 3. sınav tahmini



Şekil 4.3. Sentetik veriler, doğrusal regresyon ile 2. sınavdan 3. sınav tahmini

4.2. K-En Yakın Komşu Uygulama Sonuçları

Öğrencilerin girdiği sınavdan, sonraki sınav performanslarını tahmin etmek için gerçek veriler kullanılarak elde edilen performans ölçümleri Tablo 4.3'te gösterilmiştir.

Tablo 4.3. Gerçek veriler ve k-en yakın komşu ile elde edilen değerler.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	23,578	1139	0.389	4,899
1. sınavdan 3. sınav tahmini	19,523	721	0.497	7,626
2. sınavdan 3. sınav tahmini	25,331	749	0.45	9,742

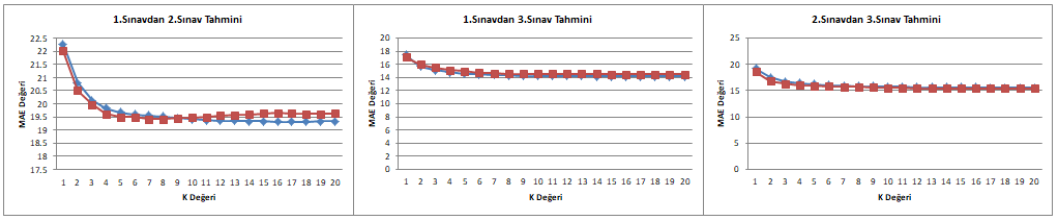
Gerçek verilerden üretilen sentetik veriler kullanılarak elde edilen performans ölçümleri Tablo 4.4'te gösterilmiştir. Sentetik veriler sayesinde elde edilen sonuçların daha iyi olduğu görülmüştür.

Tablo 4.4. Sentetik veriler ve k-en yakın komşu ile elde edilen değerler.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	19,493	859	0.39	0.766
1. sınavdan 3. sınav tahmini	14,927	483	0.539	1,155
2. sınavdan 3. sınav tahmini	15,889	554	0.478	1,179

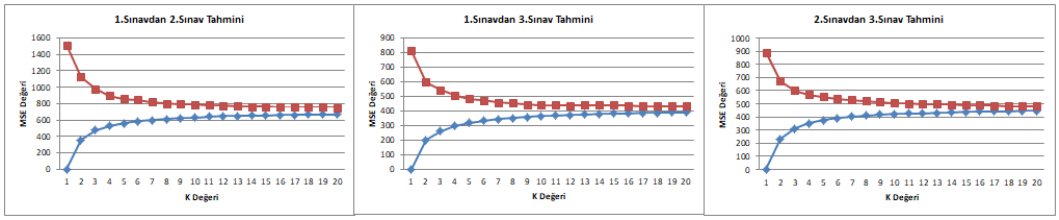
Burada daha iyi tahmin üretebilmek için sonuç üretirken bakılan komşu sayısı parametresi olan k değerine 1 ile 20 arasında değerler verilerek sonuçların değişimi gözlenmiştir. Şekil 4.4, Şekil 4.5, Şekil 4.6 ve Şekil 4.7'de sistemin eğitimden sonra uygulanan k katmanlı çapraz doğrulaması ile elde edilen değerler mavi renk ile gösterilmiştir. Eğitilmiş sisteme test verilerinin uygulanmasıyla elde edilen değerler ise kırmızı ile gösterilmiştir.

Şekil 4.4'te k değeri en düşük değeri aldığında MAE değerinin en yüksek seviyede olduğu ve k değeri arttıkça test verileri kullanılarak elde edilen MAE değerlerinin düştüğü, bir süre sonra dengelendiği görülmüştür. MAE değerinin sıfıra yaklaştığı için hata oranının azaldığı görülmüştür.



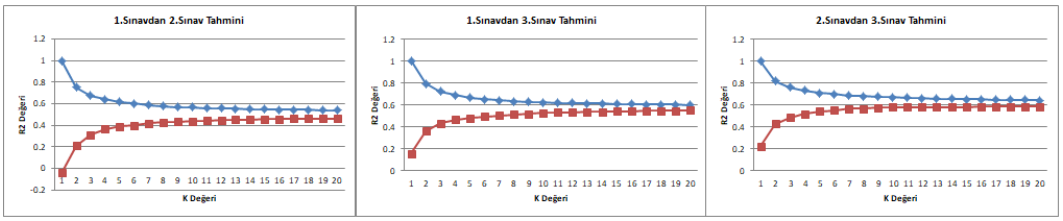
Şekil 4.4. Eğitim ve test verilerine göre k değeri ile MAE ilişkisi.

Şekil 4.5'te k değeri en düşük değeri aldığında eğitilmiş sistemde MSE değerinin en düşük, test verileri uygulandığında ise en yüksek seviyede olduğu görülmüştür. K değeri arttıkça test verileri kullanılarak elde edilen MSE değerlerinin düştüğü, bir süre sonra dengelendiği yani MSE değerinin sıfıra yaklaştığı için hata oranını azaldığı görülmüştür.



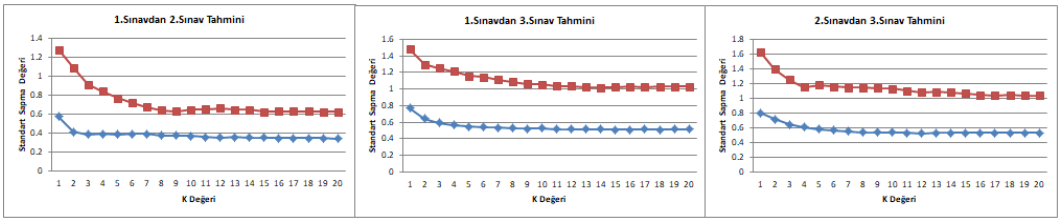
Şekil 4.5. Eğitim ve test verilerine göre k değeri ile MSE ilişkisi.

Şekil 4.6'da k değeri en düşük değeri aldığında eğitilmiş sistemde R^2 değerinin en yüksek, test verileri uygulandığında ise en düşük seviyede olduğu görülmüştür. K değeri arttıkça test verileri kullanılarak elde edilen R^2 sonuçlarının arttığı, bir süre sonra dengelendiği yani 1'e yaklaştığı için tahmin doğruluğunun arttığı görülmüştür.



Şekil 4.6. Eğitim ve test verilerine göre k değeri ile R^2 ilişkisi.

Şekil 4.7'de k değeri en düşük değeri aldığında eğitilmiş sistemde ve test verilerinin uygulanması sonucunda standart sapmanın yüksek olduğu, k değerinin yükselmesiyle standart sapma değerlerinin düştüğü, bir süre sonra dengelendiği görülmüştür. Standart sapma veriler arasındaki yakınlıkları ya da uzaklıkları hakkında fikir verir. Standart sapma değeri küçüldükçe yayılmanın azaldığı anlaşılır.



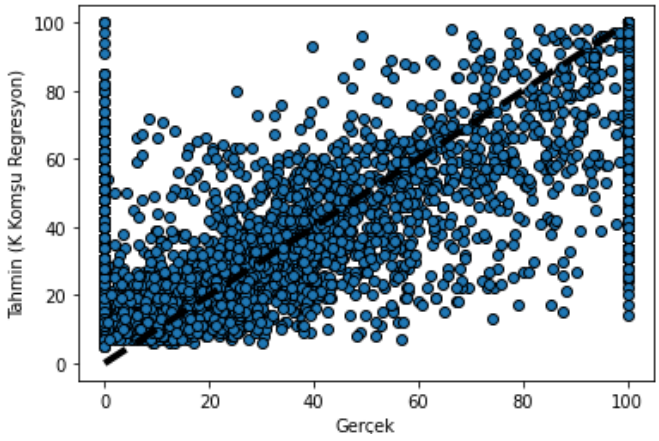
Şekil 4.7. Eğitim ve test verilerine göre k değeri ile standart sapma ilişkisi.

K değeriyle ilgili analizlerden sonra algoritmadan en iyi performansı alabilmek amacıyla yani en iyi k değerini bulabilmek için Python programındaki GridSearchCV yöntemi kullanılmıştır ve 35 olarak bulunmuştur. Tablo 4.5'te, k=35 uygulanarak elde edilen sonuçlar görülmektedir. Elde edilen değerlerin parametre olmadan elde edilen sonuçlara göre daha iyi olduğu görülmüştür.

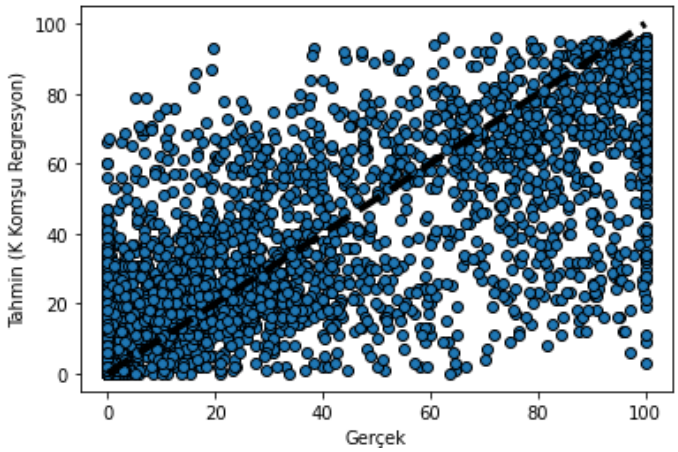
Tablo 4.5. KNN algoritmasında k=35 değeri uygulanarak bulunan sonuçlar.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	19.828	746	0.47	0.603
1. sınavdan 3. sınav tahmini	14.452	427	0.589	1.027
2. sınavdan 3. sınav tahmini	15.587	470	0.557	1.016

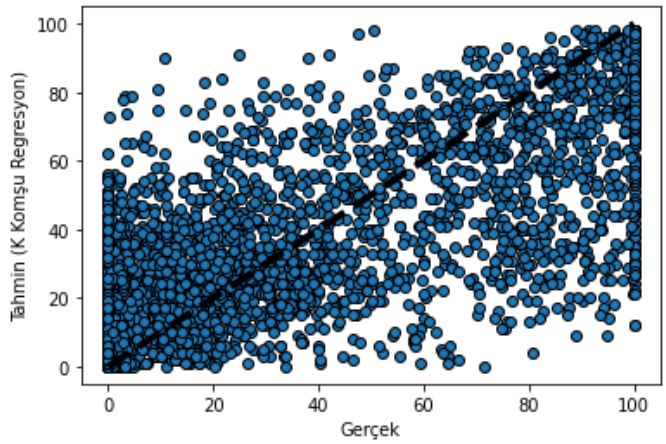
Tablo 4.5'teki elde edilen değerlere göre test verileri kullanılarak elde edilen tahmin dağılım grafikleri Şekil 4.8, Şekil 4.9 ve Şekil 4.10'da görülmektedir.



Şekil 4.8. KNN ile 1. sınavdan 2. sınav performans tahmin grafiği.



Şekil 4.9. KNN ile 1. sınavdan 3. sınav performans tahmin grafiği.



Şekil 4.10. KNN ile 2. sınavdan 3. sınav performans tahmin grafiği.

4.3. Karar Ağacı Uygulama Sonuçları

Karar ağacı algoritması kullanılarak öğrencilerin girdiği sınavdan, sonraki sınav performanslarını tahmin etmek için gerçek veriler kullanılarak elde edilen performans ölçümleri Tablo 4.6'da gösterilmiştir.

Tablo 4.6. Gerçek veriler ve karar ağacı ile elde edilen değerler.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	29.659	1623	0.073	11.529
1. sınavdan 3. sınav tahmini	19.711	943	0.242	8.118
2. sınavdan 3. sınav tahmini	20.906	1354	-0.094	10.048

Gerçek verilerden üretilen sentetik veriler kullanılarak elde edilen performans ölçümleri Tablo 4.7'de gösterilmiştir. Sentetik veriler sayesinde elde edilen sonuçların daha iyi olduğu görülmüştür.

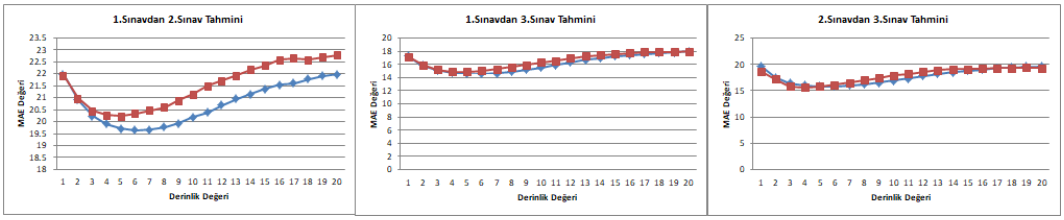
Tablo 4.7. Sentetik veriler ve karar ağacı ile elde edilen değerler.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	22.802	1483	-0.034	1.536
1. sınavdan 3. sınav tahmini	18.003	759	0.27	1.417
2. sınavdan 3. sınav tahmini	19.354	904	0.145	1.621

Burada daha iyi tahmin üretebilmek için derinlik değerine 1 ile 20 arasında değerler verilerek sonuçların değişimi gözlenmiştir. Şekil 4.11, Şekil 4.12, Şekil 4.13 ve Şekil 4.14'te eğitimden sonra uygulanan k katmanlı çapraz doğrulaması ile elde edilen değerler mavi renk ile gösterilmiştir. Test verilerinin uygulanmasıyla elde edilen değerler ise kırmızı ile gösterilmiştir.

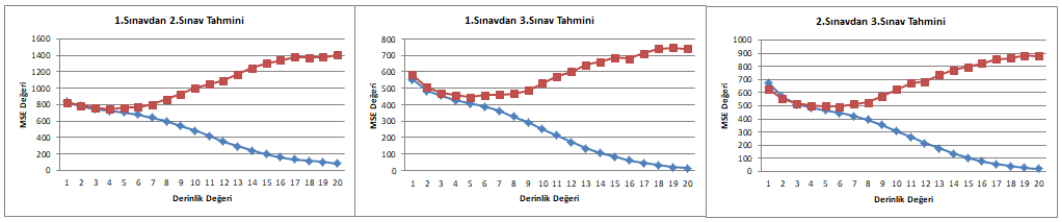
Şekil 4.11'de derinlik değeri en düşük değeri aldığında MAE değerinin en yüksek seviyede olduğu ve k değeri arttıkça test verileri kullanılarak elde edilen MAE

değerlerinin düştüğü, bir noktadan sonra ise yükseldiği görülmüştür. MAE değerinin sıfıra yaklaştığı için hata oranının azaldığı, MAE değerinin yükseldiği yerlerde ise hata oranının da arttığı görülmüştür.



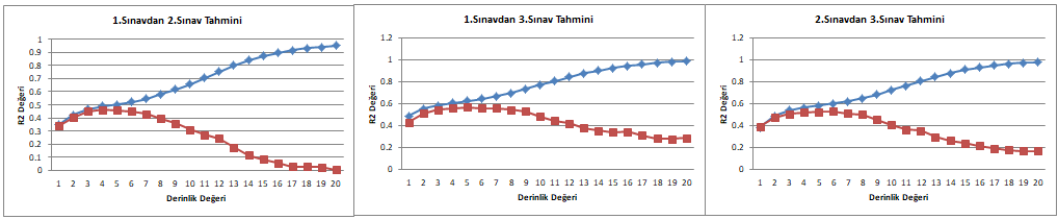
Şekil 4.11. Eğitim ve test verilerine göre derinlik ile MAE ilişkisi.

Şekil 4.12'de derinlik değeri arttıkça eğitilmiş sistemden elde edilen MSE değerlerinin düştüğü, test verileri uygulandığında elde edilen değerlerin ise bir süre düşüp daha sonra yükselişe geçtiği görülmüştür. MSE değerinin düştüğü bölgede hatanın azaldığı, yükseldiği yerlerde ise hata oranının arttığı görülmüştür.



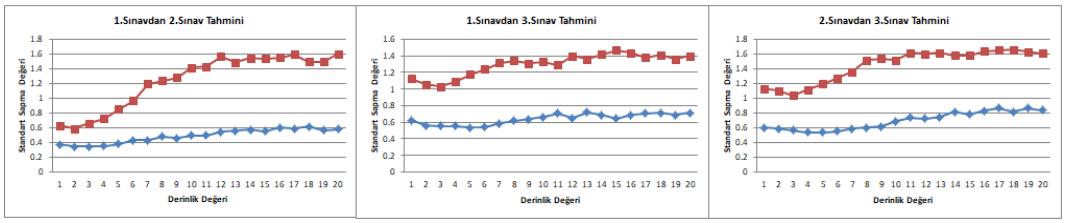
Şekil 4.12. Eğitim ve test verilerine göre derinlik ile MSE ilişkisi.

Şekil 4.13'te derinlik değeri arttıkça eğitilmiş sistemden elde edilen R^2 değerlerinin yükseldiği, test verileri uygulandığında elde edilen değerlerin ise bir süre yükselip daha sonra düşüşe geçtiği görülmüştür. R^2 değerinin yükseldiği bölgede tahmin doğruluğunun arttığı, düştüğü yerlerde ise tahmin doğruluğunun azaldığı görülmüştür.



Şekil 4.13. Eğitim ve test verilerine göre derinlik ile R^2 ilişkisi.

Şekil 4.14'te derinlik değeri arttıkça standart sapma değerlerinin genel olarak arttığı görülmüştür. Standart sapma, veriler arasındaki yakınlıkları ya da uzaklıkları hakkında fikir verir. Standart sapma değeri küçüldükçe yayılmanın azaldığı anlaşılır.



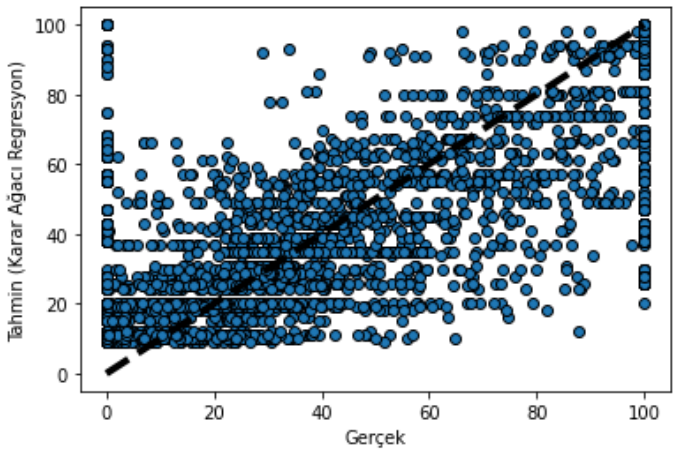
Şekil 4.14. Eğitim ve test verilerine göre derinlik ile standart sapma ilişkisi.

Derinlik değeriyle ilgili analizlerden sonra algoritmadan en iyi performansı alabilmek ve aşırı uyum sorununu önlemek amacıyla en iyi derinlik (max_depth) ve düğümün bölünmeden önce gerekli örnek sayısı (min_samples_split) değerlerini bulabilmek amacıyla Python programındaki GridSearchCV yöntemi kullanılmıştır. En iyi derinlik değeri 5 ve örnek sayısı değeri 18 olarak bulunmuştur. Tablo 4.8'de max_depth=5 ve min_samples_split=18 uygulanarak elde edilen sonuçlar görülmektedir. Elde edilen değerlerin parametre olmadan elde edilen sonuçlardan daha iyi olduğu görülmüştür.

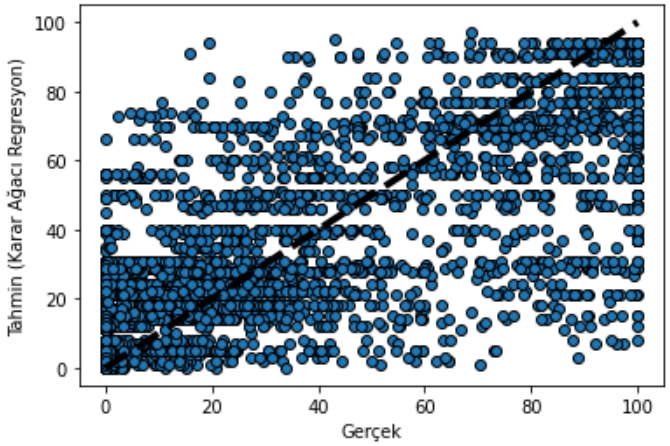
Tablo 4.8. Max_depth=5 ve Min_samples_split=18 ile bulunan sonuçlar.

	MAE	MSE	R^2	Standart Sapma
1. sınavdan 2. sınav tahmini	20.227	762	0.456	0.81
1. sınavdan 3. sınav tahmini	14.713	446	0.57	1.138
2. sınavdan 3. sınav tahmini	15.813	492	0.532	1.171

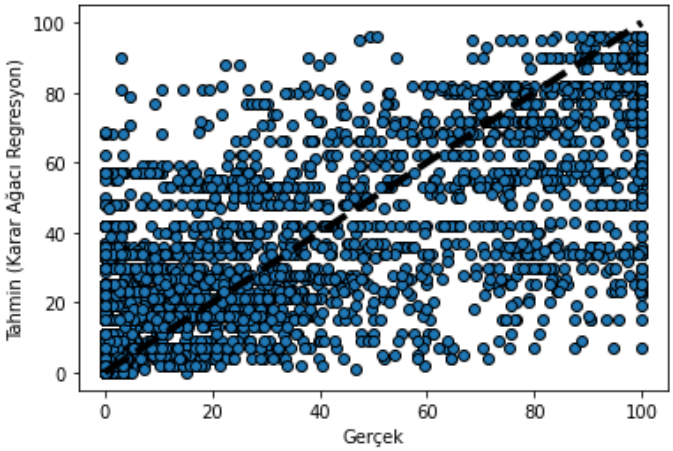
Tablo 4.8'deki elde edilen değerlere göre test verileri kullanılarak elde edilen tahmin dağılım grafikleri Şekil 4.15, Şekil 4.16 ve Şekil 4.17'de görülmektedir.



Şekil 4.15. Karar ağacı ile 1. sınavdan 2. sınav performans tahmin grafiği.



Şekil 4.16. Karar ağacı ile 1. sınavdan 3. sınav performans tahmin grafiği.



Şekil 4.17. Karar ağacı ile 2. sınavdan 3. sınav performans tahmin grafiği.

5. SONUÇ VE ÖNERİLER

Öğrenciler eğitim hayatları boyunca birçok sınava girmekte ve başarı durumlarını girdiği sınavdaki sonuçlara göre değerlendirilmektedir. Sınavdan başarısız olduğu konular tekrar edilerek eksik öğrenmeler telafi edilmeye çalışılmaktadır.

Eğitim alanında makine öğrenmesi yardımıyla başarı tahmini çalışmaları yapılmaktadır. Ancak çalışmalar incelendiğinde genellikle öğrencinin sınav ve demografik yapısıyla ilgili özelliklerine göre geliştirilen sistem öğrencinin başarılı olup olmayacağı konusunda tahmin üretmektedir.

Bu çalışma ile öğrencinin "başarılı" ya da "başarısız" olmasını tahmin etmek yerine öğrencinin girdiği sınavdaki kazanım performansına göre sonraki sınavlarındaki kazanımlardan elde edeceği performans durumları yüzdelik değer olarak tahmin edilmek istenmiştir. Sayısal tahminler üretileceği için regresyon yöntemleri kullanılmıştır. Doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılarak elde edilen analiz sonuçları karşılaştırılmıştır.

Bu çalışmada öncelikle mevcut veriler ile analizler yapılmıştır. Daha sonra mevcut verilerin kısıtlı olmasından dolayı toplanan veriler kullanılarak üretilen sentetik verilerle analiz yapılmıştır. Sentetik verilerle elde edilen değerlerin daha iyi olduğu görülmüştür. Bu durum verilerin kısıtlı olduğu durumlarda, gerçek verilerden üretilen sentetik verilerle çalışmanın daha iyi sonuçlar alınmasında etkili olduğunu göstermiştir.

Sistemin eğitiminden sonra k katmanlı çapraz doğrulama uygulandığında özellikle karar ağacı algoritmasında aşırı uyma sorunu olduğu gözlenmiştir. Aşırı uyma sorunu oluştuğundan karar ağacı algoritmasında R^2 değeri çok yüksek çıkmış ve sanki üretilen tahminlerin tamamen doğru olduğu gibi göstermiştir. Ancak test verileri uygulandığında çok farklı sonuçlar elde edilmiştir. Karar ağacı ve KNN algoritmalarında parametrelerin eklenmesiyle daha anlamlı tahmin başarı oranları elde edilmiştir. Aşırı uyma sorunundan dolayı parametresiz hesaplamalarda özellikle karar ağacı algoritmasında yanıltıcı sonuçlar çıkabileceği görülmüştür.

Regresyon yöntemlerine herhangi bir parametre eklenmeden elde edilen analiz sonuçlarına göre MAE, MSE, R^2 ve standart sapma değerlerine bakıldığında girilen sınavdaki verilere göre sonraki sınavın kazanım performanslarını tahmin etmede doğrusal regresyonun daha iyi olduğu görülmüştür. Ancak KNN algoritmasındaki k parametresi değeri ve karar ağacı algoritmasındaki derinlik ile düğüm bölünmesi için gerekli örnek sayısı parametrelerinin en iyi değerlerinin bulunarak eklenmesiyle bu algoritmalardan elde edilen sonuçlar daha iyileşmiştir. Bunun sonucunda üç algoritmanın performansları birbirine yaklaşmıştır. Kullanılan algoritmalardan son olarak elde edilen sonuçlara bakıldığında, çok az farklar olsa da KNN algoritması daha iyi performans göstermiştir.

Analiz sonucunda bulunan R^2 değerlerinin yükseltilmesi için daha fazla gerçek veri toplanabilir. Gerçek veri sayısının fazla olması sentetik veri üretim kalitesini de arttıracaktır. Ayrıca sınavlarda ölçülen kazanımlar arasındaki puan dağılımı dengelendiğinde daha iyi tahminler üretilebilecektir.

Bu çalışmada elde edilen sonuçlara göre; öğrencilerin girdiği sınav verilerinin makine öğrenmesi ile işlendikten sonra, gireceği sınavlardaki kazanım performansını tahmin etmek amacıyla bir uygulama tasarlanabilir. Sınavlardaki soruları kazanımlarla eşleştirilip, öğrencinin her kazanımdan aldığı puanlar bu uygulamaya girilir. Tasarlanan uygulama, üretilen tahminlere göre öğrenciye neler yapması gerektiği ve hangi konulara yoğunlaşması gerektiğiyle ilgili tavsiyeler sunabilir. Ayrıca ödev, alıştırma, test, video, eğitsel oyun gibi gelecekteki eksik kazanımlarını gidermesine yönelik yönlendirmeler bu uygulama tarafından sunulabilir. Yine tasarlanan uygulama üzerinden öğrencinin çalışma süreçleri takip edilebilir. Hatta oyunlaştırma kullanılarak bu uygulamanın öğrenci açısından daha eğlenceli hale getirilerek ilgisini daha çok çekmesi sağlanabilir.

Bu çalışmanın devamı olarak, tüm derslerin sınavlarındaki kazanım puan dağılımları ve öğrencilerin sınavlardan elde ettiği performans verileri alınarak dersler arasındaki performans bağlantıları ve bir dersteki performansın diğer derslerdeki durumuna etkisi yine makine öğrenmesi yöntemleri ile araştırılabilir. Öğrencilerin gelecekte derslerinde karşılaşabileceği performans düşüklükleri tahmin edilerek, diğer derslerdeki kazanım süreçleri bu tahminlere göre düzenlenebilir. Böylece öğrencilerin gelecekteki performans düşüklükleri erkenden önlenerek eğitim kalitesinin arttırılabileceği düşünülmüştür.

KAYNAKLAR

[1] Aydemir, E. (2019). Ders Geçme Notlarının Veri Madenciliği Yöntemleriyle Tahmin Edilmesi . Avrupa Bilim ve Teknoloji Dergisi , (15) , 70-76 . DOI: 10.31590/ejosat.518899.

[2] Gök, M. (2017). Makine Öğrenmesi Yöntemleri İle Akademik Başarinin Tahmin Edilmesi . Gazi University Journal of Science Part C: Design and Technology , 5 (3) , 139-148.

[3] Güner, N. & Çomak, E. (2011). Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi . Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi , 17 (2) , 87-96.

[4] Abbasoğlu, B. (2020). Ortaokul Öğrencilerinin Akademik Başarılarının Eğitsel Veri Madenciliği Yöntemleri İle Tahmini . Veri Bilimi , 3 (1) , 1-10.

[5] Aghalarova, S. & Bozkurt Keser, S. (2022). Öğrencilerin Akademik Performanslarının Tahmin Edilmesi için AutoML Tekniğinin Uygulanması . El-Cezeri , 9 (2) , 394-412 . DOI: 10.31202/ecjse.946505.

[6] Başer, S. H. , Hökelekli, O. & Adem, K. (2020). Ortaöğretimde Öğrenim Gören Öğrenci Performanslarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi . Bilgisayar Bilimleri ve Teknolojileri Dergisi , 1 (1) , 22-27.

[7] Şengür, D. & Tekin, A. (2014). Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini . Bilişim Teknolojileri Dergisi , 6 (3) , 7-16.

[8] Altun, M. , Kayıkçı, K. & Irmak, S. (2019). Sınıf Öğretmenliği Öğrencilerinin Mezuniyet Notlarının Regresyon Analizi ve Yapay Sinir Ağları Yöntemleriyle Tahmini / Estimation of Graduation Grades of Primary Education Students by Using Regression Analysis and Artificial Neural Networks . e-Uluslararası Eğitim Araştırmaları Dergisi , 10 (3) , 29-43 . DOI: 10.19160/ijer.624839.

[9] Aydemir, E. , Kaysi, F. & Gülseçen, S. (2019). Üniversite Öğrencilerinin Türk Dili Dersi Sınav Sonuçlarının Sınava Hazırlık Düzeylerine Göre Tahminlenmesi . Alphanumeric Journal , 7 (2) , 351-356 . DOI: 10.17093/alphanumeric.583502.

[10] Kanchana, J. D., Amarasinghe, G., Nanayakkara, V., & Perera, A. S. (2021, December). A Data Mining Approach for Early Prediction Of Academic Performance of Students. In 2021 IEEE International Conference on Engineering, Technology & Education (TALE) (pp. 01-08). IEEE..

[11] Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A Machine Learning Algorithm Framework For Predicting Students Performance: A Case Study Of Baccalaureate Students In Morocco. Education And Information Technologies, 24(6), 3577-3589.

[12] Hasib, K. M., Rahman, F., Hasnat, R., & Alam, M. G. R. (2022, January). A Machine Learning And Explainable Ai Approach For Predicting Secondary School Student Performance. In 2022 Ieee 12th Annual Computing And Communication Workshop And Conference (Ccwc) (Pp. 0399-0405). Ieee.

[13] Bujang, S. D. A., Selamat, A., & Krejcar, O. (2021, February). A Predictive Analytics Model For Students Grade Prediction By Supervised Machine Learning. In Iop Conference Series: Materials Science And Engineering (Vol. 1051, No. 1, P. 012005). Iop Publishing.

[14] Ezz, M., & Elshenawy, A. (2020). Adaptive Recommendation System Using Machine Learning Algorithms For Predicting Student's Best Academic Program. Education And Information Technologies, 25(4), 2733-2746.

[15] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An Overview And Comparison Of Supervised Data Mining Techniques For Student Exam Performance Prediction. Computers & Education, 143, 103676.

[16] Trakunphutthirak, R., & Lee, V. C. (2022). Application Of Educational Data Mining Approach For Student Academic Performance Prediction Using Progressive Temporal Data. Journal Of Educational Computing Research, 60(3), 742-776.

[17] Buraimoh, E., Ajoodha, R., & Padayachee, K. (2021, April). Application Of Machine Learning Techniques To The Prediction Of Student Success. In 2021 Ieee International Iot, Electronics And Mechatronics Conference (Iemtronics) (Pp. 1-6). Ieee.

[18] Auwal, A., Ibrahim, A. A., & Bayat, O. (2020). Developing Classifier For The Prediction Of Students' Performance Using Data Mining Classification Techniques. Aurum Journal Of Engineering Systems And Architecture, 4(1), 73-91.

[19] Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions Of Machine Learning Models Towards Student Academic Performance Prediction: A Systematic Review. Applied Sciences, 11(21), 10007.

[20] Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early Segmentation Of Students According To Their Academic Performance: A Predictive Modelling Approach. Decision Support Systems, 115, 36-51.

[21] Yağcı, M. (2022). Educational Data Mining: Prediction Of Students' Academic Performance Using Machine Learning Algorithms. Smart Learning Environments, 9(1), 1-19.

[22] Almasri, A., Celebi, E., & Alkhawaldeh, R. S. (2019). Emt: Ensemble Meta-Based Tree Model For Predicting Student Performance. Scientific Programming, 2019.

[23] Tuba, K. O. C., & Pelin, A. K. I. N. (2022). Estimation Of High School Entrance Examination Success Rates Using Machine Learning And Beta Regression Models. Journal Of Intelligent Systems: Theory And Applications, 5(1), 9-15.

[24] Yavuzarslan, M. & Erol, Ç. (2022). Öğrenme Yönetim Sistemi Log Kayıtlarının Akademik Başarı Tahmininde Kullanılması . Bilişim Teknolojileri Dergisi , 15 (2) , 199-207 . DOI: 10.17671/gazibtd.837884.

[25] Ram, M. S., Srija, V., Bhargav, V., Madhavi, A., & Kumar, G. S. (2021, September). Machine Learning Based Student Academic Performance Prediction. In 2021 Third International Conference On Inventive Research In Computing Applications (Icirca) (Pp. 683-688). Ieee.

[26] Wakelam, E., Jefferies, A., Davey, N., & Sun, Y. (2020). The Potential For Student Performance Prediction In Small Cohorts With Minimal Available Attributes. British Journal Of Educational Technology, 51(2), 347-370.

[27] Haridas, M., Gutjahr, G., Raman, R., Ramaraju, R., & Nedungadi, P. (2020). Predicting School Performance And Early Risk Of Failure From An Intelligent Tutoring System. Education And Information Technologies, 25(5), 3995-4013.

[28] Sixhaxa, K., Jadhav, A., & Ajoodha, R. (2022, January). Predicting Students Performance In Exams Using Machine Learning Techniques. In 2022 12th International Conference On Cloud Computing, Data Science & Engineering (Confluence) (Pp. 635-640). Ieee.

[29] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student Performance Prediction Model Based On Supervised Machine Learning Algorithms. In Iop Conference Series: Materials Science And Engineering (Vol. 928, No. 3, P. 032019). Iop Publishing.

[30] A. Öztürk, "Açık Ve Uzaktan Öğrenmede Öğrenenlerin Davranış Örüntülerinin Ve Profillerinin Modellenmesi, Akademik Performanslarının Tahmin Edilmesi Ve Performans Değerlendirme Panelinin Etkilerinin İncelenmesi", Anadolu Üniversitesi, Sosyal Bilimler Enstitüsü, Uzaktan Eğitim Anabilim Dalı, 2022.

[31] A. Selvi, "Bilecik İlinde İlköğretimden Liseye Geçiş Sınavlarında Makine Öğrenmesi Yöntemleri İle Öğrenci Başarısının Tahmini", Bilecik Şeyh Edebali Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2020.

[32] M. Sulak, "Yapay Zeka Teknikleri İle Açık Öğretim Lisesi Öğrencilerinin Mezuniyet Tahmini", Karabük Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2021.

[33] Ş. Özdemir, "Eğitimde Veri Madenciliği Ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama", İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı, 2016.

[34] S. Can, "Lise Öğrencilerinin Üniversiteye Giriş Başarılarının Eğitsel Veri Madenciliği İle Tahmin Edilmesi", Beykent Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Bilim Dalı, 2021.

[35] C. Can, "Eğitim Alanında Makine Öğrenimi Sınıflandırma Algoritmalarının İncelenmesi", Akdeniz Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme Ve Değerlendirme Programı, 2021.

[36] K. B. Olgun, "Ters Yüz Sınıflardaki Video İzleme Davranışları İncelenerek Veri Madenciliği İle Başarının Tahmin Edilmesi", İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı, 2021.

[37] S. Şahin, "Makine Öğrenmesi Yöntemleri İle Ortaokul Öğrenci Başarılarının Tespiti Ve Bir Uygulama", İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Enformatik Anabilim Dalı, 2021.

[38] S. Özarslan, "Öğrenci Performansının Veri Madenciliği İle Belirlenmesi", Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, 2014.

[39] O. Kayhan, "Uzaktan Eğitim Öğrencilerinin Mezuniyet Durumlarının Veri Madenciliği Yöntemleri İle Tahmini: Amasya Üniversitesi Örneği", Amasya Üniversitesi, Fen Bilimleri Enstitüsü, Teknoloji Ve İnovasyon Yönetimi Anabilim Dalı, 2019.

[40] F. Altınsoy, "Uzaktan Eğitim Öğrencilerinin Başarılarının Yapay Zeka Teknikleri İle Tahmini", Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2019.

[41] H. Göker, "Üniversite Giriş Sınavında Öğrencilerin Başarılarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi", Gazi Üniversitesi, Bilişim Enstitüsü, Bilgisayar Eğitimi, 2012.

[42] S. Yurdakul, "Veri Madenciliği İle Lise Öğrenci Performanslarının Değerlendirilmesi", Kırıkkale Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2015.

[43] B. Aydemir, "Veri Madenciliği Yöntemleri Kullanarak Meslek Yüksek Okulu Öğrencilerinin Akademik Başarım Tahmini", Pamukkale Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı, 2017.

[44] S. Deshmukh, "A Machıne Learnıng Approach To Predıct Retentıon Of Computer Scıence Students At Unıversıty Of Nevada, Las Vegas", University Of Nevada, Las Vegas, Howard R. Hughes College Of Engineering, Department Of Computer Science, 2015.

[45] H. N. Bastem, "Student Academıc Performance Predıctıon Vıa Artıfıcıal Intellıgence Usıng Machıne Learnıng Algorıthms", Çankaya Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Bölümü, 2021.

[46] N. K. Olga, "The Desıgn Of A Predıctıve Model For The Academıc Performance Of Unıversıty Students", Unıversıty Of Johannesburg, Faculty Of Management, It Management, 2016.

[47] A. Bah, "Comparıson Of Predıctıon Algorıthms For Student Performance Predıctıon", The Graduate Schoool Of Natural And Applied Sciences, The Department Of Software Engineering, 2018.

[48] Savaş, S. , Topaloğlu, N. & Yılmaz, M. (2012). Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri . İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi , 11 (21) , 1-23.

[49] Shearer, C. (2000). The Crisp-Dm Model: The New Blueprint For Data Mining. Journal Of Data Warehousing, 5(4), 13-22.

[50] Akpınar, H. (2000). Veri Tabanlarında Bilgi Keşfi Ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 29(1), 1-22.

[51] Aruğaslan, E. & Çivril, H. (2021). Türkiye'de eğitim alanında yapılan veri madenciliği ve yapay zeka çalışmaları . Uluslararası Teknolojik Bilimler Dergisi , 13 (2) , 81-89.

[52] A. Deveci ve M. F. Esen , "Medikal Sentetik Veri Üretimiyle Veri Dengelemesi", İstatistik ve Uygulamalı Bilimler Dergisi, sayı. 5, ss. 17-27, Haz. 2022, doi:10.52693/jsas.1105599.

[53] https://gretel.ai/blog/what-is-synthetic-data, Erişim Tarihi: 10.09.2022.

[54] https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42, Erişim Tarihi:10.09.2022.

[55] https://gretel.ai/gretel-synthetics-faqs/how-does-gretel-synthetics-create-artificial-data, Erişim Tarihi:10.09.2022.

[56] https://gretel.ai/blog/introducing-gretel-amplify, Erişim Tarihi: 24.09.2022

[57] https://gretel.ai/gretel-synthetics-faqs/is-there-an-architecture-diagram, Erişim Tarihi:10.09.2022.

[58] S. Ercan, "Destek Vektör Makinelerini Kullanan Teknik Seçmeli Öneri Sistemi", Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Ve Bilişim Mühendisliği, 2020.

[59] https://erdincuzun.com/makine_ogrenmesi/makine-ogrenmesi-metotlari, Erişim Tarihi: 13.09.2022.

[60] https://tr.wikipedia.org/wiki/regresyon_analizi, Erişim Tarihi: 11.09.2022.

[61] https://machinelearningmastery.com/multi-output-regression-models-with-python, Erişim Tarihi: 11.09.2022.

[62] Kumari, S., Siwach, V., Singh, Y., Barak, D., & Jain, R. (2022). A Machine Learning Centered Approach For Uncovering Excavators' Last Known Location Using Bluetooth And Underground Wsn. Wireless Communications And Mobile Computing, 2022.

[63] He, D., Kuhn, D., & Parida, L. (2016). Novel Applications Of Multitask Learning And Multiple Output Regression To Multiple Genetic Trait Prediction. Bioinformatics, 32(12), I37-I43.

[64] S. Mardikyan, "İlişki Analizinde Varsayımlardan Sapmaların Belirlenmesi Ve Çözümlenmesine Yönelik Bir Bilgisayar Programı Geliştirilmesi", İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İşletme Anabilim Dalı, Sayısal Yöntemler Bilim Dalı, 2005.

[65] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.linearregression.html, Erişim Tarihi: 25.09.2022.

[66] https://towardsdatascience.com/knn-regression-model-in-python-9868f21c9fa2, Erişim Tarihi: 13.09.2022.

[67] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.kneighborsregressor.html, Erişim Tarihi: 25.09.2022.

[68] https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn, Erişim Tarihi: 13.09.2022.

[69] https://medium.com/deep-learning-turkiye/karar-ağaçları-makine-öğrenmesi-serisi-3-a03f3ff00ba5, Erişim Tarihi: 25.09.2022

[70] https://scikit-learn.org/stable/auto_examples/tree/ plot_tree_regression.html, Erişim Tarihi: 13.09.2022.

[71] https://scikit-learn.org/stable/modules/generated/sklearn.tree.decisiontreeregressor.html, Erişim Tarihi: 25.09.2022

[72] T. Özlen, "Servikal Kanserlerin Teşhisinde Kullanılan Makine Öğrenmesi Algoritmalarının Karşılaştırmalı Analizi", İstanbul Aydın Üniversitesi, Lisansüstü Eğitim Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı, 2022.

[73] https://stanford.edu/~shervine/l/tr/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks, Erişim Tarihi:12.09.2022.

[74] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.repeatedkfold.html, Erişim Tarihi: 25.09.2022.

[75] M. F. Adak, "Elektronik Burun Verilerinin Yapay Zeka Tabanlı Algoritmalarla Sınıflandırılması", Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Ve Bilişim Mühendisliği, 2016.

[76] Aydınoğlu, A. Ç. , Bovkır, R. & Çölkesen, İ. (2023). Toplu taşınmaz değerlemede makine öğrenme algoritmalarının kullanımı ve konumsal/konumsal olmayan özniteliklerin tahmin doğruluğuna etkilerinin karşılaştırılması . Jeodezi ve Jeoinformasyon Dergisi , 10 (1) , 63-83 . DOI: 10.9733/JGG.2023R0005.T.

[77] https://www.dataquest.io/blog/understanding-regression-error-metrics, Erişim Tarihi: 13.09.2022.

[78] https://www.rapidinsight.com/blog/brushing-r-squared, Erişim Tarihi: 13.09.2022.

[79] http://meslek.eba.gov.tr/?p=ders-bilgi-formu&tur=mtal, Erişim Tarihi: 20.02.2022

[80] https://gretel.ai/blog/how-accurate-is-my-synthetic-data, Erişim Tarihi: 17.09.2022

ÖZGEÇMİŞ

Ad-Soyad : Ömer DURALİOĞLU

ÖĞRENİM DURUMU:

- **Ön Lisans** **:** 2002, Sakarya Üniv., Hendek MYO,
  Bilgisayar Programcılığı
- **Lisans** **:** 2008, Marmara Üniv., Teknik Eğitim Fak.,
  Bilgisayar ve Kontrol Öğretmenliği
- **Lisans** **:** 2011, Anadolu Üniv., İktisat Fak.,
  Çalışma Ekonomisi ve Endüstri İlişkileri
- **Lisans** **:** 2017, İstanbul Teknik Üniv., Bilgisayar ve Bilişim Fak.,
  Bilgisayar Mühendisliği
- **Yükseklisans** **:** Devam ediyor, Sakarya Üniv., Fen Bilimleri Ens.,
  Bilgisayar ve Bilişim Mühendisliği

MESLEKİ DENEYİM:

- 2012 yılından itibaren Milli Eğitim Bakanlığı'nda öğretmen olarak çalışmaktadır.