



**T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

**MAKİNE ÖĞRENMESİ İLE ÖĞRENCİ BAŞARI DURUMUNUN TAHMİN
EDİLMESİ**

MÜSLÜME KARAKAYA

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**Şanlıurfa
2025**



**T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

**MAKİNE ÖĞRENMESİ İLE ÖĞRENCİ BAŞARI DURUMUNUN TAHMİN
EDİLMESİ**

MÜSLÜME KARAKAYA

**BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
Tez Danışmanı: Doç. Dr. İBRAHİM BERKAN AYDİLEK**

**Şanlıurfa
2025**

TEŐEKKÖR

Bu alıřmanın her ařamasında bilgi, deneyim ve rehberlięiyle bana yol g steren, akademik ve kiřisel geliřimime b y k katkılar saęlayan saygıdeęer danıřman hocam Do. Dr. İbrahim Berkan Aydilek'e, her zaman yanımda olan ve desteęini esirgemeyen aileme, s reci benim iin daha anlamlı hale getiren kıymetli arkadaşlarıma teőekk r ederim.

İÇİNDEKİLER

ÖZET	i
ABSTRACT	ii
ŞEKİLLER DİZİNİ	iii
ÇİZELGELER DİZİNİ	iv
KISALTMALAR	v
1. GİRİŞ	1
2. ÖNCEKİ ÇALIŞMALAR	5
3. GEREÇ VE YÖNTEM	12
3.1. Veri Seti	13
3.2. Veri Madenciliği	16
3.2.1. Veri Ön İşleme	17
3.3. Çalışmada Kullanılan Makine Öğrenmesi Yöntemleri	19
3.4. Önerilen Yöntemin Kullanılması	20
3.4.1. Veri Ön İşleme Adımlarının Kullanılması	20
4. BULGULAR	31
4.1. Değerlendirme Ölçütleri	31
4.2. Önerilen Yöntemin Web Üzerinde Uygulanması	44
5. TARTIŞMA	47
6. SONUÇLAR	56
7. ÖNERİLER	63
KAYNAKLAR	66
ÖZGEÇMİŞ	71

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ İLE ÖĞRENCİ BAŞARI DURUMUNUN TAHMİN EDİLMESİ

MÜSLÜME KARAKAYA

HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Tez Danışman: Doç. Dr. İBRAHİM BERKAN AYDİLEK

Yıl: 2025, Sayfa : 63

Bu çalışmada, öğrenci başarısının tahmin edilmesi amacıyla Kaggle platformundan beş farklı veri seti kullanılmıştır. Veri seti 1, 6607 gözlem ve 20 değişken; Veri seti 2, 10.000 gözlem ve 15 değişken; Veri seti 3, 1388 gözlem ve 5 değişken; Veri seti 4, 10.000 gözlem ve 6 değişken; Veri seti 5 ise 2392 gözlem ve 15 değişken içermektedir. Veri setleri üzerinde eksik ve aykırı değer analizi gerçekleştirilmiş, gerekli düzeltmeler yapılmıştır. Kategorik değişkenler label encoding yöntemiyle sayısal formata dönüştürülmüş, ardından korelasyon analizi yapılarak düşük korelasyona sahip değişkenler çıkarılmıştır. Ridge, Lineer, Lasso, Random Forest ve Destek Vektör Regresyonu (DVR) algoritmaları kullanılarak 10 kat çapraz doğrulama ile modellerin performansları belirleme katsayısı (R^2), ortalama mutlak hata (MAE) ve kök ortalama kare hata (RMSE) metrikleri üzerinden değerlendirilmiştir. Standartlaştırma işlemi Z-skor yöntemiyle uygulanarak modeller tekrar eğitilmiş, hiperparametre optimizasyonu ise GridSearchCV ile gerçekleştirilmiştir. En iyi sonuçlar veri seti 1’de Random Forest algoritmasıyla (R^2 : 0.9553, MAE: 0.5586, RMSE: 0.7211), veri seti 2’de DVR ile (R^2 : 0.9407, MAE: 3.6525, RMSE: 4.2676), veri seti 3’te Random Forest ile (R^2 : 0.9940, MAE: 0.5378, RMSE: 0.7358), veri seti 4’te DVR ile (R^2 : 0.9999, MAE: 0.0762, RMSE: 0.1858) ve veri seti 5’te Lineer Regresyon ile (R^2 : 1.000, MAE: 0.000, RMSE: 0.000) elde edilmiştir. Son aşamada, öğrenci başarısının tahmin edilmesini sağlayan modelin gerçek veriler üzerinde nasıl çalıştığını gözlemlemek amacıyla bir web uygulaması geliştirilmiştir. Bu uygulama, öğrencilere özel başarı tahminleri sunarak kişiselleştirilmiş eğitim süreçlerinde karar destek mekanizması olarak kullanılabilirliğini göstermiştir. Çalışma sonucunda, veri ön işleme, model seçimi ve optimizasyon adımlarının tahmin doğruluğunu önemli ölçüde artırdığı ve makine öğrenimi modellerinin öğrenci başarısını yüksek doğrulukla tahmin edebildiği ortaya konmuştur.

ANAHTAR KELİMELER: Veri Madenciliği, Makine öğrenimi, Öğrenci Başarısı Tahmini, Kişiselleştirilmiş Eğitim Sistemleri

ABSTRACT

MASTER THESIS

PREDICTION OF STUDENT SUCCESS STATUS WITH MACHINE LEARNING

MÜSLÜME KARAKAYA

HARRAN UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

COMPUTER ENGINEERING DEPARTMENT

Thesis Supervisor: Assoc. Prof. Dr. İBRAHİM BERKAN AYDİLEK

Year: 2025, Page : 63

In this study, five different data sets from the Kaggle platform were used to predict student success. Data set 1 includes 6607 observations and 20 variables; Data set 2 includes 10,000 observations and 15 variables; Data set 3 includes 1388 observations and 5 variables; Data set 4 includes 10,000 observations and 6 variables; Data set 5 includes 2392 observations and 15 variables. Missing and outlier analysis was performed on the data sets, and necessary corrections were made. Categorical variables were converted to numerical format by label encoding method, and then correlation analysis was performed to remove variables with low correlation. The performances of the models were evaluated via coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE) metrics with 10-fold cross-validation using Ridge, Linear, Lasso, Random Forest and Support Vector Regression (SVR) algorithms. The models were re-trained by applying the standardization process with the Z-score method, and hyperparameter optimization was performed with GridSearchCV. The best results were obtained on dataset 1 with Random Forest algorithm (R^2 : 0.9553, MAE: 0.5586, RMSE: 0.7211), on dataset 2 with DVR (R^2 : 0.9407, MAE: 3.6525, RMSE: 4.2676), on dataset 3 with Random Forest (R^2 : 0.9940, MAE: 0.5378, RMSE: 0.7358), on dataset 4 with DVR (R^2 : 0.9999, MAE: 0.0762, RMSE: 0.1858) and on dataset 5 with Linear Regression (R^2 : 1.000, MAE: 0.000, RMSE: 0.000). In the final stage, a web application was developed to observe how the model that allows the prediction of student success works on real data. This application has demonstrated its usability as a decision support mechanism in personalized education processes by providing special success predictions for students. As a result of the study, it was revealed that data preprocessing, model selection and optimization steps significantly increased the prediction accuracy and that machine learning models can predict student success with high accuracy.

KEYWORDS: Data Mining, Machine learning, Student Performance Prediction, Personalized Education Systems

ŞEKİLLER DİZİNİ

Şekil 3.1.	Önerilen Yöntemin Akış Diyagramı	13
Şekil 3.2.	Veri Madenciliği bilgi keşif süreci(Fayyad vd., 1996, s.41),(Irmak, 2009)	17
Şekil 3.3.	Veri seti 1, değişkenler arasındaki ilişkiyi gösteren ısı haritası	23
Şekil 3.4.	Veri seti 2, değişkenler arasındaki ilişkiyi gösteren ısı haritası	25
Şekil 3.5.	Veri seti 3, değişkenler arasındaki ilişkiyi gösteren ısı haritası	26
Şekil 3.6.	Veri seti 4, değişkenler arasındaki ilişkiyi gösteren ısı haritası	27
Şekil 3.7.	Veri seti 5, değişkenler arasındaki ilişkiyi gösteren ısı haritası	29
Şekil 4.1.	Veri seti 1, aykırı değer grafiği	33
Şekil 4.2.	Veri seti 3, aykırı değer grafiği	33
Şekil 4.3.	Kullanıcının Dosya Yükleyebildiği ve Analiz Edebildiği Sayfa	45
Şekil 4.4.	Kullanıcının Dosya Yükleyebildiği ve Analiz Edebildiği Sayfa	46
Şekil 4.5.	Kullanıcının Öğrenci Notunu Tahmin Edebildiği Arayüz	46
Şekil 5.1.	Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları	47
Şekil 5.2.	Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları	48
Şekil 5.3.	Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları	50
Şekil 5.4.	Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları	51
Şekil 5.5.	Önerilen Yönteme Göre Çalışılan Veri setleri ve Sonuçları	52
Şekil 6.1.	Veri seti 3, Orijinal Veri Setinde Bulunan Gerçek Not Dağılımı	59
Şekil 6.2.	Web Uygulaması Tarafından Tahmin Edilen Not Dağılımı	60

ÇİZELGELER DİZİNİ

Çizelge 3.1.	Veri seti 1, öznitelik ve veri tipleri	14
Çizelge 3.2.	Veri seti 2, öznitelik ve veri tipleri	15
Çizelge 3.3.	Veri seti 3, öznitelik ve veri tipleri	15
Çizelge 3.4.	Veri seti 4, öznitelik ve veri tipleri	16
Çizelge 3.5.	Veri seti 5, öznitelik ve veri tipleri	16
Çizelge 3.6.	Veri seti 1, 'Exam_Score' ile diğer özelliklerin korelasyonları	22
Çizelge 3.7.	Veri seti 2, 'Exam_Score(%)' ile diğer özelliklerin korelasyonları	24
Çizelge 3.8.	Veri seti 3, 'Grades' ile diğer özelliklerin korelasyonları	25
Çizelge 3.9.	Veri seti 4, 'Performance Index' ile diğer özelliklerin korelasyonları	26
Çizelge 3.10.	Veri seti 5, 'GPA' ile diğer özelliklerin korelasyonları	28
Çizelge 4.1.	Veri seti 1, 10 Kat Çaprazlama Model Sonuçları	34
Çizelge 4.2.	Veri seti 2, 10 Kat Çaprazlama Model Sonuçları	35
Çizelge 4.3.	Veri seti 3, 10 Kat Çaprazlama Model Sonuçları	35
Çizelge 4.4.	Veri seti 4, 10 Kat Çaprazlama Model Sonuçları	36
Çizelge 4.5.	Veri seti 5, 10 Kat Çaprazlama Model Sonuçları	36
Çizelge 4.6.	Veri seti 1, Standartlaştırma Sonrası Model Sonuçları	37
Çizelge 4.7.	Veri seti 2, Standartlaştırma Sonrası Model Sonuçları	37
Çizelge 4.8.	Veri seti 3, Standartlaştırma Sonrası Model Sonuçları	37
Çizelge 4.9.	Veri seti 4, Standartlaştırma Sonrası Model Sonuçları	38
Çizelge 4.10.	Veri seti 5, Standartlaştırma Sonrası Model Sonuçları	38
Çizelge 4.11.	Veri seti 1, GridSearhCV Sonrası Model Sonuçları	39
Çizelge 4.12.	Veri seti 2, GridSearhCV Sonrası Model Sonuçları	39
Çizelge 4.13.	Veri seti 3, GridSearhCV Sonrası Model Sonuçları	39
Çizelge 4.14.	Veri seti 4, GridSearhCV Sonrası Model Sonuçları	40
Çizelge 4.15.	Veri seti 5, GridSearhCV Sonrası Model Sonuçları	40
Çizelge 4.16.	Veri seti 1, Standartlaştırma ve GridSearhCV Sonrası Model Sonuçları	41
Çizelge 4.17.	Veri seti 2, Standartlaştırma ve GridSearhCV Sonrası Model Sonuçları	41
Çizelge 4.18.	Veri seti 3, Standartlaştırma ve GridSearhCV Sonrası Model Sonuçları	42
Çizelge 4.19.	Veri seti 4, Standartlaştırma ve GridSearhCV Sonrası Model Sonuçları	42
Çizelge 4.20.	Veri seti 5, Standartlaştırma ve GridSearhCV Sonrası Model Sonuçları	42
Çizelge 4.21.	Veri seti 1, Önerilen Yöntemin Ortalama Sonuçları	43
Çizelge 4.22.	Veri seti 2, Önerilen Yöntemin Ortalama Sonuçları	43
Çizelge 4.23.	Veri seti 3, Önerilen Yöntemin Ortalama Sonuçları	44
Çizelge 4.24.	Veri seti 4, Önerilen Yöntemin Ortalama Sonuçları	44
Çizelge 4.25.	Veri seti 5, Önerilen Yöntemin Ortalama Sonuçları	44

KISALTMALAR

DVM : Destek Vektör Makineleri (Support Vector Machines)

R2 : R-Squared (R-Kare)

RF : Rastgele Orman (Random Forest)

1. GİRİŞ

Teknoloji; sağlık, eğitim ve birçok farklı alanda etkili bir şekilde kullanılmakta olup, özellikle eğitim süreçlerinde öğrencilerin akademik başarılarını artırmada önemli bir rol oynamaktadır (Wekerle vd., 2022). Eğitimde teknolojinin kullanımı, öğrenme materyallerine daha kolay erişim sağlamakla kalmayıp, aynı zamanda öğrenciler için kişiselleştirilmiş öğrenme deneyimlerinin oluşturulmasına da katkıda bulunmaktadır. Bilgisayar teknolojilerinin sürekli gelişimi, veri birikim sistemlerinin eğitim dâhil birçok alanda yaygın olarak kullanılmasını sağlamış ve verilerin saklanması hem ekonomik hem de pratik hale getirmiştir (Kalikov, 2006). Ancak, biriken ham verilerin anlamlı bilgilere dönüştürülmesi ve eğitim süreçlerinde etkin bir şekilde kullanılabilmesi, veri madenciliği yöntemleri ile mümkün olmaktadır (Özbay, 2015).

Veri madenciliği, karmaşık verilerin üretildiği ve saklandığı birçok alanda kullanılan bir tekniktir (Erten, 2015). Verilerin işlenerek anlamlı bilgilere dönüştürülmesi, karar alma süreçlerini iyileştirmek, sistemleri optimize etmek ve geleceğe yönelik tahminlerde bulunmak açısından büyük önem taşımaktadır. Veri madenciliği; iş zekası, pazarlama, sağlık, eğitim, finans ve üretim gibi pek çok sektörde etkili bir şekilde kullanılmaktadır. Eğitim alanında kullanılan veri madenciliği ve istatistikler, öğrencilerin öğrenme yöntemlerini belirlemek, bilgi ve başarılarını tahmin etmek ve gerçek hayatta öğrenme yetilerini geliştirmek açısından dönüşüm yaratıcı bir potansiyele sahiptir.

Eğitimde veri madenciliğinin ilerlemesi, sağlık, finans ve ekonomi gibi diğer alanlarla kıyaslandığında önemli bir artış göstermiştir. Öğrenme biçimleri, geçmişten günümüze araştırılan bir konu olmuştur. Veri madenciliği tekniklerini kullanarak öğrencilerin başarılarının kişiselleştirilmesi üzerine yapılan bir çalışmada, her öğrenci için uygun öğrenme yönteminin seçildiği, öğrencilerin okul performanslarına göre uyarlamalar yapıldığı ve böylece derslere daha etkin katılım sağlandığı görülmüştür (Magdin, 2015).

Arthur Samuel, bireylerin bilgisayarlara değil de bilgisayarların kendi kendine öğrenebileceği fikrini savunmuştur (Samuel, 1959). Günümüzde sık sık karşımıza çıkan makine öğrenmesi bu fikirle yerini almıştır. Makine öğrenmesi önceden deneyimlenen bilgi ve tecrübeleri kullanarak yeni bir yaklaşım sergilemektedir (Alpaydın, 2020). Eğitimde kullanılan makine öğrenmesi ise öğrencilerin performanslarını tahmin edebilmektedir. Makine öğrenmesi model ve teknikleri her

birey için öğrenebilme zorluklarını belirlemekte ve destek verici yöntemler önerebilmektedir (Anozie & Junker, 2006). Bu çalışmada, öğrenci başarısına etki eden faktörleri tespit etmek ve öğrenci başarılarını doğru bir şekilde tahmin etmek amacıyla makine öğrenmesi yöntemlerinin kullanılması hedeflenmektedir.

Eğitim sistemleri, günümüzde öğrencilerin demografik bilgilerini ve akademik başarılarını kapsayan geniş bir veri havuzunu elektronik platformlarda depolamaktadır. Bu veriler, genellikle öğrenme yönetim sistemleri (ÖYS) ve öğrenci bilgi sistemleri (ÖBS) gibi dijital kaynaklardan toplanmaktadır. Eğitim alanında hızla büyüyen bu veri hacmi, öğrencilerin akademik başarılarını artırmak ve öğrenme süreçlerini geliştirmek için büyük bir potansiyel sunmaktadır. Yapılan çalışmalar (Shorfuzzaman ve ark., 2019; Viberg ve ark., 2018), bu verilerin analitik yöntemlerle işlenmesi yoluyla, eğitim çıktılarının iyileştirilmesine ve öğrencilere daha etkili, kişiselleştirilmiş öğrenme deneyimleri sunulmasına olanak tanıdığını göstermektedir. Eğitimdeki bu dijitalleşme ve veri odaklı yaklaşım, öğrencilerin ihtiyaçlarını daha iyi anlamaya ve onların potansiyellerini ortaya çıkarmaya yardımcı olabilecek önemli bir araç haline gelmiştir.

Geleneksel eğitim yöntemleri, bireysel farklılıkları göz ardı ederek genellikle sınıfın genel seviyesine göre şekillendirilmektedir. Ancak ideal bir eğitim yaklaşımı, her öğrencinin bireysel özelliklerini dikkate alarak oluşturulan ders içeriklerini temel almalıdır. Bu noktada, yapay zekâ teknolojilerinin eğitimde kullanımı, öğrencilere daha özelleştirilmiş bir öğrenme deneyimi sunmanın yanı sıra, öğretmenlere de öğrencilerin performanslarını daha ayrıntılı şekilde izleme ve etkili geri bildirim sağlama fırsatı sunmaktadır. Özellikle makine öğrenimi ve veri madenciliği gibi yapay zekâ yöntemleri, öğrenci verilerinin kapsamlı analizini mümkün kılarak eğitim süreçlerinin geliştirilmesine önemli katkılar sağlamaktadır (Chen ve ark., 2022). Bu yenilikçi yaklaşımlar, eğitimde bireyselleştirilmiş çözümlere olan ihtiyacı karşılamada kritik bir rol oynamaktadır.

Makine öğrenimi, bir bilgisayarın belirli bir problemi, o problemle ilgili verilere dayanarak modellemesine olanak tanıyan algoritmaların genel adıdır. Bu algoritmalar, mevcut veri setine ve kullanılan yöntemlere göre en iyi performansı hedefleyecek şekilde tasarlanmaktadır. Bu nedenle, çeşitli makine öğrenimi teknikleri geliştirilmiştir. Bu tekniklerden bazıları şunlardır: k-en yakın komşu algoritması, Naive Bayes sınıflandırıcı, karar ağaçları, lojistik regresyon, k-ortalamlar algoritması, destek vektör makineleri ve yapay sinir ağları. Bu yöntemlerden bazıları tahmin ve öngörü üzerine yoğunlaşırken, bazıları veri

kümelemesi veya sınıflandırma işlevlerini gerçekleştirebilmektedir (Atalay ve Çelik, 2017).

Makine öğrenimi, bir sistemin, önceki verilerden ve deneyimlerden yararlanarak kendi kendine öğrenmesini sağlayan bir süreçtir. Başka bir ifadeyle, makine öğrenimi, bilgisayarların doğrudan verilere dayanarak bilgi edinmesini ve bu bilgiler ışığında sorunları çözme yeteneği kazanmasını içermektedir (Ratner, 2000). Bu yöntem, veri odaklı problem çözümünde oldukça etkili bir araçtır.

Eğitim süreçlerinde, öğrencilerin akademik başarılarını etkileyen faktörlerin doğru bir şekilde analiz edilmesi, kişiselleştirilmiş öğrenme yaklaşımlarının geliştirilmesi açısından büyük bir önem taşımaktadır. Geleneksel eğitim sistemleri, genellikle tüm öğrencilere aynı yöntemlerle yaklaşarak eğitimde tek tip bir model benimsemekte ve bireysel farklılıkları yeterince dikkate almamaktadır. Ancak, öğrencilerin öğrenme biçimleri, çalışma alışkanlıkları, sosyo-ekonomik durumları ve akademik geçmişleri birbirinden farklı olduğundan, her öğrenci için en uygun öğrenme stratejilerinin belirlenmesi gerekmektedir. Bu bağlamda, eğitimde veri odaklı yaklaşımlar ve yapay zeka teknikleri, öğrencilerin akademik gelişimini daha iyi anlamak ve onların bireysel ihtiyaçlarına yönelik özelleştirilmiş çözümler sunmak için güçlü bir araç olarak değerlendirilmektedir.

Bu çalışmada, öğrenci başarılarının makine öğrenimi algoritmalarıyla tahmin edilmesi amaçlanmıştır. Farklı veri setleri kullanılarak, öğrencilerin bireysel özellikleri, çalışma alışkanlıkları, sosyal etkinliklere katılımları, uyku düzenleri, ders çalışma süreleri ve geçmiş akademik performansları gibi unsurlar detaylı bir şekilde incelenmiş ve bu faktörlerin başarı üzerindeki etkileri analiz edilmiştir. Veri madenciliği teknikleri ve makine öğrenimi algoritmaları kullanılarak en iyi tahmin performansı veren model belirlenmiş ve bu model aracılığıyla öğrencilerin akademik başarılarının önceden tahmin edilmesi sağlanmıştır. Böylece, öğrenci başarısını etkileyen değişkenler arasında güçlü korelasyonlar tespit edilerek, öğretmenlerin ve eğitimcilerin karar alma süreçlerine katkı sağlanmıştır.

Bunun yanı sıra, elde edilen modelin eğitim süreçlerine entegre edilebilmesi amacıyla öğretmenler için kullanıcı dostu bir web arayüzü geliştirilmiştir. Bu sistem, öğretmenlerin öğrencilerin başarı durumlarını önceden görmelerine ve akademik riskleri belirleyerek erken müdahalede bulunmalarına olanak tanımaktadır. Öğrencinin bireysel özellikleri ve öğrenme geçmişi dikkate alınarak geliştirilen bu tahmin sistemi sayesinde, düşük performans gösterme ihtimali yüksek olan

öğrenciler önceden belirlenebilmekte ve böylece zamanında rehberlik hizmetleri sağlanarak akademik başarılarının artırılması hedeflenmektedir. Ayrıca, bu sistem aracılığıyla öğretmenler, öğrenciler için en uygun öğrenme stratejilerini belirleyebilir ve kişiselleştirilmiş eğitim planları oluşturabilir.

Bu araştırmanın temel katkılarından biri, öğrenci başarısını etkileyen çok boyutlu faktörleri analiz ederek eğitimin kişiselleştirilmesine yönelik bilimsel bir yaklaşım sunmasıdır. Öğrencilerin akademik gelişimini desteklemek, bireysel farklılıkları göz önünde bulundurarak eğitim süreçlerini daha verimli hale getirmek ve başarısızlık risklerini en aza indirmek için veri madenciliği ve makine öğrenimi tekniklerinin etkin bir şekilde nasıl kullanılabileceği ortaya konulmuştur. Çalışmanın sonucunda elde edilen bulgular, öğrenci başarısını tahmin etmeye yönelik geliştirilen sistemlerin gelecekte daha kapsamlı ve hassas hale getirilmesine katkı sağlayacak niteliktedir. Bu bağlamda, önerilen modelin ve geliştirilen sistemin, eğitimde daha bilinçli kararların alınmasına yardımcı olacağı ve kişiselleştirilmiş öğrenme modellerine önemli bir referans oluşturacağı öngörülmektedir.

2. ÖNCEKİ ÇALIŞMALAR

Can, Özdil ve ark. (2018), üniversite öğrencilerinin ders başarılarına etki eden faktörleri belirlemek için lojistik regresyon analizinden faydalanmıştır. Araştırmada, bir devlet üniversitesinden alınan 5820 öğrencinin ders değerlendirme anketleri analiz edilmiştir. Bu anket, Likert ölçeğiyle hazırlanan 28 tutum/görüş sorusu ve 5 demografik bilgi sorusundan oluşmaktadır. Veriler, ders tekrar durumu ve memnuniyet gibi bağımlı değişkenlere göre iki ayrı başarı ölçütü çerçevesinde kategorize edilmiştir. Çapraz doğrulama yöntemiyle yapılan analizler sonucunda, doğru sınıflandırma oranı %91,6 olarak hesaplanmıştır. (Can, Özdil, & Yılmaz, 2018)

Zilyas ve Yılmaz (2023), ortaokul düzeyindeki öğrencilerin eğitimdeki başarılarını tahmin etmek amacıyla makine öğrenimi algoritmalarını uygulamıştır. Çalışmada, 8. sınıf öğrencilerinden alınan 13 soruluk bir anket kullanılmıştır. Ankette, öğrencilerin başarı düzeylerini etkileyebilecek unsurlar arasında aile gelir seviyesi, ders çalışma süreleri, uyku düzeni ve özel ders alımı gibi etkenler yer almıştır. Toplam 519 öğrencinin verisi eğitim (%80) ve test (%20) olarak ayrılmıştır. Rastgele Orman, Çoklu Doğrusal Regresyon, KNN, Karar Ağaçları, Torbalanmış Ağaçlar ve Gradyen Artırıcı Regresyon algoritmaları denenmiştir. Performans değerlendirmesi için RMSE, MSE ve R^2 metrikleri kullanılmıştır. En başarılı yöntem, %88 doğruluk ve %98 R^2 değeriyle Rastgele Orman algoritması olmuştur. Bulgular, aile gelir düzeyi ve çalışma sürelerinin Türkçe ders notları üzerinde en önemli etkilere sahip olduğunu göstermiştir. (Zilyas & Yılmaz, 2023)

Güvenç, Sakal ve ark. (2022), "Bilişim Sistemleri Mühendisliğine Giriş" dersini alan 71 üniversite öğrencisinin dönem başındaki bilgisayar becerilerinin dönem sonundaki başarıları üzerindeki etkisini incelemiştir. Veri setindeki dengesizlik, SMOTE yöntemi kullanılarak giderilmiş ve toplamda 640 eğitim, 300 test verisi oluşturulmuştur. KNN, SVM, Rastgele Orman, Karar Ağaçları, Lojistik Regresyon ve Naive Bayes algoritmaları analiz için kullanılmıştır. En iyi sonuç, %97.66 doğruluk oranıyla KNN ve SVM algoritmalarıyla elde edilmiştir. Çalışma, makine öğrenimi ve veri artırma tekniklerinin başarı tahmininde ve eksikliklerin belirlenmesinde etkili olduğunu göstermiştir. (Güvenç, Sakal, Çetin, & Özkaraca, 2022)

Aghalarova ve Keser (2021), ortaokul öğrencilerinin Matematik ve Portekizce derslerindeki akademik başarılarını tahmin etmek için Yapay Sinir Ağları (YSA)

algoritmasını kullanmıştır. Dengeli bir model oluşturmak adına SMOTE tekniği uygulanmış ve %94.8 doğruluk oranına ulaşılmıştır. Matematik dersi için F1 skoru %94.6, Portekizce için ise %92.8 olarak rapor edilmiştir. (Aghalarova & Keser, 2021)

Adak ve Duralioğlu (2023), İstanbul Ataşehir'de bir meslek lisesinde öğrenim gören 87 öğrencinin Nesne Tabanlı Programlama dersine ait sınav performanslarını tahmin etmek amacıyla bir çalışma gerçekleştirmiştir. Veri eksikliği sebebiyle sentetik veri oluşturulmuş ve bu verinin doğruluğu kontrol edilmiştir. Doğrusal Regresyon, KNN ve Karar Ağaçları algoritmalarıyla yapılan analizlerde, en başarılı sonuçlar KNN yöntemiyle elde edilmiştir. Çalışma, eğitimde başarı tahmini için makine öğrenimi ve sentetik veri kullanımının etkili bir yaklaşım sunduğunu göstermiştir. (Adak & Duralioğlu, 2023)

Alkan (2024), kamuya açık bir veri setini kullanarak, 1000 öğrencinin matematik, okuma ve yazma notlarını ve sosyo-demografik bilgilerini göz önünde bulundurarak sınav başarılarını tahmin etmeye yönelik bir çalışma yapmıştır. Çalışma kapsamında KNN, Karar Ağaçları, Naive Bayes, Rastgele Orman, Destek Vektör Makinesi (DVM), Lojistik Regresyon ve Lineer Diskriminant Analizi gibi algoritmalar kullanılarak analizler gerçekleştirilmiş ve en yüksek doğruluk oranı %67 ile DVM algoritması ile elde edilmiştir. Sonuçlar, makine öğrenmesi tekniklerinin eğitimde başarı tahmini için etkili bir araç olduğunu göstermektedir. (Alkan, 2024)

Bakan ve Kanbay (2024), UCI veri tabanındaki "Higher Education Students Performance Evaluation" veri seti üzerinde, öğrencilerin akademik başarılarını etkileyen faktörleri inceleyerek makine öğrenmesi modelleri geliştirmiştir. Çalışmada 145 öğrenciye ait 31 özellik kullanılarak KNN, Naive Bayes, Rastgele Orman, DVM, Karar Ağaçları ve Boosting algoritmaları uygulanmış ve en yüksek doğruluk %96 ile Rastgele Orman algoritmasıyla elde edilmiştir. Boyut azaltma tekniklerinin de kullanılmasıyla modelin başarısı artırılmış ve çalışmanın eğitimde başarı tahmini açısından oldukça etkili olduğu belirlenmiştir. (Bakan & Kanbay, 2024)

Er ve arkadaşları (2023), 8. sınıf öğrencilerinin LGS (2023) sınavı puanlarını tahmin etmek için bir araştırma yapmışlardır. Araştırmada, 29 farklı özellik içeren bir veri seti kullanılarak Parçacık Sürü Optimizasyonu (PSO), Ateşböceği Algoritması (ABA) ve Guguk Kuşu Arama Algoritması (GKAA) gibi yöntemlerle önemli özellikler belirlenmiş ve Destek Vektör Regresyonu (SVR), Gauss Süreç Regresyonu

(GPR) ve Karar Ağaçları algoritmaları ile tahmin modelleri geliştirilmiştir. GKAA ve SVR kombinasyonu, %80 eğitim-%20 test oranıyla $R^2 = 0.866$ skoru ile en yüksek başarıyı elde etmiştir. Çalışmada doğruluk oranı, MSE, RMSE ve R^2 gibi başarı metrikleri kullanılmış ve öğrencilerin 6., 7. ve 8. sınıf ortalamaları, aile gelir durumu, kardeş sayısı ve özürsüz devamsızlık sayısı gibi faktörlerin başarı üzerindeki etkisi vurgulanmıştır. (Er ve diğerleri, 2023)

Deepika ve Sathyanarayana (2018), öğrencilerin akademik başarılarını tahmin etmek amacıyla UCI Öğrenci Performansı Veri Seti ve Kaggle Kalboard 360 E-Öğrenme Veri Seti gibi iki farklı veri seti kullanarak çeşitli makine öğrenimi algoritmalarını karşılaştırmışlardır. UCI veri seti, Portekiz'deki iki liseden elde edilen sosyal, akademik ve demografik bilgileri içerirken, Kaggle veri seti ise ebeveyn katılımı ve öğrencilerin davranışsal özellikleri gibi veriler sunmaktadır. Çalışmada, Karar Ağaçları (DT), Rastgele Orman (RF), Yapay Sinir Ağları (ANN), Destek Vektör Makineleri (SVM) ve Naive Bayes (NB) algoritmaları kullanılmıştır.

Elde edilen sonuçlara göre, UCI veri setinde en yüksek doğruluk oranı %94 ile Rastgele Orman modeliyle elde edilmiştir. Kaggle veri setinde ise, öğrencilerin davranışsal özelliklerinin de modele dahil edilmesiyle Yapay Sinir Ağları (ANN) modeli %79.1 doğruluk oranı ile en iyi performansı sergilemiştir. Genel olarak, Karar Ağaçları (DT) algoritması, yorumlanabilirliği ve güçlü performans özellikleriyle dikkat çekmiştir. Bu çalışma, makine öğrenmesi tekniklerinin eğitim verilerini analiz etmede ne kadar etkili ve faydalı olabileceğini göstermektedir. (Deepika & Sathyanarayana, 2018)

Salal, Abdullaev ve arkadaşları (2019), Portekiz'deki iki liseden (Gabriel Pereira ve Mousinho da Silveira) toplanan 649 öğrenciye ait verileri kullanarak öğrenci performansını tahmin etmeye yönelik bir çalışma yapmıştır. UCI Machine Learning Repository'den alınan bu veri seti, öğrencilerin akademik, demografik, sosyal ve okul ile ilgili 33 farklı özelliği içermektedir. Çalışmada, Naive Bayes, Karar Ağaçları (J48), Rastgele Orman, Rastgele Ağaç, REPTree, JRip, OneR, Basit Lojistik Regresyon ve ZeroR algoritmaları gibi çeşitli makine öğrenimi yöntemleri uygulanmıştır.

Sonuçlar, en yüksek doğruluk oranını %76.73 ile REPTree ve OneR algoritmalarının elde ettiğini göstermiştir. Ayrıca, başarı tahmininde en etkili özelliğin ikinci dönem notu (G2) olduğu belirlenmiştir. Bu çalışma, veri madenciliği yöntemlerinin öğrenci performansı analizinde ne kadar güçlü ve etkili bir araç

olabileceğini ortaya koymuştur. (Salal, Abdullaev & Kumar, 2019)

Suleiman, Okunade ve arkadaşları (2024), Kaggle platformundan elde edilen ve 10.000 öğrenciye ait verilerle oluşturulan "Student Performance Dataset" kullanarak öğrencilerin akademik başarılarını etkileyen faktörleri incelemişlerdir. Çalışmada, doğrusal regresyon modeli, akademik başarıyı tahmin etmek ve değişkenler arasındaki ilişkileri anlamak amacıyla uygulanmıştır. Veri ön işleme sürecinde, eksik veriler düzeltilmiş, kategorik değişkenler sayısallaştırılmış ve veri seti %80 eğitim, %20 test şeklinde bölünmüştür. Değerlendirme metrikleri olarak R^2 , MSE ve MAE gibi yaygın ölçütler kullanılmıştır.

Çalışmanın sonuçları, en yüksek R^2 skorunun 0.98, MSE değerinin 4.22 ve MAE değerinin ise 1.62 olduğunu göstermiştir. Akademik başarıyı en çok etkileyen faktörler arasında çalışma saatleri, önceki akademik başarı ve çözülmüş örnek sorular öne çıkmıştır. Korelasyon analizine göre, önceki notların başarı üzerindeki etkisi en yüksek korelasyon değeri olan 0.92 ile en güçlü faktör olarak belirlenmiş, çalışma saatleri ise 0.37 ile orta düzeyde pozitif bir korelasyona sahip olmuştur. Diğer faktörler (dış etkinlikler, uyku saatleri, örnek sorular) ise başarı üzerinde daha düşük etkiye sahip olarak bulunmuştur. Bu çalışma, makine öğrenmesi yöntemlerinin öğrenci performansı tahmin etmekte ne kadar etkili bir araç olabileceğini göstermektedir. (Suleiman, Okunade, Dada & Ezeanya, 2024)

Chenrui Pei (2024), Kaggle'dan elde edilen 2.392 lise öğrencisinin verilerini kullanarak öğrencilerin akademik başarılarını tahmin etmek amacıyla bir model geliştirmiştir. Çalışma, öğrencilerin not ortalamalarını (GPA) etkileyen beş ana faktörü incelemiştir: demografik bilgiler, çalışma alışkanlıkları, ebeveyn katılımı, ders dışı etkinlikler ve genel akademik başarı. Veri setinde yer alan demografik değişkenler arasında yaş, cinsiyet, etnik köken ve ebeveynlerin eğitim durumu gibi bilgiler bulunurken, davranışsal faktörler arasında haftalık çalışma süresi, devamsızlık, özel ders alma durumu, ebeveyn desteği, spor, müzik ve gönüllülük gibi ders dışı etkinlikler yer almaktadır.

Tahmin modeli, çoklu doğrusal regresyon kullanılarak oluşturulmuş ve modelin doğruluğunu artırmak için adım adım regresyon yöntemiyle anlamlı olmayan değişkenler çıkarılmıştır. Sonuçlar, haftalık çalışma süresi, ebeveyn desteği ve ders dışı etkinliklerin akademik başarı üzerinde pozitif bir etki yarattığını, devamsızlık sayısının ise en güçlü negatif etkiye sahip olduğunu ortaya koymuştur. Modelin performansı, $R^2 = 0.954$, RMSE = 0.19691, AIC = -781.4 ve BIC = -737.0

gibi yüksek doğruluk oranlarıyla değerlendirilmiştir. Bu bulgular, öğrencilerin akademik başarılarını artırmaya yönelik stratejiler geliştirilmesinde önemli bir temel sağlamaktadır. (Pei, 2024)

Staneviciene ve arkadaşları (2024) tarafından yapılan araştırma, Moodle sanal öğrenme platformunda toplanan öğrenci verileriyle akademik başarı tahmini yapılmasına odaklanmıştır. Araştırma, CRISP-DM (Cross-Industry Standard Process for Data Mining) metodolojisi kullanılarak gerçekleştirilmiştir. Bu çalışmada, öğrencilerin çevrimiçi etkinlikleri, örneğin oturum açma sıklığı, modül tıklama sayısı, bireysel ve grup bazında yapılan görevler, test sonuçları ve forum etkileşimleri gibi veriler analiz edilmiştir. Kullanılan makine öğrenimi teknikleri arasında Karar Ağaçları, Bayes Sınıflandırıcı, KNN, Destek Vektör Makineleri ve Rastgele Orman gibi popüler algoritmalar bulunmuştur.

Çalışma sonuçlarına göre, en yüksek başarıyı Rastgele Orman algoritması elde etmiştir, F-Measure değeri 0.873 ve doğruluk oranı %81 olarak belirlenmiştir. Destek Vektör Makineleri ise ikinci en iyi performansı göstererek F-Measure değeri 0.862 ile sonuçlanmıştır. Ayrıca, araştırma, öğrenci devamsızlık oranlarının yüksek olması ve düşük etkileşim seviyelerinin öğrencinin başarısızlıkla doğrudan ilişkili olduğunu ortaya koymuştur. Bu bulgular, e-öğrenme ortamlarında öğrenci başarısını artırmak amacıyla veri madenciliği ve makine öğrenimi tekniklerinin oldukça etkili bir şekilde kullanılabileceğini göstermektedir. (Staneviciene ve diğerleri, 2024)

Dubey ve arkadaşları (2024) tarafından gerçekleştirilen çalışmada, engelli ve engelsiz bireylerin eğitim başarılarını tahmin etmek için 2013-2020 yılları arasında Office for National Statistics'ten alınan veriler kullanılarak bir model geliştirilmiştir. Bu çalışmada, 21 ile 38 yaş arasındaki bireylerin eğitim seviyeleri, öğrenme platformlarındaki etkileşimleri ve bireysel özellikleri gibi veriler analiz edilmiştir. Eğitim başarılarını tahmin etmek için Lineer Regresyon, Rastgele Orman, Karar Ağaçları, Destek Vektör Makineleri (SVM) ve XGBoost gibi çeşitli makine öğrenimi algoritmaları uygulanmıştır. Modeldeki önemli değişkenler, Recursive Feature Elimination (RFE) tekniğiyle belirlenmiş ve modeller, %80 eğitim ve %20 test oranında değerlendirilmiştir.

En yüksek doğruluk oranı, %90 ile Rastgele Orman algoritmasından elde edilmiştir. Çalışma sırasında kullanılan metrikler arasında R^2 , MSE ve RMSE yer almış, bu metrikler aracılığıyla modelin performansı ölçülmüştür. Ayrıca, araştırma, öğrenme platformlarındaki düşük etkileşim oranlarının engelli bireylerin eğitim

başarısını olumsuz etkilediğini ortaya koymuştur. Bu bulgu, engelli bireyler için daha destekleyici ve etkileşimli eğitim ortamlarının gerekliliğini vurgulamaktadır. Araştırma, eğitim politikalarının iyileştirilmesi ve engelli bireylerin eğitim süreçlerine yönelik daha etkili stratejilerin geliştirilmesinin önemini bir kez daha hatırlatmaktadır. (Dubey ve diğerleri, 2024)

Jingbo ve Yujie (2024) tarafından gerçekleştirilen çalışmada, üniversite öğrencilerinin akademik başarısızlık risklerini önceden tespit edebilmek için bir "Akademik Erken Uyarı Modeli" geliştirilmiştir. Çalışmada, öğrencilerin demografik bilgileri, akademik performansları ve davranışsal özelliklerinden oluşan geniş bir veri seti kullanılmıştır. Yapay Sinir Ağları (ANN), Karar Ağaçları (DT) ve Destek Vektör Makineleri (SVM) gibi farklı makine öğrenimi algoritmaları, öğrencilerin akademik başarısızlık risklerini belirlemek amacıyla uygulanmıştır.

Veri hazırlık aşamasında eksik veriler düzeltilmiş, keşifsel veri analizi (EDA) ile veri seti incelenmiş ve özellik seçimi yapılmıştır. Model, doğruluk, hatırlama ve F1 skoru gibi önemli performans metrikleriyle değerlendirilmiş ve en yüksek başarıyı %94 doğruluk oranıyla Karar Ağaçları algoritması göstermiştir. Bu bulgu, Karar Ağaçları'nın akademik başarısızlık riski tahmininde güçlü bir araç olduğunu ortaya koymaktadır. Sonuçlar, öğrencilerin akademik başarılarının artırılmasında erken müdahale ve doğru analizlerin önemini vurgulamaktadır. Bu araştırma, eğitim sistemlerinde öğrencilerin potansiyel sorunlarını erken tespit etmek ve buna göre destekleyici önlemler almak için etkili bir yaklaşım sunmaktadır. (Jingbo & Yujie, 2024)

Rabelo, Rodrigues ve arkadaşları (2024), eğitim yönetimini geliştirmeye yönelik olarak eğitim veri madenciliği (EDM) ve öğrenme analitiği (LA) tekniklerinin nasıl kullanılabileceğini incelemişlerdir. Çalışmada, farklı e-öğrenme platformlarından (örneğin, Moodle) elde edilen öğrenci davranışları, etkileşimleri ve öğrenme süreçlerini içeren veri setleri analiz edilmiştir. Öğrencilerin bağlılık seviyelerini artırmak ve öğrenme yollarını kişiselleştirmek amacıyla, karar ağaçları, kümeleme teknikleri ve tahminsel modeller gibi çeşitli yöntemler uygulanmıştır.

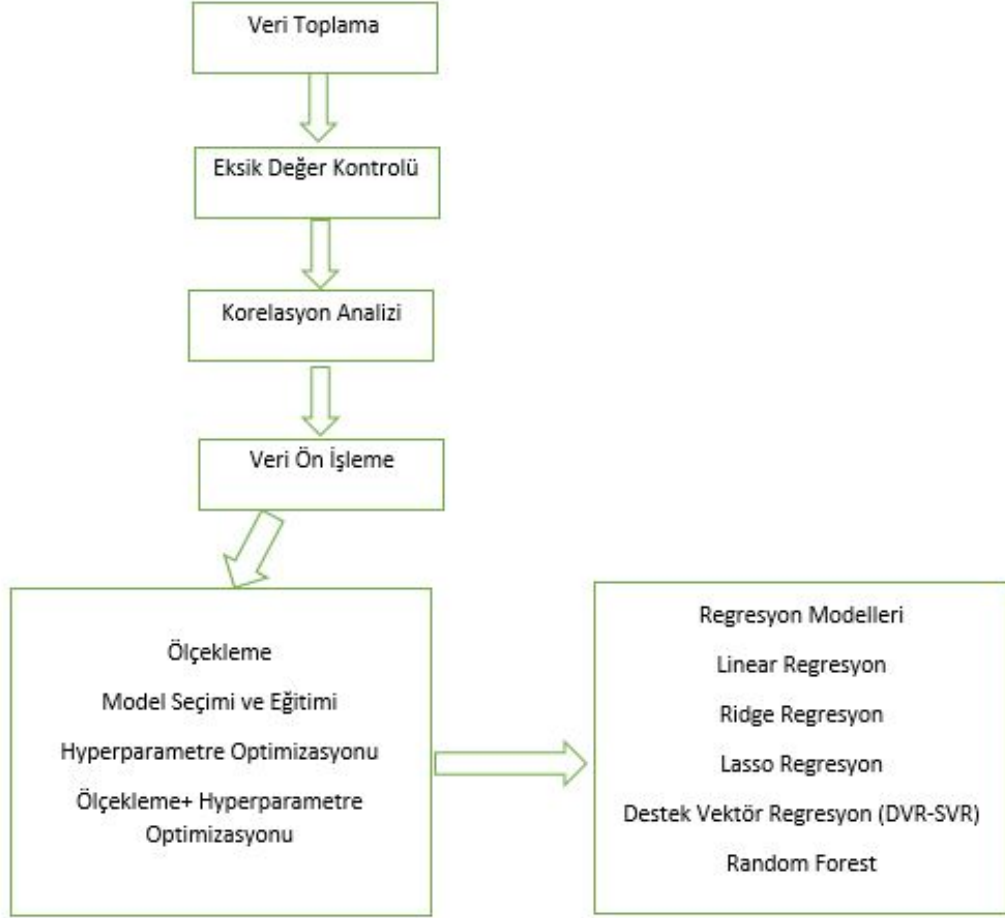
Veri ön işleme, özellik seçimi ve model doğrulama aşamalarından geçirilen bu modeller, öneri sistemleri ve performans tahmini gibi alanlarda olumlu sonuçlar elde etmiştir. Sonuçlar, veri madenciliği tekniklerinin öğrenci başarı oranlarını ve platformlardaki etkileşimi artırmada etkili bir araç olduğunu ortaya koymaktadır. Bu çalışma, eğitimde daha verimli ve kişiselleştirilmiş öğrenme deneyimleri oluşturmak

için veri madenciliği yöntemlerinin önemini vurgulamaktadır. (Rabelo, Rodrigues, Nobre, Isotani & Zárata, 2024)

3. GEREÇ VE YÖNTEM

Bu çalışmada, Kaggle'dan alınan beş veri seti kullanılmıştır. Bu veri setlerinde, öğrenci başarısını etkileyen çeşitli faktörler yer almaktadır. Çalışmanın amacı, bu faktörleri kullanarak öğrencilerin akademik başarılarını tahmin etmektir. Veri ön işleme aşamasında, eksik veri kontrolü yapılmış, kategorik veriler sayısal verilere dönüştürülmüş ve özelliklerin ölçeklendirilmesi sağlanmıştır.

Veri seti üzerinde yapılan analizlerde, farklı makine öğrenimi algoritmaları kullanılmıştır. Bu algoritmalar arasında; Lineer Regresyon (LR), Ridge Regresyon, Lasso Regresyon, Destek Vektör Regresyonu(DVR-SVR), Random Forest yer almaktadır. Her bir modelin başarımı, R^2 skoru, MAE, RMSE gibi metriklerle değerlendirilmiştir. Model optimizasyonu ve hiperparametre ayarlamaları, GridSearchCV yöntemi ile gerçekleştirilmiştir. En yüksek doğruluk oranına ulaşmak için modeller üzerinde çapraz doğrulama (cross-validation) teknikleri de uygulanmıştır. Ayrıca, model başarısının artması için özellik mühendisliği yöntemleri kullanılmış ve modelin genellenebilirliğini artırmak amacıyla farklı özellik kombinasyonları denenmiştir. Şekil 3.1. de önerilen yöntemin akış diyagramı gösterilmiştir.



Şekil 3.1. Önerilen Yöntemin Akış Diyagramı

3.1. Veri Seti

Çalışma da kullanılan ilk veri seti kaggle platformundan alınmış olup toplamda 6.607 kayıt ve 20 öznitelikten oluşmaktadır. (Alaa, t.y.). Veri setinde yer alan bağımsız değişkenler ve veri türleri Çizelge 3.1. de gösterilmiştir:

Çizelge 3.1. Veri seti 1, öznitelik ve veri tipleri

Öznitelik	Açıklama	Veri Tipleri
Hours_Studied	Çalışma Süresi	int64
Attendance	Devamsızlık Durumu	int64
Parental_Involvement	Ebeveyn Katılımı	object
Access_to_Resources	Kaynaklara Erişim	object
Extracurricular_Activities	Ders Dışı Aktiviteler	object
Sleep_Hours	Uyku Süresi	int64
Previous_Scores	Önceki Başarı Durumu	int64
Motivation_Level	Motivasyon Seviyesi	object
Internet_Access	İnternet Erişimi	object
Tutoring_Sessions	Özel Ders Sayısı	int64
Family_Income	Aile Geliri	object
Teacher_Quality	Öğretmen Kalitesi	object
Scholl_Type	Okul Türü	object
Peer_Influence	Akran Etkisi	object
Physical_Activity	Fiziksel Aktivite Düzeyi	int64
Learning_Disalibities	Öğrenme Güçlükleri	object
Parental_Education_Level	Ebeveyn Eğitim Seviyesi	object
Distance_from_Home	Ev-Okul Arası Mesafe	object
Gender	Cinsiyet	object
Exam_Score	Sınav Notu	int64

Kullanılan ikinci veri seti ise; Kaggle platformundan alınmış olup öğrenci akademik başarısını etkileyen faktörleri analiz etmek amacıyla oluşturulmuştur. (Shamim, t.y.). Veri seti, 10.000 kayıt ve 15 öznitelikten oluşmaktadır. İçeriğinde öğrencilerin akademik performanslarını, öğrenme alışkanlıklarını ve günlük yaşam faktörlerini içeren çeşitli değişkenler bulunmaktadır. Veri setinde yer alan bağımsız değişkenler ve veri türleri Çizelge 3.2. de gösterilmiştir.

Çizelge 3.2. Veri seti 2, öznitelik ve veri tipleri

Öznitelik	Açıklama	Veri Tipleri
Student_ID	Öğrenci Kimlik Numarası	int
Age	Öğrencinin Yaşı	int
Gender	Öğrenci Cinsiyeti	String
Study_Hours_per_Week	Haftalık Çalışma Süresi	Float
Preferred_Learning_Style	Tercih Edilen Öğrenme Stili	String
Online_Courses_Completed	Tamamlanan Çevrimiçi Kurs Sayısı	int
Participation_in_Discussions	Tartışmalara katılım Düzeyi	Float
Assignment_Completion_Rate(%)	Ödev Tamamlama Oranı	Float
Exam_Score(%)	Sınav Puanı	Float
Attendance_Rate(%)	Derse Devam Oranı	Float
Use_of_Educational_Tech	Eğitim Teknolojisi Kullanım Durumu	String
Self_Reported_Stress_Level	Öğrencinin Bildirdiği Stres Seviyesi	int
Time_Spent_on_Social_Media	Haftalık Sosyal Medya Kullanımı	Float
Sleep_Hours_per_Night	Günlük Uyku Süresi	Float
Final_Grade	Final Notu	Float

Kullanılan üçüncü veri seti ise; Kaggle platformundan alınmış olup öğrenci akademik başarısını etkileyen faktörleri analiz etmek amacıyla oluşturulmuştur.(Stealth Technologies, t.y.). Veri seti, 1388 kayıt ve 5 öznitelikten oluşmaktadır. Veri setine ait öznitelikler Çizelge 3.3. de gösterilmiştir.

Çizelge 3.3. Veri seti 3, öznitelik ve veri tipleri

Öznitelik	Açıklama	Veri Tipleri
Socioeconomic Score	Sosyoekonomik Skor	Float64
Study Hours	Çalışma Saatleri	Float64
Sleep Hours	Uyku saatleri	Float64
Attendance(%)	Devam Oranı	Float64
Grades	Notlar/Başarı Durumu	Float64

Kullanılan dördüncü veri seti ise; Kaggle platformundan alınmış olup öğrenci akademik başarısını etkileyen faktörleri analiz etmek amacıyla oluşturulmuştur.(Abbas, 2024). Veri seti, 10000 kayıt ve 6 öznitelikten oluşmaktadır. Veri setine ait öznitelikler Çizelge 3.4. de gösterilmiştir.

Çizelge 3.4. Veri seti 4, öznitelik ve veri tipleri

Öznitelik	Açıklama	Veri Tipleri
Hours Studied	Çalışılan Saat	int64
Previous Scores	Önceki Notlar	int64
Extracurricular Activities	Dış etkinliklere katılım	object
Sleep Hours	Uyku saatleri	int64
Sample Question Papers Practiced	Çalışılan örnek soru sayısı	int64
Performance Index	Performans indeksi(Hedef değişken)	float64

Çalışmada kullanılan son veri ise; Kaggle platformundan alınmış olup 2.392 kayıt ve 15 öznitelik içermektedir. (El Kharoua, 2024). Veri setine ait öznitelikler Çizelge 3.5. de gösterilmektedir.

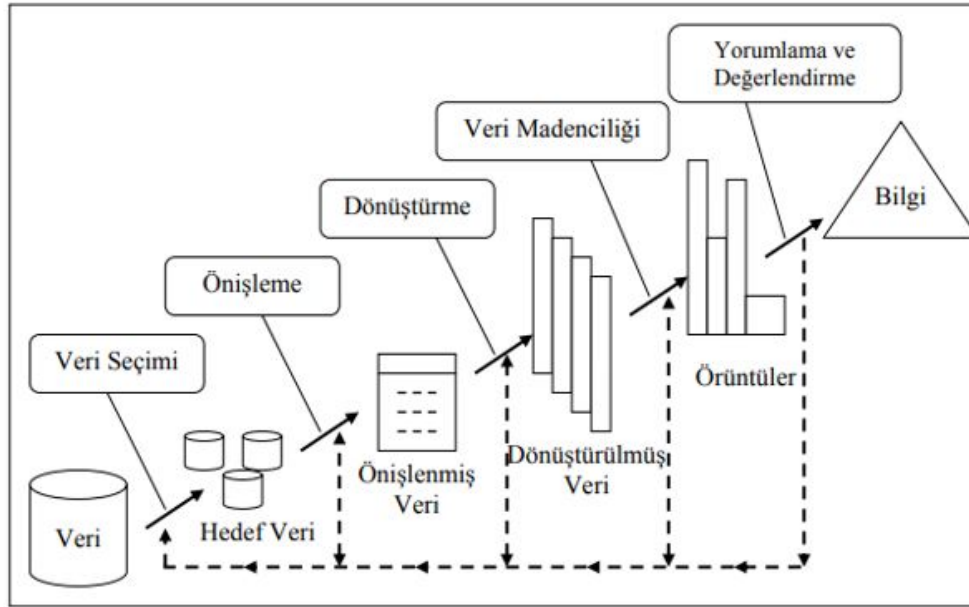
Çizelge 3.5. Veri seti 5, öznitelik ve veri tipleri

Öznitelik	Açıklama	Veri Tipleri
StudentID	Öğrenci Kimlik Numarası	int64
Age	Yaş	int64
Gender	Cinsiyet	int64
Ethnicity	Etnik Köken	int64
ParentalEducation	Ebeveyn Eğitim Seviyesi	int64
StudyTimeWeekly	Haftalık Çalışma Süresi	float64
Absences	Devamsızlık Sayısı	int64
Tutoring	Özel Ders Alımı	int64
ParentalSupport	Ebeveyn Desteği	int64
Extracurricular	Dış Etkinliklere Katılım	int64
Sports	Spor Etkinliklerine Katılım	int64
Music	Müzik Etkinliklerine Katılım	int64
Volunteering	Gönüllü Faaliyetlere Katılım	int64
GPA	Genel Not Ortalaması	float64
GradeClass	Sınıf Notu	float64

3.2. Veri Madenciliği

Günümüzde hızla artan veri yığınları, gelişmiş veri depolama teknolojileriyle birlikte daha karmaşık ilişkiler içermeye başlamıştır. Geleneksel analiz yöntemleri, bu büyük ve karmaşık veri kümelerini çözmekte yetersiz kalmaktadır. Veriler arasında önemli ilişkiler kurabilmek ve önemli bilgiler ortaya çıkarabilmek için yeni araçlar ve yöntemlerin geliştirilmesigerekli hale gelmiştir.Bu gereksinimi karşılamak için, bilgisayar,veri tabanı, analiz ve farklı disiplinlerin bir araya gelmesiyle oluşmuştur. Büyük veri kümelerini analiz ederek gizli kalmış ilişkileri ve anlamlı bilgileri keşfeden bir yöntem olarak değer kazanmıştır.(Can ve arkadaşları, 2012).

Büyük ve karmaşık veri kümelerinin analiz edilmesi, gizli kalmış ilişkilerin keşfedilmesi ve anlamlı bilgilerin ortaya çıkarılması için veri madenciliği bilgi keşif süreci kullanılmaktadır. Bu süreç, öncelikle ham verinin toplanması ve uygun veri seçimi ile başlar. Seçilen verinin analiz için uygun hale getirilmesi amacıyla ön işleme aşamasında eksik, gürültülü veya tutarsız veriler temizlenir. Daha sonra, farklı veri madenciliği teknikleri kullanılarak veriler arasındaki ilişkiler ve desenler belirlenir. Son olarak, elde edilen bulguların anlamlandırılması için yorumlama ve değerlendirme aşamasına geçilir ve sonuçlar karar destek sistemlerinde kullanılabilir hale getirilir. Şekil 3.2.'de bu sürecin görseli yer almaktadır.



Şekil 3.2. Veri Madenciliği bilgi keşif süreci(Fayyad vd., 1996, s.41),(Irmak, 2009)

3.2.1. Veri Ön İşleme

Veri ön işleme, ham verinin modele uygun hale getirilmesi için yapılan bir dizi işlemi kapsar. Bu aşama, hatalı veya eksik verilerin düzeltilmesi, gereksiz verilerin çıkarılması, sayısal olmayan verilerin dönüştürülmesi gibi adımları içerir. Temiz ve düzgün işlenmiş veri, modelin daha güvenilir sonuçlar üretmesini ve tahminlerin doğruluğunu artırmasını sağlar (Yılmaz ve ark., 2021). Ayrıca, veri ön işleme döngüsü, veri dönüştürme, birleştirme, temizleme ve ayıklama gibi işlemleri içerir. Verinin güvenilir kaynaklardan toplanması, hatalı verilerin ayıklanması ve eksik bilgilerin gözden geçirilmesi, süreçte karşılaşılabilecek sorunları azaltır ve

işlem süresini optimize eder (Bölükbaş, 2013).

Korelasyon Analizi, veri setindeki öz nitelikler arasındaki en yüksek değere sahip olan ve birbirlerinden ayrı özellikleri içerenleri belirleyerek seçim yapmayı amaçlar (Emhan ve ark, 2019).

Pearson Katsayısı (r) Formülü:

Verilen formül:

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (3.1)$$

Veri temizleme ve azaltma işlemleri, gürültü, eksik veriler ve gereksiz özelliklerden kaynaklanan sorunları ortadan kaldırırsa da, işlenmiş veri seti hâlâ analiz için uygun olmayabilir. Uygun olmayan veri yapıları, modelin doğruluğunu ve verimliliğini olumsuz etkileyebilir, bu yüzden verinin algoritmalara uygun formata dönüştürülmesi gereklidir. Veri dönüştürme işlemleri arasında normalleştirme ve veri entegrasyonu önemlidir (Çetin ve Yıldız, 2021). Kategorik verilerin sayısal verilere dönüştürülmesinde etiket kodlama (Label Encoding) ve One Hot kodlama gibi teknikler kullanılır. Etiket kodlama, sıralı kategorilerde tercih edilir ve her kategori bir tamsayıya dönüştürülür (Mwamba et al., 2021). Z skor normalizasyonu ise, her bir değeri ortalama ve standart sapmaya göre dönüştürerek normalleştirme sağlar (Oğuzlar, 2003).

Veri büyüdükçe karmaşıklığı artar ve işlem süresi uzar, bu nedenle verilerin boyutunun küçültülmesi önemlidir, ancak veri kalitesinin bozulmaması gerekir. Özellik seçimi ve çıkarma teknikleri, bu tür zorlukların üstesinden gelmek için kullanılır (Çetin ve Yıldız, 2021). Özellik seçimi, doğruluktan ödün vermeksizin, gereksiz verilerin çıkarılmasını ve model doğruluğunun artırılmasını amaçlar. Bu süreçte, hedef değişkenle güçlü ilişkisi olan özellikler seçilir ve gereksiz özellikler elenir (Microsoft, n.d.). Normalizasyon, verilerin farklı ölçekte saklandığı durumlarda kullanılır, örneğin finansal verilerde (Tunç & Ülger, 2016). Z-Skoru standartlaştırma ise, verinin ortalama ve standart sapmasına dayalı olarak yapılan yaygın bir yöntemdir ve verilerin daha anlamlı ve hızlı işlenmesini sağlar (Khemka, 2003).

3.3. Çalışmada Kullanılan Makine Öğrenmesi Yöntemleri

Makine öğrenmesi, bir problemi veriye dayanarak çözmek amacıyla kullanılan algoritmaların genel adıdır. Birçok farklı yöntem geliştirilmiş olup, bunlar arasında naive bayes, destek vektör makineleri, ve random forest gibi algoritmalar yer alır. Bu yöntemler, kümelere ayırma veya sınıflandırma gibi işlemler için kullanılır. Öğrenme stratejilerine göre üçe ayrılır: denetimsiz, denetimli ve pekiştirmeli öğrenme. Denetimli öğrenmede model, girdi ve hedef arasındaki bağıntıyı öğrenerek doğru tahminler yapar. Denetimsiz öğrenmede, yalnızca girdi verileriyle kümeler oluşturulur. Pekiştirmeli öğrenmede ise modelin çıktıları, iyilik veya kötülük gibi kriterlerle geri bildirimle değerlendirilir (Atalay ve Çelik, 2017).

Doğrusal regresyon analizi, hedeflenen bir değişkeni, daha kolay veya daha erken ölçülebilen diğer değişkenler kullanılarak tahmin etmeye yönelik bir model oluşturma yöntemidir. (Alpar, 2010). Örneğin, bir öğrencinin yıl sonu akademik başarısını, doğrudan yıl sonu sınavına girmeden; öğrencinin ders çalışma süresi, önceki sınav notları, uyku düzeni ve katıldığı etkinlik sayısı gibi başarı ile ilişkili olduğu bilinen diğer değişkenlere dayanarak tahmin etmek amaçlanabilir.

Rastgele Orman, sınıflandırma ve regresyon ağaçları kullanarak büyük, esnek ve verimli sonuçlar üreten, kullanımı kolay bir makine öğrenmesi algoritmasıdır. Bu algoritma, torbalama yöntemini temel alan topluluk öğrenme algoritmaları arasında yer almaktadır. Rastgele Orman'ın en güçlü yöntemlerinden bir tanesi, diğer kullanılan algoritmaların temel bileşenini oluşturan regresyon problemlerini de çözebilmesidir. (Kumral ve ark. 2022)

Makine öğrenmesinde, Destek Vektör Makineleri (DVM), verileri analiz ederek sınıflandırma amaçları için de kullanılabilen bir algoritma olarak yer almaktadır. DVM'ler, eğitim verilerinde her örneğin belirli bir kategoriye ait olduğu bir dizi örnek üzerinde çalışarak, yeni verileri bu kategorilere atayabilen bir model oluşturur. Bu model, ikili doğrusal sınıflandırma yapan bir yapıdır ve sınıflandırma işlemini olasılık hesaplamadan gerçekleştirir. (Jordan ve ark. 2008). 1995 yılında Vapnik tarafından geliştirilen Destek Vektör Makineleri (DVM) Vapnik teorisiyle ilişkilendirilen denetimli makine öğrenimi algoritmasını tasarlamış, bir sonraki aşamada tahmin yapabilmek için DVR (Destek Vektör Regresyonu) oluşturmuştur. (Drucker ve diğerleri, 1997)

Ridge regresyonu, bir olayın meydana gelmesinde etkili olan tüm

değişkenlerin modele dahil edilmesini sağlayarak, bu faktörlerin toplam etkisinin daha doğru bir şekilde incelenmesine olanak tanır. Bu yöntem, özellikle çoklu doğrusal bağıllık (multicollinearity) sorunlarının yaşandığı durumlarda, normal en küçük kareler (EKK) yöntemine göre daha başarılı sonuçlar elde eder. Ridge regresyonu, modelin parametrelerini düzenler ve böylece yüksek korelasyona sahip bağımsız değişkenlerin model üzerindeki etkisini azaltarak daha sağlam ve güvenilir tahminler yapar. Bu sayede, modelin genelleme yeteneği artar ve overfitting (aşırı uyum) riski azaltılır. (Marquardt & Snee, 1975)

Lasso regresyon modeli , 1996 yılında Tibshirani tarafından ortaya atılmış ve yanlı tahmin yöntemleri arasında yer alır. Ridge regresyonuna benzer şekilde, Lasso da katsayılar üzerine ceza terimi ekleyerek çalışır. Ancak bu yöntem, bazı katsayıları sıfıra indirerek, gereksiz değişkenlerin modelden çıkarılmasını sağlar. (Orhan ve Vergili,2022)

3.4. Önerilen Yöntemin Kullanılması

Kaggle platformundan beş adet farklı veri seti temin edilmiştir. İlk olarak, bu veri setleri üzerinde ön işleme işlemleri gerçekleştirilmiştir. Veri ön işleme aşamalarının ilk adımı olarak, veri setlerinde eksik değerlerin bulunup bulunmadığı kontrol edilmiştir. Eksik değerlerin varlığı durumunda model performansının olumsuz etkilenebileceği dikkate alınarak bu kontrol titizlikle yapılmıştır. Eksik değerlere rastlanmaması durumunda diğer aşamalara geçilmiştir

3.4.1. Veri Ön İşleme Adımlarının Kullanılması

Veri setinde bulunan kategorik değişkenler, makine öğrenimi algoritmalarının bu tür verilerle doğrudan çalışamaması nedeniyle sayısal değerlere dönüştürülmüştür. Bu işlem için Label Encoding yöntemi kullanılmıştır. Label Encoding, kategorik verilerin her bir sınıfını benzersiz bir tam sayı ile eşleştiren bir dönüşüm yöntemidir. Bu yöntem, kategorik değişkenlerin basit ve hızlı bir şekilde işlenmesini sağlamış ve modellerin bu verilerle daha etkili çalışmasına olanak tanımıştır.

Model performansının daha doğru bir şekilde değerlendirilmesi amacıyla on katlı çapraz doğrulama yöntemi kullanılmıştır. Bu işlemde, kullanılan seti on parçaya eşit bir şekilde bölünmüş ve her kat, test seti olarak kullanılarak modelin eğitimi gerçekleştirilmiştir. Bu yöntem, modelin daha güvenilir bir performans ölçümü sağlar ve overfitting riskini azaltır.

Daha sonra, veri setindeki değişkenler bağımlı ve bağımsız değişkenler olarak ayrılmıştır. Bağımsız değişkenler X, öğrenci performansını etkileyen çeşitli faktörleri temsil ederken, bağımlı değişken Y, öğrenci performansını ifade etmektedir.

Modellerin performansını artırmak amacıyla bağımsız değişkenler arasında korelasyon analizi yapılmıştır. Bu analizle, değişkenler arasındaki ilişkiler incelenmiştir. Analiz sonucunda yüksek korelasyona sahip olan değişkenler birleştirilmiş, düşük korelasyona sahip olan değişkenler ise çıkarılmıştır. Bu işlem, modelin daha anlamlı ve etkili bir şekilde öğrenmesini sağlamak için gerçekleştirilmiştir. Çizelge 3.6, Çizelge 3.7, Çizelge 3.8, Çizelge 3.9 ve Çizelge 3.10'da hedef değişken ile bağımsız değişkenler arasındaki korelasyon analizi sonuçları gösterilmiştir. Şekil 3.3, Şekil 3.4, Şekil 3.5, Şekil 3.6 ve Şekil 3.7'de ise değişkenler arasındaki ilişkileri görselleştiren ısı haritası yer almaktadır.

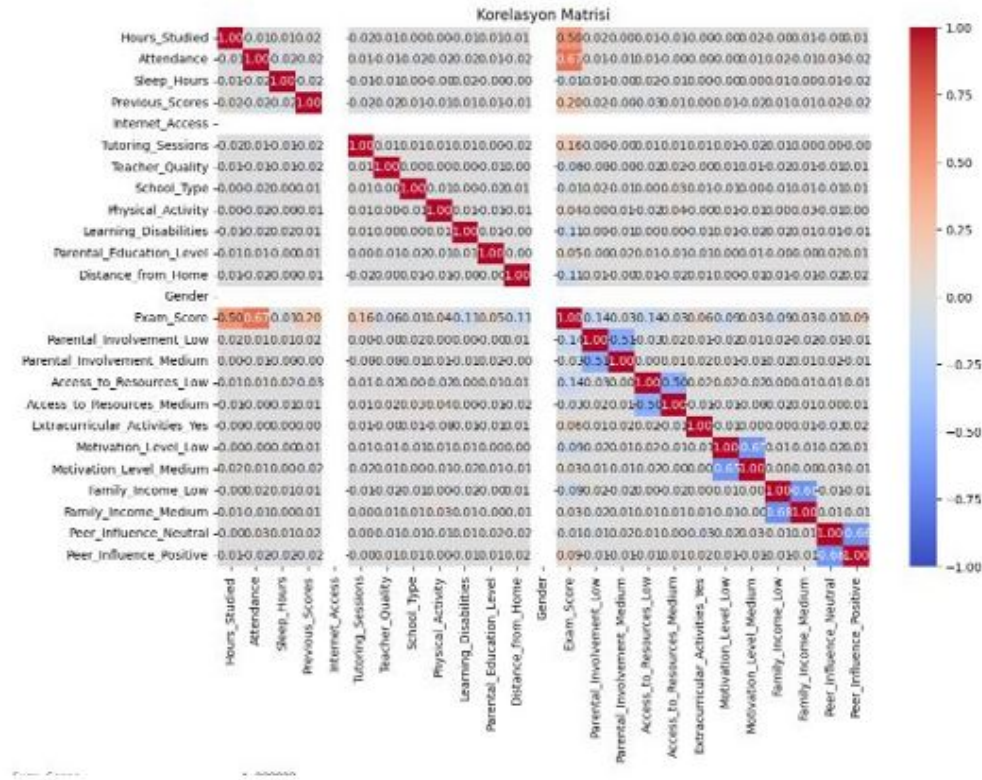
Çizelge 3.6, Veri seri 1'de yer alan "Exam_Score" ile diğer özellikler arasındaki korelasyonları göstermektedir. Bu tablo, her bir özelliğin "Exam_Score" ile ilişkisini sayısal olarak ortaya koyarak, hangi özelliklerin sınav puanı üzerinde daha güçlü bir etkiye sahip olduğunu görmemizi sağlar.

Çizelge 3.6. Veri seti1, 'Exam_Score' ile diğer özelliklerin korelasyonları

Hours_Studied	0.502556
Attendance	0.673602
Parental_Involvement	-0.030021
Access_to_Resources	-0.032350
Extracurricular_Activities	0.064757
Sleep_Hours	-0.013483
Previous_Scores	0.199388
Motivation_Level	0.031081
Internet_Access	NaN
Tutoring_Sessions	0.161826
Family_Income	0.029020
Teacher_Quality	-0.06111
Scholl_Type	-.0.0147
Peer_Influence	0.090791
Physical_Activity	0.044705
Learning_Disalibities	-0.10670
Parental_Education_Level	0.051562
Distance_from_Home	-0.10571
Gender	NaN
Exam_Score	1.00000

Isı haritası, Veri seti 1'de yer alan değişkenler arasındaki ilişkileri görsel olarak göstermektedir. Her bir hücre, iki değişken arasındaki korelasyonu renk ile temsil eder; koyu renkler güçlü bir ilişkiyi, açık renkler ise zayıf bir ilişkiyi göstermektedir. Bu harita, değişkenler arasındaki korelasyonların hızlıca analiz edilmesini sağlayarak, güçlü ilişkiler ve potansiyel etkileyen faktörlerin tespit edilmesine olanak tanımaktadır. Sınav puanı gibi hedef değişkenin anlaşılması ve bu değişkenin diğer faktörlerle nasıl etkileşimde bulunduğu gözlemlenmektedir.

Şekil 3.3' de, Veri seti 1'de yer alan değişkenler arasındaki ilişkileri görsel olarak göstermektedir.



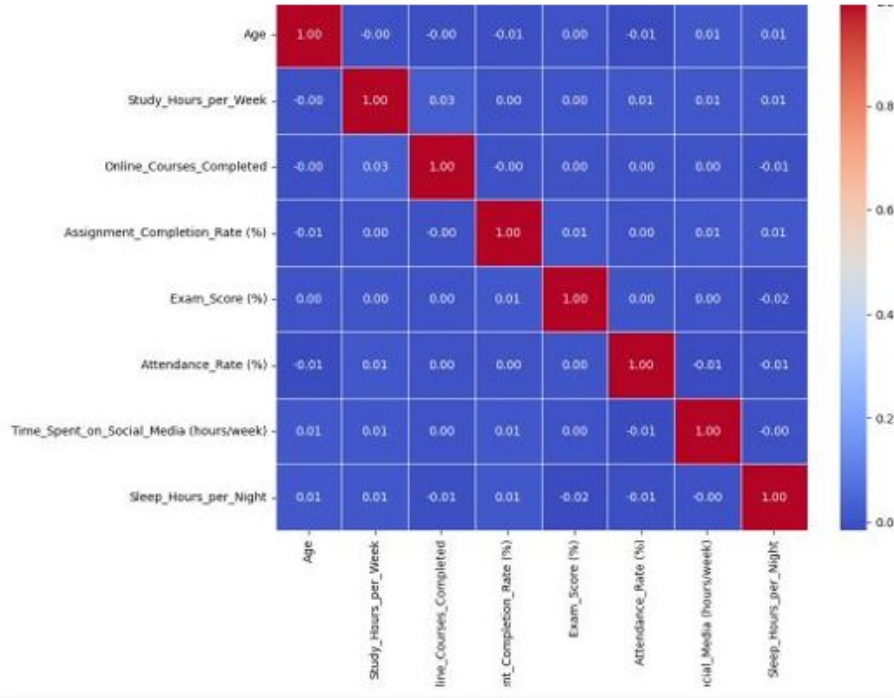
Şekil 3.3. Veri seti 1, değişkenler arasındaki ilişkiyi gösteren ısı haritası

Çizelge 3.7, Veri seti 2'de yer alan "Exam_Score(%)" ile diğer özellikler arasındaki korelasyonları göstermektedir. Bu korelasyonlar, sınav puanlarının yüzdesel değerinin diğer faktörlerle nasıl ilişkili olduğunu anlamamıza yardımcı olmaktadır. Verilen özellikler ile sınav yüzdesi arasındaki güçlü veya zayıf ilişkiler, öğrencilerin performansını etkileyebilecek faktörlerin belirlenmesine olanak sağlar. Bu bilgiler, hangi değişkenlerin sınav başarısını etkilediğini analiz etmek ve eğitim stratejilerini geliştirmek için kullanılabilir.

Çizelge 3.7. Veri seti 2, 'Exam_Score(%)' ile diğer özelliklerin korelasyonları

Age	0.0016
Gender	0.0006
Study_Hours_per_Week	0.0040
Preferred_Learning_Style	-0.0107
Online_Courses_Completed	0.0020
Participation_in_Discussions	0.0015
Assignment_Completion_Rate(%)	0.0104
Exam_Score(%)	1.0000
Attendance_Rate(%)	0.0037
Use_of_Educational_Tech	-0.0054
Self_Reported_Stress_Level	-0.0255
Time_Spent_on_Social_Media	0.0008
Sleep_Hours_per_Night	-0.0162
Final_Grade	-0.9679

Isı haritası, Veri seti 2'de yer alan değişkenler arasındaki ilişkileri görsel olarak sunmaktadır. Her bir hücre, iki değişken arasındaki korelasyonu renklerle ifade eder; koyu renkler güçlü, açık renkler ise zayıf ilişkileri gösterir. Bu görsel, veri setindeki değişkenlerin birbirleriyle nasıl etkileşime girdiğini hızlıca analiz etmemize olanak sağlar ve önemli bağlantıları tespit etmek için faydalıdır.



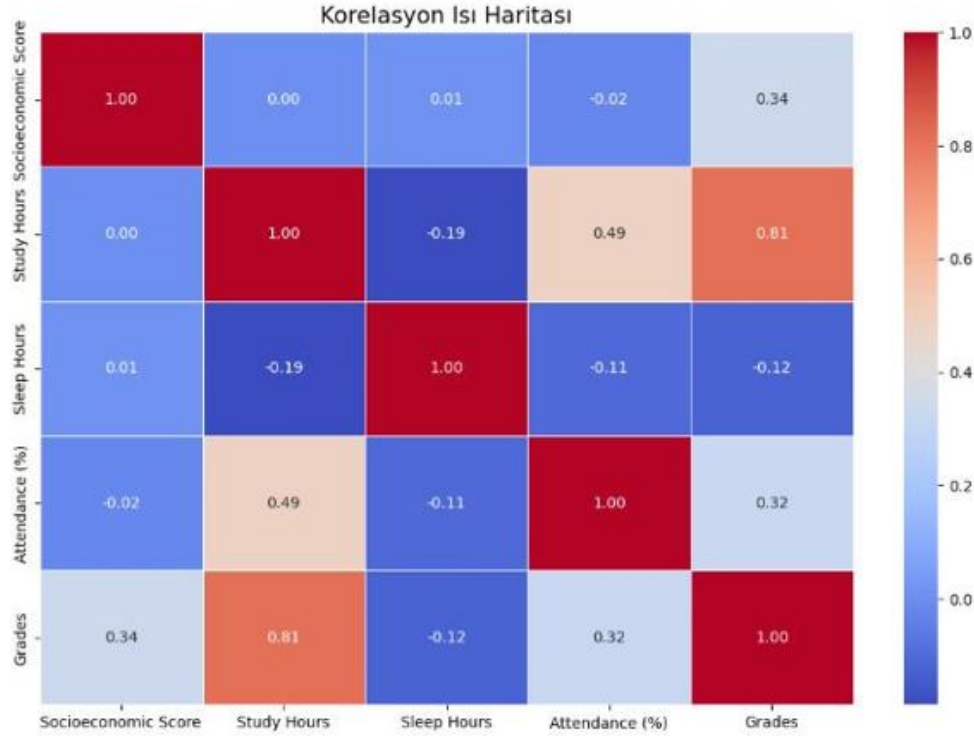
Şekil 3.4. Veri seti 2, değişkenler arasındaki ilişkiyi gösteren ısı haritası

Çizelge 3.8'de, Veri seti 3'de yer alan "Grades" ile diğer özellikler arasındaki korelasyonları göstermektedir. Bu korelasyonlar, "Grades" ile diğer faktörlerin nasıl ilişkili olduğunu hızlıca anlamamıza yardımcı olur ve öğrencilerin başarı düzeyini etkileyen önemli değişkenlerin tespit edilmesini sağlar.

Çizelge 3.8. Veri seti 3, 'Grades' ile diğer özelliklerin korelasyonları

Socioeconomic Score	0.3350
Study Hours	0.8108
Sleep Hours	-0.1246
Attendance(%)	0.3219
Grades	1.0000

Isı haritası, Veri seti 3'de yer alan değişkenler arasındaki ilişkileri görsel olarak sunmaktadır. Değişkenler arasındaki korelasyonlar, renkler aracılığıyla gösterilir; koyu renkler güçlü ilişkileri, açık renkler ise zayıf ilişkileri belirtir. Bu harita, değişkenler arasındaki etkileşimleri hızlı bir şekilde analiz etmeye yardımcı olur.



Şekil 3.5. Veri seti 3, değişkenler arasındaki ilişkiyi gösteren ısı haritası

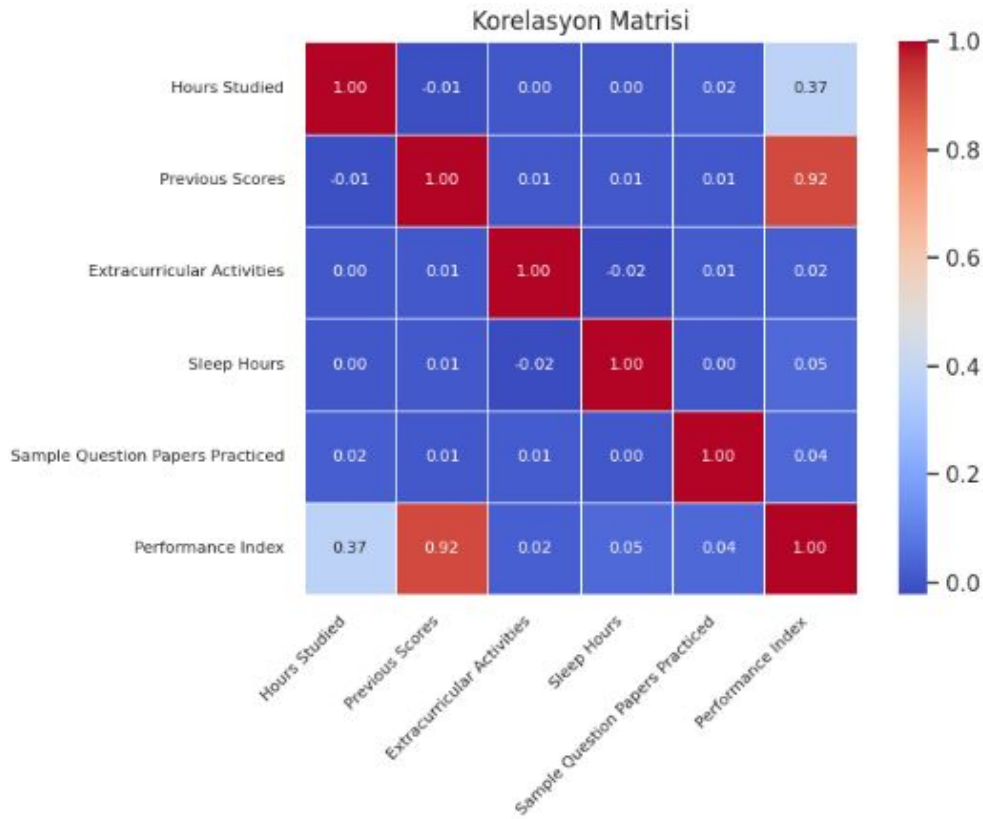
Çizelge 3.9, Veri seti 4'de yer alan "Performance Index" ile diğer özellikler arasındaki korelasyonları göstermektedir. Bu korelasyonlar, "Performance Index" ile diğer faktörler arasındaki ilişkiyi ortaya koyarak, öğrencilerin performansını etkileyebilecek önemli değişkenlerin belirlenmesine yardımcı olur. Bu sayede, performans göstergesi ile güçlü bir ilişkiye sahip olan faktörler tespit edilebilir.

Çizelge 3.9. Veri seti 4, 'Performance Index' ile diğer özelliklerin korelasyonları

Performance Index	1.0000
Previous Scores	0.9151
Hours Studies	0.3737
Sleep Hours	0.0481
Sample Question Papers Practiced	0.0432
Extracurricular Activities	0.0245

Isı haritası, Veri seti 4'de yer alan değişkenler arasındaki ilişkileri görsel olarak sunmaktadır. Değişkenler arasındaki korelasyonlar renklerle gösterilir; koyu renkler güçlü, açık renkler ise zayıf ilişkileri işaret eder. Bu harita, değişkenler arasındaki etkileşimleri hızlıca analiz etmek için kullanılır ve önemli ilişkilerin

belirlenmesini sağlar.



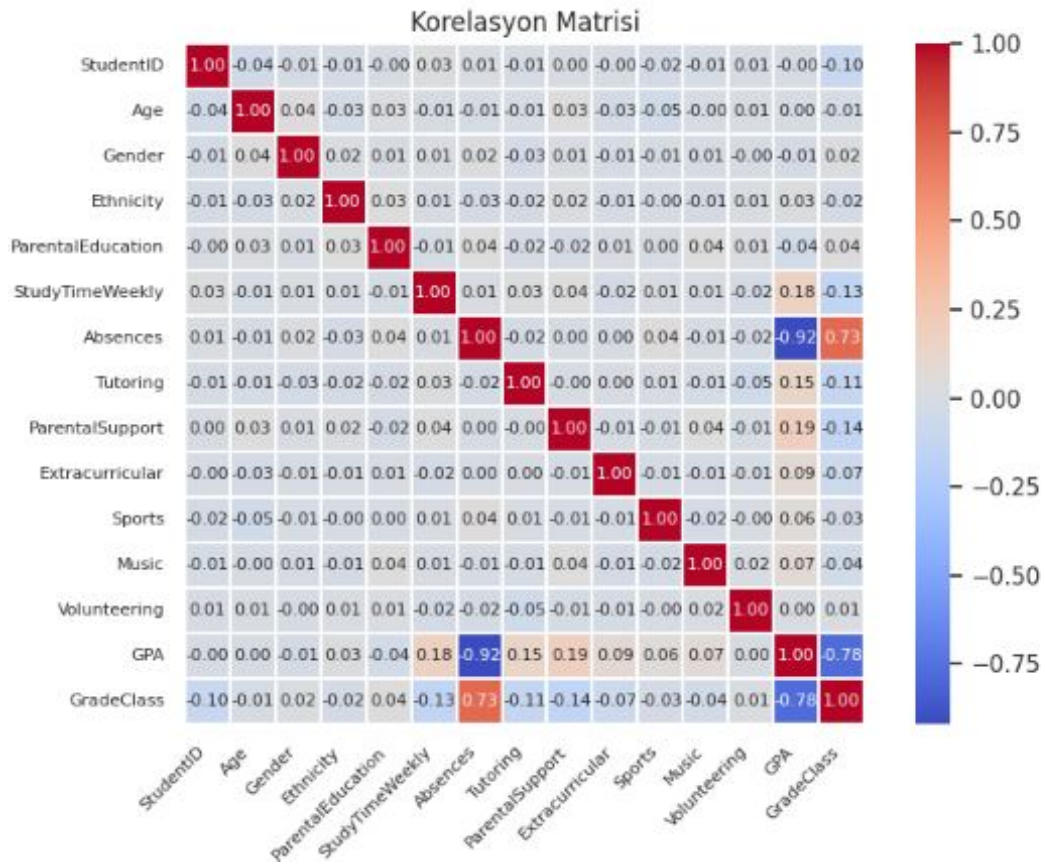
Şekil 3.6. Veri seti 4, değişkenler arasındaki ilişkiyi gösteren ısı haritası

Çizelge 3.10, Veri seti 5'te yer alan "GPA" ile diğer özellikler arasındaki korelasyonları göstermektedir. Bu korelasyonlar, "GPA" ile diğer faktörler arasındaki ilişkiyi ortaya koyarak, öğrencilerin akademik başarılarını etkileyebilecek önemli değişkenlerin belirlenmesine yardımcı olur.

Çizelge 3.10. Veri seti 5, 'GPA' ile diğer özelliklerin korelasyonları

StudentID	-0.0026
Age	0.0002
Gender	-0.0133
Ethnicity	0.0277
ParentalEducation	-0.0358
StudyTimeWeekly	0.1792
Absences	-0.9193
Tutoring	0.1451
ParentalSupport	0.1907
Extracurricular	0.0940
Sports	0.0578
Music	0.0733
Volunteering	0.0032
GradeClass	-0.7828

Isı haritası, Veri seti 5'te yer alan değişkenler arasındaki ilişkileri görsel olarak sunmaktadır. Harita, her bir değişkenin diğerleriyle olan korelasyonlarını renkler aracılığıyla gösterir; koyu renkler güçlü ilişkileri, açık renkler ise zayıf ilişkileri ifade eder. Bu görsel, değişkenler arasındaki etkileşimleri hızlıca analiz etmeye olanak tanır ve önemli ilişkilerin tespit edilmesini sağlar.



Şekil 3.7. Veri seti 5, değişkenler arasındaki ilişkiyi gösteren ısı haritası

Veri setindeki bağımsız değişkenler, ölçek farklılıklarının modeller üzerindeki olumsuz etkilerini azaltmak amacıyla standartlaştırma işlemi ile ölçeklendirilmiştir. Standartlaştırma, her bir değişkenin ortalamasını 0 ve standart sapmasını 1 yapmak için kullanılan bir tekniktir. Bu işlemde, her bir değerden ortalama çıkarılır ve ardından bu fark, değişkenin standart sapmasına bölünerek Z-skoru hesaplanır. Z-skoru, veri noktalarının ortalama değere olan uzaklıklarını standart birimler (standart sapma) cinsinden ifade eder. Bu sayede, veri setindeki tüm bağımsız değişkenler aynı ölçek aralığına getirilmiştir.

$$z = \frac{x - \mu}{\sigma} \quad (3.2)$$

Dönüştürülmüş ve ölçeklendirilmiş veri seti üzerinde farklı makine öğrenimi algoritmaları uygulanmıştır. Uygulanan modeller arasında Lasso Regresyon, Ridge Regresyon, Linear Regresyon, Random Forest Regresyon, Destek Vektör Regresyonu (Support Vector Regression, DVR), yer almıştır. Her bir model eğitilmiş ve sonuçlar detaylı bir şekilde kaydedilmiştir. Bu adımda, modellerin doğruluğu ve

genellenebilirliği üzerine odaklanılmıştır.

Makine öğrenimi modellerinin performansını artırmak amacıyla hiperparametre optimizasyonu yapılmıştır. Bu işlem için GridSearchCV yöntemi kullanılmıştır. GridSearchCV, belirli hiperparametre değerlerini sistematik olarak deneyerek en uygun parametre kombinasyonlarını belirlemiştir. Her bir model için bu optimizasyon süreci gerçekleştirilmiş ve sonuçlar kaydedilmiştir.

Son adım olarak, hem ölçeklendirme hem de GridSearchCV yöntemleri bir arada kullanılarak makine öğrenimi modelleri yeniden eğitilmiştir. Bu yaklaşım, modellerin genel performansını artırmayı ve daha güvenilir sonuçlar elde etmeyi amaçlamıştır. Elde edilen nihai sonuçlar, karşılaştırma ve değerlendirme amacıyla kaydedilmiştir.

4. BULGULAR

4.1. Değerlendirme Ölçütleri

Bu çalışmada, öğrenci performansını tahmin etmek amacıyla farklı makine öğrenimi algoritmaları kullanılmış ve bu modellerin performansı çeşitli metriklerle değerlendirilmiştir. Değerlendirme metrikleri arasında (MAE), (RMSE) ve R^2 skoru yer almaktadır. MAE, modelin tahminlerinin gerçek değerlerden farklı olduğunu ölçen bir metriktir ve daha düşük bir MAE değeri, modelin daha doğru tahminler yaptığına işaret eder. RMSE de hata büyüklüğünü kare kök olarak ölçerken, R^2 skoru, modelin veriye ne kadar uyum sağladığını gösteren bir diğer önemli değerlendirme ölçütüdür.

Her bir modelin eğitim verisi üzerinde yapılan 10 katlı çapraz doğrulama ile elde edilen sonuçlar sunulmuştur. Bu yaklaşım, modelin genellenebilirliğini değerlendirirken daha güvenilir sonuçlar elde edilmesini sağlamıştır.

Modelin doğruluğunu değerlendiren ve varyansla paralel olarak kullanılan yaygın ölçüm grubu, regresyon modelinin eğitim verilerine ne kadar yakın olduğunu belirlemeye yönelik mesafe bazlı ölçütlerdir. Bu grubun temel üyelerinden ikisi, (MAE) ve (RMSE) gibi yöntemlerdir. Bu ölçütler, modelin tahminlerinin gerçek verilere ne kadar yakın olduğunu anlamamıza yardımcı olur. (Sammut & Webb, 2010a). Belirleme katsayısı (R^2), 1 ile sınırlıdır ve mükemmel uyumu ifade eder. R^2 'nin alt sınırı yoktur ve 0 değeri, tüm eğitim noktalarının hedef değerinin ortalamasını temsil eden yatay çizgiyle elde edilen uyumu gösterir. R^2 için negatif değerler, ortalama çizgiden daha kötü bir uyumu belirtir. Bu nedenle, R^2 'nin anlamlı aralığı 0 ile 1 arasında kabul edilir.

$$MAE = \frac{1}{m} \sum_{i=1}^m |x_i - y_i| \quad (4.1)$$

$$R^2 = \sum_{i=1}^m (y_i - \bar{y})^2 \quad (4.2)$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - y_i)^2} \quad (4.3)$$

İlk aşamada çalışma da kullanılan beş farklı veri seti üzerinde eksik değer analizi gerçekleştirilmiştir. İlk olarak, Veri seti 1'deki eksik değerler incelenmiş ve Teacher_Quality (Öğretmen Kalitesi), Parental_Education_Level (Ebeveyn Eğitim Seviyesi) ve Distance_from_Home (Ev-Okul Arası Mesafe) değişkenlerinde sırasıyla 78, 90 ve 67 adet eksik gözlem bulunduğu tespit edilmiştir. Diğer değişkenlerde eksik değer bulunmamaktadır. Eksik değerlerin giderilmesi amacıyla uygun yöntemler tercih edilmiş; sayısal değişkenler, veri setinin aşırı uç değerlerden etkilenmesini önlemek için medyan (median) değeri ile doldurulmuş, kategorik değişkenler ise en sık tekrar eden değer olan mod (mode) kullanılarak tamamlanmıştır. Bu işlemler sonucunda, Veri seti 1'de eksik değer kalmadığı belirlenmiştir. Veri seti 2, üzerinde gerçekleştirilen analizde ise herhangi bir eksik veri bulunmamış, tüm değişkenlerin eksiksiz olması analiz sürecinin güvenilirliği açısından önemli bir avantaj sağlamıştır. Benzer şekilde, Veri seti 3'de de eksik değer kontrolü yapılmış ve veri setinin tamamında herhangi bir eksik gözleme rastlanmadığı tespit edilmiştir. Ayrıca, Veri seti 4 ve Veri seti 5, üzerinde yapılan eksik değer analizleri sonucunda da eksik gözlem bulunmadığı görülmüştür. Böylece, çalışmada kullanılan veri setlerinde eksik değerlerin uygun yöntemlerle giderildiği ya da eksiksiz olduğu doğrulanarak analiz sürecinin güvenilirliği artırılmıştır.

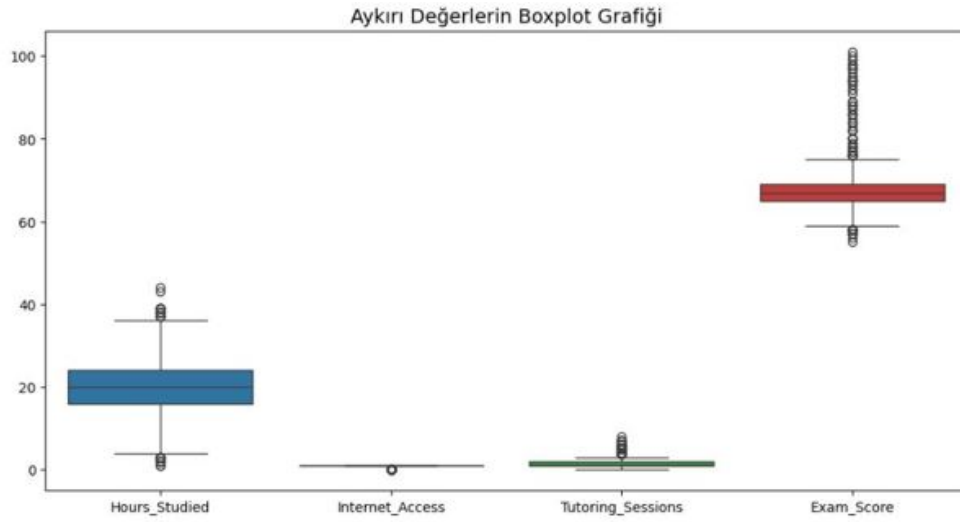
Daha sonra, veri setinde bulunan kategorik değişkenler, makine öğrenimi algoritmalarında kullanılabilir hale getirilmesi amacıyla Label Encoding yöntemi ile sayısal değerlere dönüştürülmüştür. Bu işlem sonrasında bağımlı ve bağımsız değişkenler belirlenmiştir.

Aykırı değerlerin belirlenmesi için Z-Skoru (Z-Score) yöntemi kullanılmıştır. Z-Skoru, her bir veri noktasının ortalamadan kaç standart sapma uzaklıkta olduğunu hesaplayarak aşırı uç değerleri tespit etmeye yardımcı olmaktadır.

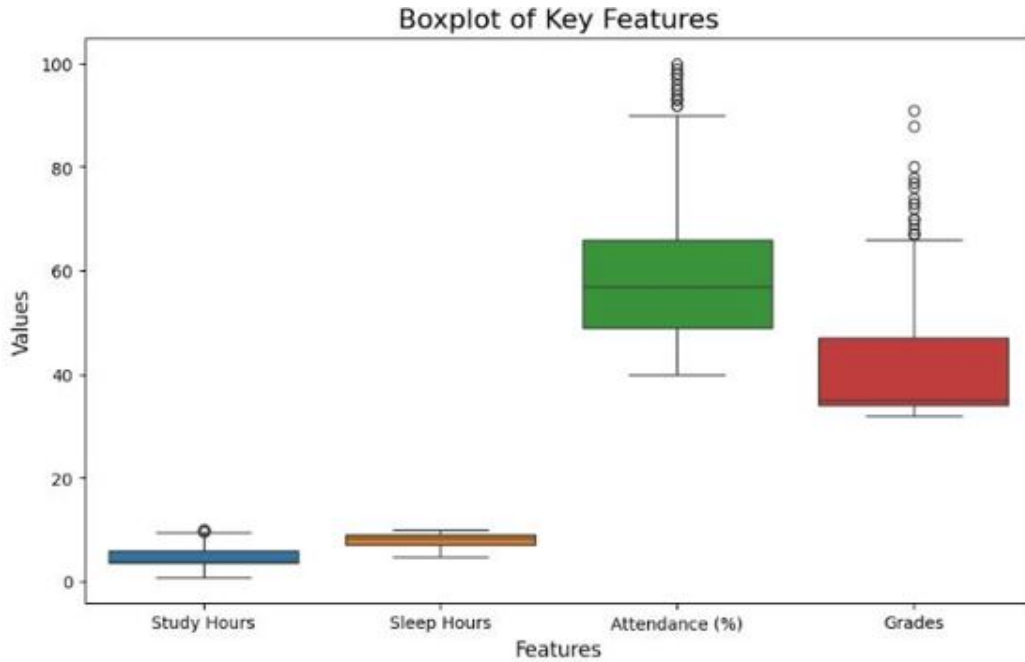
Veri setindeki sayısal değişkenlerin dağılımını incelemek ve olası aykırı değerleri görselleştirmek amacıyla kutu grafiği (boxplot) oluşturulmuştur. Bu sayede, verinin genel yapısı analiz edilmiş ve aykırı değerlerin varlığı görsel olarak değerlendirilmiştir. Tespit edilen aykırı değerlerin veri setinin genel dağılımını olumsuz etkilemesini önlemek amacıyla, bu değerlerin yerine ilgili sütunun medyanı atanmıştır. Medyan değeri, aşırı uç değerlerden etkilenmediği için tercih edilmiştir. Bu işlemler sonucunda, veri setindeki aykırı değerler ortadan kaldırılarak daha

dengeli bir dağılım elde edilmiştir.

Yapılan aykırı değer analizi sonucunda, Veri seti 1 ve Veri seti 3'te aykırı değerlere rastlanmıştır. Bu aykırı değerler, Şekil 4.1.ve Şekil 4.2.'de boxplot grafikleri ile görselleştirilmiştir. Diğer veri setleri üzerinde gerçekleştirilen analizlerde ise herhangi bir aykırı değere rastlanmamıştır.



Şekil 4.1. Veri seti 1, aykırı değer grafiği



Şekil 4.2. Veri seti 3, aykırı değer grafiği

Modelin doğruluğunu daha sağlam bir şekilde değerlendirmek amacıyla on

katlı çapraz doğrulama yöntemi kullanılmıştır. Çapraz doğrulama, modelin farklı veri alt kümelerinde test edilmesiyle genelleme yeteneğinin artırılmasını hedefler. Bu teknik, modelin eğitim verisine aşırı derecede uyum gösterme (overfitting) riskini en aza indirir, çünkü her adımda farklı eğitim ve test veri setleri oluşturulur. On katlı çapraz doğrulamada, verinin %90'lık kısmı eğitim için ayrılırken, geri kalan %10'luk bölüm test verisi olarak kullanılır ve bu süreç toplamda 10 kez tekrarlanır.

Çalışmanın devamında, öğrenci performansını tahmin etmek amacıyla çeşitli makine öğrenimi algoritmaları kullanılmıştır. Bu algoritmalar arasında Linear Regression, Ridge Regression, Lasso Regression, Destek Vektör Regresyonu (DVR), Random Forest yer almıştır. Her bir algoritma, model performansının karşılaştırılması ve en iyi sonuçların elde edilmesi amacıyla ayrı ayrı uygulanmıştır.

Modellerin daha sağlam ve genelleştirilebilir sonuçlar üretebilmesi için 10 katlı çapraz doğrulama yöntemi uygulanmıştır. Çizelge 4.1, Çizelge 4.2, Çizelge 4.3, Çizelge 4.4 ve Çizelge 4.5'de sonuçlar gösterilmiştir. Bu yöntemle veri seti farklı katmanlara ayrılarak, her model tüm katmanlarda ayrı ayrı eğitilip test edilmiştir. Bu süreçte, kullanılan algoritmalar olan Linear Regression, Ridge Regression, Lasso Regression, Destek Vektör Regresyonu (DVR), Random Forest uygulanmış ve modellerin performans sonuçları çapraz doğrulama ile analiz edilmiştir. Elde edilen sonuçlar, modellerin doğruluğunu ve güvenilirliğini değerlendirmek amacıyla kaydedilmiştir.

Çizelge 4.1' de Veri seti 1'e , 10 kat çapraz doğrulama yöntemiyle değerlendirilen model sonuçları gösterilmiştir.. Bu çizelgede, her bir katmanda modelin performans metrikleri (R^2 , MAE, RMSE vb.) hesaplanarak modelin genel başarısı analiz edilmiştir. Çapraz doğrulama, modelin farklı veri bölümlerindeki tutarlılığını ve genelleme kabiliyetini görmek için kullanılmıştır.

Çizelge 4.1. Veri seti 1, 10 Kat Çaprazlama Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.9269	0.5545	0.8901
Lasso Regression	0.7497	1.3010	1.5474
DVR	0.7982	1.1525	1.4792
Random Forest	0.8469	0.9801	1.2882
Linear Regression	0.9269	0.5545	0.8901
Ortalama	0.8497	0.9085	1.219

Çizelge 4.2' de Veri seti 2'ye, 10 kat çapraz doğrulama yöntemiyle

değerlendirilen model sonuçları gösterilmiştir.. Bu çizelgede, her bir katmanda modelin performans metrikleri (R^2 , MAE, RMSE vb.) hesaplanarak modelin genel başarısı analiz edilmiştir. Çapraz doğrulama, modelin farklı veri bölümlerindeki tutarlılığını ve genelleme kabiliyetini görmek için kullanılmıştır.

Çizelge 4.2. Veri seti 2, 10 Kat Çaprazlama Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.9368	3.8271	4.4291
Lasso Regression	0.7891	6.5703	8.0946
DVR	0.8847	4.8186	5.9829
Random Forest	0.9319	3.9364	4.5955
Linear Regression	0.9368	3.8271	4.4291
Ortalama	0.8959	4.5959	5.5062

Çizelge 4.3' de Veri seti 3'e, 10 kat çapraz doğrulama yöntemiyle değerlendirilen model sonuçları gösterilmiştir. Bu çizelgede, her bir katmanda modelin performans metrikleri (R^2 , MAE, RMSE vb.) hesaplanarak modelin genel başarısı analiz edilmiştir. Çapraz doğrulama, modelin farklı veri bölümlerindeki tutarlılığını ve genelleme kabiliyetini görmek için kullanılmıştır.

Çizelge 4.3. Veri seti 3, 10 Kat Çaprazlama Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.7665	4.5161	3.4742
Lasso Regression	0.6532	5.5256	4.1771
DVR	0.1980	8.4193	5.5912
Random Forest	0.9784	1.3689	0.9734
Linear Regression	0.7664	4.5161	3.4771
Ortalama	0.6725	4.8692	3.5386

Çizelge 4.4' de Veri seti 4'e , 10 kat çapraz doğrulama yöntemiyle değerlendirilen model sonuçları gösterilmiştir. Bu çizelgede, her bir katmanda modelin performans metrikleri (R^2 , MAE, RMSE vb.) hesaplanarak modelin genel başarısı analiz edilmiştir. Çapraz doğrulama, modelin farklı veri bölümlerindeki tutarlılığını ve genelleme kabiliyetini görmek için kullanılmıştır.

Çizelge 4.4. Veri seti 4, 10 Kat Çaprazlama Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9887	1.6179	2.0379
Lasso Regression	0.9868	1.7529	2.2050
DVR	0.9858	1.8142	2.2846
Random Forest	0.9851	1.8629	2.3435
Linear Regression	0.9887	1.6179	2.0379
Ortalama	0.9870	1.7332	2.1818

Çizelge 4.5' de Veri seti 5'e , 10 kat çapraz doğrulama yöntemiyle değerlendirilen model sonuçları gösterilmiştir. Bu çizelgede, her bir katmanda modelin performans metrikleri (R², MAE, RMSE vb.) hesaplanarak modelin genel başarısı analiz edilmiştir. Çapraz doğrulama, modelin farklı veri bölümlerindeki tutarlılığını ve genelleme kabiliyetini görmek için kullanılmıştır.

Çizelge 4.5. Veri seti 5, 10 Kat Çaprazlama Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9507	0.1623	0.1996
Lasso Regression	0.8274	0.3058	0.3785
DVR	0.9091	0.1930	0.2440
Random Forest	0.7797	0.2282	0.2930
Linear Regression	0.9507	0.1623	0.1996
Ortalama	0.8835	0.2103	0.2629

Çalışmanın bir sonraki aşamasında, veri setindeki değişkenler arasındaki ilişkileri değerlendirmek amacıyla korelasyon analizi yapılmış, yüksek korelasyona sahip özellikler belirlenip bazı veriler çoğaltılarak zenginleştirilmiştir. Düşük korelasyona sahip değişkenler ise analizden çıkarılmıştır.

Standartlaştırma işlemi, Z-skoru yöntemiyle gerçekleştirilmiş ve bu sayede veriler aynı ölçeğe getirilerek modelin performansı artırılmıştır. Sonrasında, makine öğrenimi modelleri çalıştırılmış ve elde edilen sonuçlar değerlendirilmiştir. Çizelge 4.6, Çizelge 4.7, Çizelge 4.8, Çizelge 4.9, ve Çizelge 4.10'da, standartlaştırma sonrası elde edilen model sonuçları sunulmuştur.

Çizelge 4.6' da Veri seti 1'e uygulanan Z-skor yöntemi ile standartlaştırılmış veriler kullanılarak elde edilen model sonuçları gösterilmiştir. R², MSE ve MAE gibi performans metrikleri hesaplanmıştır.

Çizelge 4.6. Veri seti 1, Standartlaştırma Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9183	0.8858	0.6176
Lasso Regression	0.5263	5.1368	1.8152
DVR	0.9237	0.8272	0.5944
Random Forest	0.8439	1.6922	0.9927
Linear Regression	0.9183	0.8858	0.6176
Ortalama	0.8261	1.8856	0.9275

Çizelge 4.7' de Veri seti 2'ye uygulanan Z-skör yöntemi ile standartlaştırılmış veriler kullanılarak elde edilen model sonuçları gösterilmiştir. R², MSE ve MAE gibi performans metrikleri hesaplanmıştır.

Çizelge 4.7. Veri seti 2, Standartlaştırma Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9368	3.8308	4.4337
Lasso Regression	0.9344	3.8799	4.5177
DVR	0.1900	13.6445	15.8788
Random Forest	0.9342	3.8994	4.5347
Linear Regression	0.9368	3.8308	4.4337
Ortalama	0.7864	5.8171	6.7597

Çizelge 4.8' de Veri seti 3'e uygulanan Z-skör yöntemi ile standartlaştırılmış veriler kullanılarak elde edilen model sonuçları gösterilmiştir. R², MSE ve MAE gibi performans metrikleri hesaplanmıştır.

Çizelge 4.8. Veri seti 3, Standartlaştırma Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.7660	3.4767	4.5138
Lasso Regression	0.7403	3.6220	4.7695
DVR	0.9031	1.7033	2.9128
Random Forest	0.9792	0.9557	1.3258
Linear Regression	0.7660	3.4768	4.5138
Ortalama	0.8309	2.6469	3.6071

Çizelge 4.9' da Veri seti 4'e uygulanan Z-skör yöntemi ile standartlaştırılmış veriler kullanılarak elde edilen model sonuçları gösterilmiştir. R², MSE ve MAE gibi performans metrikleri hesaplanmıştır.

Çizelge 4.9. Veri seti 4, Standartlaştırma Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9907	1.4694	1.8534
Lasso Regression	0.9824	1.9678	2.5483
DVR	0.9964	0.8819	1.1540
Random Forest	0.9997	0.1678	0.3187
Linear Regression	0.9907	1.4694	1.8534
Ortalama	0.9919	1.1912	1.5455

Çizelge 4.10' da Veri seti 5'e uygulanan Z-skor yöntemi ile standartlaştırılmış veriler kullanılarak elde edilen model sonuçları gösterilmiştir. R², MSE ve MAE gibi performans metrikleri hesaplanmıştır.

Çizelge 4.10. Veri seti 5, Standartlaştırma Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	1.0000	0.0016	0.0020
Lasso Regression	0.8274	0.3059	0.3786
DVR	0.9826	0.0777	0.1000
Random Forest	1.0000	0.0011	0.0019
Linear Regression	1.0000	0.0000	0.0000
Ortalama	0.9620	0.0773	0.0965

Daha sonra, GridSearchCV kullanılarak hiperparametre optimizasyonu gerçekleştirilmiştir. Bu işlem, farklı hiperparametre kombinasyonlarının sistematik bir şekilde denenmesi ile en iyi model parametrelerinin belirlenmesini sağlamıştır. Tüm makine öğrenimi algoritmaları üzerinde bu optimizasyon yöntemi uygulanmış ve her bir modelin en iyi parametre seti ile sonuçlar elde edilmiştir. Hiperparametre optimizasyonu, modellerin daha verimli ve doğru sonuçlar üretmesini sağlamak amacıyla yapılmıştır. Çizelge 4.11, Çizelge 4.12, Çizelge 4.13, Çizelge 4.14 ve Çizelge 4.15' de veri setlerine uygulanan GridsearchCV sonrası model sonuçları gösterilmiştir.

Çizelge 4.11' de Veri seti 1'e uygulanan hiperparametre optimizasyonunun GridSearchCV yöntemi ile yapıldığı model sonuçları gösterilmiştir. Bu yöntemle farklı hiperparametre kombinasyonları test edilmiş ve en iyi performansı veren parametreler belirlenmiştir. Çizelgede, optimize edilen modelin R², MSE ve MAE gibi performans metrikleri yer almaktadır.

Çizelge 4.11. Veri seti 1, GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9080	0.6484	1.0346
Lasso Regression	0.9040	0.6821	1.0571
DVR	0.9009	0.5975	1.0737
Random Forest	0.9552	0.5556	0.7220
Linear Regression	0.9080	0.6484	1.0346
Ortalama	0.9152	0.6264	0.9844

Çizelge 4.12’ de Veri seti 2’ye uygulanan hiperparametre optimizasyonunun GridSearchCV yöntemi ile yapıldığı model sonuçları gösterilmiştir. Bu yöntemle farklı hiperparametre kombinasyonları test edilmiş ve en iyi performansı veren parametreler belirlenmiştir. Çizelgede, optimize edilen modelin R², MSE ve MAE gibi performans metrikleri yer almaktadır.

Çizelge 4.12. Veri seti 2, GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9371	4.4252	3.8225
Lasso Regression	0.9380	3.7731	4.3953
DVR	0.9106	4.3548	5.2702
Random Forest	0.9350	3.8420	5.2702
Linear Regression	0.9380	3.7783	4.3973
Ortalama	0.9317	4.0347	4.6311

Çizelge 4.13’ de Veri seti 3’e uygulanan hiperparametre optimizasyonunun GridSearchCV yöntemi ile yapıldığı model sonuçları gösterilmiştir. Bu yöntemle farklı hiperparametre kombinasyonları test edilmiş ve en iyi performansı veren parametreler belirlenmiştir. Çizelgede, optimize edilen modelin R², MSE ve MAE gibi performans metrikleri yer almaktadır.

Çizelge 4.13. Veri seti 3, GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.7743	3.4567	4.4964
Lasso Regression	0.7725	3.4542	4.4966
DVR	0.7571	3.3532	4.6643
Random Forest	0.9940	0.5378	0.7358
Linear Regression	0.7743	3.4567	4.4964
Ortalama	0.8144	2.8517	3.7779

Çizelge 4.14’ de Veri seti 4’e uygulanan hiperparametre optimizasyonunun

GridSearchCV yöntemi ile yapıldığı model sonuçları gösterilmiştir. Bu yöntemle farklı hiperparametre kombinasyonları test edilmiş ve en iyi performansı veren parametreler belirlenmiştir. Çizelgede, optimize edilen modelin R^2 , MSE ve MAE gibi performans metrikleri yer almaktadır.

Çizelge 4.14. Veri seti 4, GridSearchCV Sonrası Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.9907	1.4686	1.8526
Lasso Regression	0.9905	1.4856	1.8730
DVR	0.9979	0.7006	0.8842
Random Forest	0.9984	0.5725	0.7567
Linear Regression	0.9907	1.4686	1.8526
Ortalama	0.9936	1.1394	1.2436

Çizelge 4.15' de Veri seti 5'e uygulanan hiperparametre optimizasyonunun GridSearchCV yöntemi ile yapıldığı model sonuçları gösterilmiştir. Bu yöntemle farklı hiperparametre kombinasyonları test edilmiş ve en iyi performansı veren parametreler belirlenmiştir. Çizelgede, optimize edilen modelin R^2 , MSE ve MAE gibi performans metrikleri yer almaktadır.

Çizelge 4.15. Veri seti 5, GridSearchCV Sonrası Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	1.0000	0.0001	0.0002
Lasso Regression	0.9052	0.2276	0.2818
DVR	0.9999	0.0049	0.0071
Random Forest	1.0000	0.0021	0.0028
Linear Regression	1.0000	0.0000	0.0000
Ortalama	0.9810	0.0471	0.0584

Son işlem olarak, standartlaştırma ve hiperparametre optimizasyonu (GridSearchCV) sırasıyla uygulanmıştır. İlk olarak, veri seti üzerinde standartlaştırma işlemi gerçekleştirilmiş, ardından GridSearchCV kullanılarak modellerin hiperparametreleri optimize edilmiştir. Bu işlemler, birbirini takip eden adımlar olarak tek bir kod bloğunda gerçekleştirilmiş, her iki işlem de model performansını iyileştirmek amacıyla uygulanmıştır. Standartlaştırma ile veriler uyumlu hale getirilmiş ve GridSearchCV ile en iyi hiperparametre kombinasyonları elde edilmiştir. Sonuç olarak, bu iki işlemde elde edilen parametrelerle model sonuçları tekrar gözden geçirilmiştir. Çizelge 4.16, Çizelge 4.17, Çizelge 4.18, Çizelge 4.19 ve Çizelge 4.20'de hem standartlaştırma hem de GridSearchCV uygulanmış model sonuçları gösterilmektedir.

Çizelge 4.16’ da Veri seti 1’e, Z-skör yöntemi ile standartlaştırma ve GridSearchCV ile hiperparametre optimizasyonunun birlikte uygulanmasıyla elde edilen model sonuçlar gösterilmiştir. Bu süreçte modelin R^2 , RMSE ve MAE değerleri hesaplanmıştır.

Çizelge 4.16. Veri seti 1, Standartlaştırma ve GridSearchCV Sonrası Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.9080	0.6484	1.0346
Lasso Regression	0.9070	0.6660	1.0406
DVR	0.9011	0.5971	1.0727
Random Forest	0.9553	0.5586	0.7211
Linear Regression	0.9080	0.6484	1.0346
Ortalama	0.9159	0.6237	0.9807

Çizelge 4.17’ de Veri seti 2’ye, Z-skör yöntemi ile standartlaştırma ve GridSearchCV ile hiperparametre optimizasyonunun birlikte uygulanmasıyla elde edilen model sonuçlar gösterilmiştir. Bu süreçte modelin R^2 , RMSE ve MAE değerleri hesaplanmıştır.

Çizelge 4.17. Veri seti 2, Standartlaştırma ve GridSearchCV Sonrası Model Sonuçları

Model	R^2 Score	MAE	RMSE
Ridge Regression	0.9404	3.6802	4.2788
Lasso Regression	0.9405	3.6679	4.2759
DVR	0.9407	3.6525	4.2676
Random Forest	0.9352	3.7783	4.3973
Linear Regression	0.9379	3.7783	4.3973
Ortalama	0.9469	3.7114	4.3234

Çizelge 4.18’ de Veri seti 3’e, Z-skör yöntemi ile standartlaştırma ve GridSearchCV ile hiperparametre optimizasyonunun birlikte uygulanmasıyla elde edilen model sonuçlar gösterilmiştir. Bu süreçte modelin R^2 , RMSE ve MAE değerleri hesaplanmıştır.

Çizelge 4.18. Veri seti 3, Standartlaştırma ve GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.7660	3.4767	4.5137
Lasso Regression	0.7656	3.4702	4.5200
DVR	0.9675	1.0969	1.6735
Random Forest	0.9799	0.9171	1.3184
Linear Regression	0.7660	3.4768	4.5138
Ortalama	0.9489	2.4875	3.3079

Çizelge 4.19’ da Veri seti 4’e, Z-skör yöntemi ile standartlaştırma ve GridSearchCV ile hiperparametre optimizasyonunun birlikte uygulanmasıyla elde edilen model sonuçlar gösterilmiştir. Bu süreçte modelin R², RMSE ve MAE değerleri hesaplanmıştır.

Çizelge 4.19. Veri seti 4, Standartlaştırma ve GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	0.9907	1.4686	1.8526
Lasso Regression	0.9905	1.4839	1.8743
DVR	0.9999	0.0762	0.1858
Random Forest	0.9985	0.5714	0.7529
Linear Regression	0.9907	1.4686	1.8526
Ortalama	0.9941	1.0137	1.3036

Çizelge 4.20’ de Veri seti 5’e, Z-skör yöntemi ile standartlaştırma ve GridSearchCV ile hiperparametre optimizasyonunun birlikte uygulanmasıyla elde edilen model sonuçlar gösterilmiştir. Bu süreçte modelin R², RMSE ve MAE değerleri hesaplanmıştır.

Çizelge 4.20. Veri seti 5, Standartlaştırma ve GridSearchCV Sonrası Model Sonuçları

Model	R ² Score	MAE	RMSE
Ridge Regression	1.0000	0.0001	0.0001
Lasso Regression	0.9881	0.0844	0.1000
DVR	1.0000	0.0047	0.0056
Random Forest	1.0000	0.0022	0.0028
Linear Regression	1.0000	0.0000	0.0000
Ortalama	0.9976	0.0183	0.0217

İlk olarak, 10 kat çapraz doğrulama ve standartlaştırma işlemleri gerçekleştirilmiştir. Ardından, GridSearchCV ile hiperparametre optimizasyonu yapılmıştır. Son olarak, hem standartlaştırma hem de GridSearchCV birlikte uygulanarak elde edilen sonuçların ortalamaları R², MAE ve RMSE metrikleriyle

Çizelge 4.21, Çizelge 4.22, Çizelge 4.23, Çizelge 4.24 ve Çizelge 4.25'de sunulmuştur.

Çizelge 4.21' de Veri seti 1'e , 10 kat çapraz doğrulama, standartlaştırma, GridSearchCV ile hiperparametre optimizasyonu ve standartlaştırma ile GridSearchCV'nin birlikte uygulanmasıyla elde edilen sonuçların ortalamaları gösterilmiştir. Çizelgede, bu yöntemlerin ortalama R^2 , RMSE ve MAE değerleri sunulmaktadır.

Çizelge 4.21. Veri seti 1, Önerilen Yöntemin Ortalama Sonuçları

Ortalama	R^2 Score	MAE	RMSE
10 Kat Çapraz Doğrulama	0.8497	0.9085	1.219
Standartlaştırma	0.8261	1.8856	0.9275
GridsearchCV	0.9152	0.6264	0.9844
Standartlaştırma+GridsearchCV	0.9159	0.6237	0.9807

Çizelge 4.22' de Veri seti 2'ye , 10 kat çapraz doğrulama, standartlaştırma, GridSearchCV ile hiperparametre optimizasyonu ve standartlaştırma ile GridSearchCV'nin birlikte uygulanmasıyla elde edilen sonuçların ortalamaları gösterilmiştir. Çizelgede, bu yöntemlerin ortalama R^2 , RMSE ve MAE değerleri sunulmaktadır.

Çizelge 4.22. Veri seti 2, Önerilen Yöntemin Ortalama Sonuçları

Ortalama	R^2 Score	MAE	RMSE
10 Kat Çapraz Doğrulama	0.8959	4.5959	5.5062
Standartlaştırma	0.7864	5.8171	6.7597
GridsearchCV	0.9317	4.0347	4.6311
Standartlaştırma+GridsearchCV	0.9469	3.7114	4.3234

Çizelge 4.23' de Veri seti 3'e , 10 kat çapraz doğrulama, standartlaştırma, GridSearchCV ile hiperparametre optimizasyonu ve standartlaştırma ile GridSearchCV'nin birlikte uygulanmasıyla elde edilen sonuçların ortalamaları gösterilmiştir. Çizelgede, bu yöntemlerin ortalama R^2 , RMSE ve MAE değerleri sunulmaktadır.

Çizelge 4.23. Veri seti 3, Önerilen Yöntemin Ortalama Sonuçları

Ortalama	R ² Score	MAE	RMSE
10 Kat Çapraz Doğrulama	0.6725	4.8692	3.5386
Standartlaştırma	0.8309	2.6469	3.6071
GridsearchCV	0.8144	2.8517	3.7779
Standartlaştırma+GridsearchCV	0.9489	2.4875	3.3079

Çizelge 4.24' de Veri seti 4'e , 10 kat çapraz doğrulama, standartlaştırma, GridSearchCV ile hiperparametre optimizasyonu ve standartlaştırma ile GridSearchCV'nin birlikte uygulanmasıyla elde edilen sonuçların ortalamaları gösterilmiştir. Çizelgede, bu yöntemlerin ortalama R², RMSE ve MAE değerleri sunulmaktadır.

Çizelge 4.24. Veri seti 4, Önerilen Yöntemin Ortalama Sonuçları

Ortalama	R ² Score	MAE	RMSE
10 Kat Çapraz Doğrulama	0.9870	1.7332	2.1818
Standartlaştırma	0.9919	1.1912	1.5455
GridsearchCV	0.9936	1.1394	1.2436
Standartlaştırma+GridsearchCV	0.9941	1.0137	1.3036

Çizelge 4.25' de Veri seti 5'e, 10 kat çapraz doğrulama, standartlaştırma, GridSearchCV ile hiperparametre optimizasyonu ve standartlaştırma ile GridSearchCV'nin birlikte uygulanmasıyla elde edilen sonuçların ortalamaları gösterilmiştir. Çizelgede, bu yöntemlerin ortalama R², RMSE ve MAE değerleri sunulmaktadır.

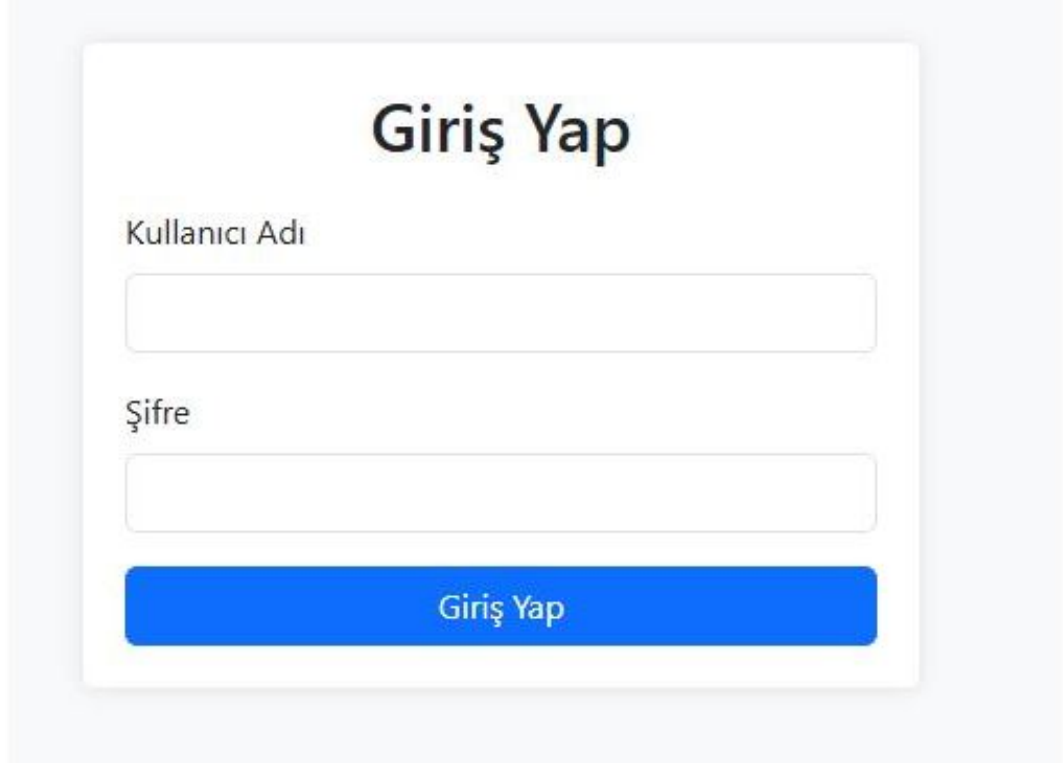
Çizelge 4.25. Veri seti 5, Önerilen Yöntemin Ortalama Sonuçları

Ortalama	R ² Score	MAE	RMSE
10 Kat Çapraz Doğrulama	0.8835	0.2103	0.2629
Standartlaştırma	0.9620	0.0773	0.0965
GridsearchCV	0.9810	0.0471	0.0584
Standartlaştırma+GridsearchCV	0.9976	0.0183	0.0217

4.2. Önerilen Yöntemin Web Üzerinde Uygulanması

Önerilen Yöntemin uygulanmasından sonra ise çalışmanın uygulanabilirliğini artırmak amacıyla bir web uygulaması geliştirilmiştir. Bu web uygulaması, Python ve React kullanılarak gerçekleştirilmiştir. Kullanıcının bir öğretmen olduğu varsayılarak, sisteme bir öğretmen kullanıcı adı ve şifresi ile giriş yapması sağlanmıştır. Giriş yapıldıktan sonra kullanıcı, dosya yükleme sayfasına

yönlendirilmekte ve CSV formatında bir dosya yükleyerek öğrencilerin akademik performanslarını tahmin edebilmektedir. Şekil 4.3.'de kullanıcının kullanıcı adı ve şifresiyle giriş yapabileceği sayfa gösterilmiştir.

The image shows a login form titled "Giriş Yap" (Login). It contains two input fields: "Kullanıcı Adı" (Username) and "Şifre" (Password). Below the password field is a blue button labeled "Giriş Yap". The form is centered on a light blue background.

Şekil 4.3. Kullanıcının Dosya Yükleyebildiği ve Analiz Edebildiği Sayfa

Dosya yükleme işlemi tamamlandığında, sistem yüklenen veriyi işleyecek ve daha önceden eğitilmiş makine öğrenmesi modelleri kullanarak tahmin işlemini gerçekleştirecektir. Veri setinde bulunan bağımsız değişkenler (ders çalışma süreleri, uyku saatleri, devamsızlık gibi faktörler), modele girdi olarak verilecek ve bağımlı değişken olan akademik performans, yani öğrencinin sınav notu tahmin edilecektir. Şekil 4.4. de doğru kullanıcı adı ve şifresiyle giriş yapan kullanıcının yönlendirildiği dosya yükleyebildiği ve analiz işlemini gerçekleştirdiği sayfa gösterilmiştir.

Şekil 4.4. Kullanıcının Dosya Yükleyebildiği ve Analiz Edebildiği Sayfa

Tahmin işlemi tamamlandıktan sonra, kullanıcıya analiz sonuçları sunulacaktır. Bu sonuçlar, ilgili faktörlerin öğrencinin sınav notu üzerindeki etkisini gösteren ayrıntılı bir rapor şeklinde sunulacaktır. Kullanıcı dostu bir arayüz ile tasarlanan bu sayfada, tahmin edilen notlar gerçek değerlerle karşılaştırılarak modelin doğruluğu gözlemlenebilecektir. Ayrıca, tahmin edilen değerler grafikler ve tablolar aracılığıyla görselleştirilecek, öğretmenlerin karar alma sürecine katkı sağlayacak detaylı istatistikler sunulacaktır. Şekil 4.5. de analiz sonuçları gösterilmektedir.

Analiz Sonuçları

Başarılı! CSV dosyası analiz edildi ve tahminler oluşturuldu.

socioeconomic score	study hours	sleep hours	attendance (%)	grades	Predicted_Grades
0.95822	3.4	8.2	53.0	nan	47.28
0.85566	3.2	5.9	55.0	nan	34.91
0.68025	3.2	9.3	41.0	nan	32.68
0.25936	3.2	8.2	47.0	nan	34.36
0.60447	3.8	10.0	75.0	nan	33.17
0.9832	3.4	9.0	47.0	nan	50.74
0.56648	7.9	8.1	63.0	nan	54.03
0.93487	1.4	8.0	47.0	nan	35.23
0.4666	5.4	8.8	67.0	nan	39.11

Şekil 4.5. Kullanıcının Öğrenci Notunu Tahmin Edebildiği Arayüz

5. TARTIŞMA

Adı	Yıl	Veri Seti	Kullanılan Algoritmalar	En İyi Sonuç Veren Algoritma
Can, Özdil ve ark.	2018	Devlet üniversitesinden alınan 5820 öğrencinin ders değerlendirme anketleri	Lojistik Regresyon	Lojistik Regresyon (%91.6 doğruluk)
Zilyas ve Yılmaz	2023	8. sınıf öğrencilerinden 13 soruluk anket verisi (519 öğrenci)	Rastgele Orman, Çoklu Doğrusal Regresyon, KNN, Karar Ağaçları, Torbalanmış Ağaçlar, Gradyen Artırıcı Regresyon	Rastgele Orman (%98 R ² , %88 doğruluk)
Güvenç, Sakal ve ark.	2022	71 üniversite öğrencisinin bilgisayar becerileri ve başarıları	KNN, SVM, Rastgele Orman, Karar Ağaçları, Lojistik Regresyon, Naive Bayes	KNN ve SVM (%97.66 doğruluk)
Aghalarova ve Keser	2021	Ortaokul öğrencilerinin Matematik ve Portekizce başarıları	Yapay Sinir Ağları (YSA)	Yapay Sinir Ağları (%94.8 doğruluk)

Şekil 5.1. Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları

Literatürde, öğrenci başarılarını tahmin etmek amacıyla çeşitli makine öğrenmesi tekniklerinin kullanıldığı birçok çalışma bulunmaktadır. Bu çalışmalarda genellikle öğrenci anketleri, akademik performans verileri ve bireysel özellikler analiz edilmiştir.

Can, Özdil ve arkadaşları (2018), bir devlet üniversitesinden elde edilen 5820 öğrencinin ders değerlendirme anketlerini kullanarak lojistik regresyon modeli ile bir analiz gerçekleştirmiştir. Bu çalışmada modelin doğruluk oranı %91.6 olarak raporlanmıştır.

Zilyas ve Yılmaz (2023) ise 8. sınıf öğrencilerine uygulanan 13 soruluk bir anketten elde edilen 519 öğrencilik veri seti üzerinde çalışmıştır. Kullanılan yöntemler arasında Rastgele Orman, Çoklu Doğrusal Regresyon, K-En Yakın Komşu (KNN), Karar Ağaçları, Torbalanmış Ağaçlar ve Gradyen Artırıcı Regresyon yer almaktadır. En başarılı sonuç, %98 R² ve %88 doğruluk oranıyla Rastgele Orman algoritmasıyla elde edilmiştir.

Güvenç, Sakal ve arkadaşları (2022), üniversite öğrencilerinin bilgisayar

becerileri ile akademik başarıları arasındaki ilişkiyi incelemiştir. Bu çalışmada KNN, Destek Vektör Makineleri (SVM), Rastgele Orman, Karar Ağaçları, Lojistik Regresyon ve Naive Bayes gibi algoritmalar uygulanmıştır. En yüksek doğruluk oranı %97.66 ile KNN ve SVM modelleriyle elde edilmiştir.

Aghalarova ve Keser (2021) tarafından yapılan çalışmada ise ortaokul öğrencilerinin Matematik ve Portekizce derslerindeki başarıları incelenmiş ve bu amaçla Yapay Sinir Ağları (YSA) kullanılmıştır. Model, %94.8 doğruluk oranıyla başarılı bir performans göstermiştir.

Adak ve Duralioğlu	2023	87 meslek lisesi öğrencisinin Nesne Tabanlı Programlama ders verisi	Doğrusal Regresyon, KNN, Karar Ağaçları	KNN
Alkan	2024	Kamuya açık veri seti (1000 öğrenci)	KNN, Karar Ağaçları, Naive Bayes, Rastgele Orman, DVM, Lojistik Regresyon, LDA	DVM (%67 doğruluk)
Bakan ve Kanbay	2024	UCI "Higher Education Students Performance Evaluation" veri seti (145 öğrenci)	KNN, Naive Bayes, Rastgele Orman, DVM, Karar Ağaçları, Boosting	Rastgele Orman (%96 doğruluk)
Er ve ark.	2023	8. sınıf öğrencilerinin LGS 2023 puanları (29 özellik)	PSO, ABA, GKAA, SVR, GPR, Karar Ağaçları	GKAA + SVR ($R^2 = 0.866$)
Deepika ve Sathyanarayana	2018	UCI Öğrenci Performansı & Kaggle Kalboard 360 E-Öğrenme	Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları, SVM, Naive Bayes	RF (UCI: %94 doğruluk), ANN (Kaggle: %79.1 doğruluk)

Şekil 5.2. Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları

Adak ve Duralioğlu (2023), 87 meslek lisesi öğrencisinin Nesne Tabanlı Programlama dersi verilerini kullanarak öğrenci başarısını tahmin etmeye yönelik bir çalışma gerçekleştirmiştir. Bu çalışmada Doğrusal Regresyon, K-En Yakın Komşu (KNN) ve Karar Ağaçları yöntemleri uygulanmış; en yüksek doğruluk oranı KNN algoritmasıyla elde edilmiştir.

Alkan (2024), kamuya açık bir veri seti üzerinde (1000 öğrenci) KNN, Karar Ağaçları, Naive Bayes, Rastgele Orman, Destek Vektör Makineleri (DVM), Lojistik

Regresyon ve Doğrusal Ayrım Analizi (LDA) algoritmalarını kullanarak analizler yapmıştır. Bu çalışmada en iyi sonuç %67 doğruluk oranıyla DVM modeli tarafından sağlanmıştır.

Bakan ve Kanbay (2024), UCI veri tabanından alınan “Higher Education Students Performance Evaluation” veri seti (145 öğrenci) ile gerçekleştirdikleri çalışmada KNN, Naive Bayes, Rastgele Orman, DVM, Karar Ağaçları ve Boosting algoritmalarını karşılaştırmıştır. En başarılı sonuç, %96 doğruluk oranı ile Rastgele Orman algoritmasıyla elde edilmiştir.

Er ve arkadaşları (2023), 8. sınıf öğrencilerinin LGS 2023 puanlarına ilişkin 29 özellik içeren veri seti üzerinde PSO, ABA, GKAA, SVR, GPR ve Karar Ağaçları algoritmalarını karşılaştırmıştır. Çalışmanın sonuçlarına göre en yüksek başarı, GKAA ve SVR algoritmalarının birlikte kullanımı ile elde edilmiş ve modelin R^2 değeri 0.866 olarak raporlanmıştır.

Deepika ve Sathyanarayana (2018), UCI Öğrenci Performansı veri seti ve Kaggle'dan alınan Kalboard 360 E-öğrenme veri setini kullanarak iki farklı ortamda analiz gerçekleştirmiştir. Bu çalışmada Karar Ağaçları, Rastgele Orman, Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (SVM) ve Naive Bayes algoritmaları değerlendirilmiştir. UCI veri setinde %94 doğruluk ile Rastgele Orman en iyi sonucu verirken, Kaggle veri setinde %79.1 doğruluk oranı ile YSA öne çıkmıştır.

Bu çalışmalar, hem farklı veri setleri hem de çeşitli algoritmaların kullanımı açısından çeşitlilik göstermekte olup, farklı öğrenme düzeylerinde öğrenci başarısını etkileyen değişkenlerin makine öğrenmesi modelleriyle etkili şekilde analiz edilebileceğini göstermektedir.

Salal, Abdullaev ve ark.	2019	UCI Öğrenci Performansı (649 öğrenci, 33 özellik)	Naive Bayes, Karar Ağaçları (J48), Rastgele Orman, REPTree, OneR, Basit Lojistik Regresyon, ZeroR	REPTree ve OneR (%76.73 doğruluk)
Suleiman, Okunade ve ark.	2024	Kaggle "Student Performance Dataset" (10.000 öğrenci)	Doğrusal Regresyon	Doğrusal Regresyon ($R^2 = 0.98$)
Chenrui Pei	2024	Kaggle (2.392 lise öğrencisi)	Çoklu Doğrusal Regresyon	Çoklu Doğrusal Regresyon ($R^2 = 0.954$)
Staneviciene ve ark.	2024	Moodle e-öğrenme platformu verisi	Karar Ağaçları, Bayes Sınıflandırıcı, KNN, SVM, Rastgele Orman	Rastgele Orman ($F1 = 0.873$, %81 doğruluk)

Şekil 5.3. Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları

Salal, Abdullaev ve arkadaşları (2019), UCI Öğrenci Performansı veri setini kullanarak gerçekleştirdikleri çalışmalarında 649 öğrenciden elde edilen ve 33 özelliği içeren veriler üzerinde analiz yapmıştır. Naive Bayes, Karar Ağaçları (J48), Rastgele Orman, REPTree, OneR, Basit Lojistik Regresyon ve ZeroR gibi çeşitli makine öğrenmesi algoritmalarını karşılaştıran araştırmada, en iyi doğruluk oranı %76.73 ile REPTree ve OneR algoritmaları tarafından elde edilmiştir.

Suleiman, Okunade ve arkadaşları (2024), Kaggle üzerinden sağlanan "Student Performance Dataset" adlı veri setiyle (10.000 öğrenci) gerçekleştirdikleri çalışmada yalnızca Doğrusal Regresyon yöntemini kullanmıştır. Modelin başarımı oldukça yüksek olup, R^2 değeri 0.98 olarak raporlanmıştır.

Chenrui Pei (2024), yine Kaggle üzerinden elde edilen 2.392 lise öğrencisine ait veriler ile Çoklu Doğrusal Regresyon modeli uygulamıştır. Çalışmada modelin performansı yüksek bulunmuş, R^2 değeri 0.954 olarak elde edilmiştir.

Staneviciene ve arkadaşları (2024) ise Moodle tabanlı bir e-öğrenme platformundan alınan verilerle gerçekleştirdikleri çalışmalarında Karar Ağaçları, Bayes Sınıflandırıcı, KNN, Destek Vektör Makineleri (SVM) ve Rastgele Orman algoritmalarını değerlendirmiştir. En başarılı sonuç %81 doğruluk ve 0.873 F1 skoru ile Rastgele Orman algoritmasında gözlemlenmiştir.

Dubey ve ark.	2024	2013-2020 ONS verisi (engelli ve engelsiz bireylerin eğitimi)	Doğrusal Regresyon, Rastgele Orman, Karar Ağaçları, SVM, XGBoost	Rastgele Orman (%90 doğruluk)
Jingbo ve Yujie	2024	Üniversite öğrencilerinin akademik başarısızlık riski verisi	ANN, Karar Ağaçları, SVM	Karar Ağaçları (%94 doğruluk)
Rabelo ve ark.	2024	Moodle ve diğer e-öğrenme platformlarından toplanan veriler	Karar Ağaçları, Kümeleme Teknikleri, Tahminsel Modeller	Rastgele Orman

Şekil 5.4. Önceki Çalışmalarda Kullanılan Veri Setleri ve Sonuçları

Dubey ve arkadaşları (2024), 2013-2020 yılları arasında ONS tarafından yayımlanan verilerden yararlanarak, engelli ve engelsiz bireylerin eğitim düzeylerini karşılaştırmaya yönelik bir çalışma gerçekleştirmiştir. Doğrusal Regresyon, Rastgele Orman, Karar Ağaçları, Destek Vektör Makineleri (SVM) ve XGBoost algoritmaları kullanılarak yapılan analiz sonucunda, en yüksek doğruluk oranı %90 ile Rastgele Orman modeli tarafından elde edilmiştir.

Jingbo ve Yujie (2024), üniversite öğrencilerinin akademik başarısızlık risklerini tahmin etmeye yönelik gerçekleştirdikleri çalışmalarında, Yapay Sinir Ağları (ANN), Karar Ağaçları ve SVM algoritmalarını kullanmıştır. Çalışmada en iyi performans %94 doğruluk oranı ile Karar Ağaçları modeli tarafından sağlanmıştır. Bu sonuç, karar ağaçlarının eğitim başarısızlığı riski gibi karmaşık yapıları çıktılarını modellemede etkin bir yöntem olduğunu göstermektedir.

Rabelo ve arkadaşları (2024) tarafından yapılan bir başka çalışmada ise Moodle ve diğer e-öğrenme platformlarından toplanan veriler analiz edilmiştir. Araştırmada Karar Ağaçları, Kümeleme Teknikleri ve çeşitli tahminsel modeller değerlendirilmiş; en iyi sonuç Rastgele Orman algoritması ile elde edilmiştir. Bu durum, çevrim içi öğrenme ortamlarında öğrenci davranışlarını ve başarılarını tahmin etmek için Rastgele Orman yönteminin etkili bir yaklaşım olduğunu göstermektedir.

Veri Seti	Yıl	Veri Seti İçeriği	Kullanılan Algoritmalar	En İyi Sonuç Veren Algoritma
Veri seti 1	2025	Kaggle'dan alınan, 6607 gözlem, 20 değişken	Linear Regression, Ridge Regression, LASSO, DVR, Random Forest	Random Forest (%91.59 doğruluk)
Veri seti 2	2025	Kaggle'dan alınan, 10.000 gözlem, 15 değişken	Linear Regression, Ridge Regression, LASSO, DVR, Random Forest	DVR ($R^2 = 0.9469$)
Veri seti 3	2025	Kaggle'dan alınan, 1388 gözlem, 5 değişken	Linear Regression, Ridge Regression, LASSO, DVR, Random Forest	Random Forest ($R^2 = 0.9940$)
Veri seti 4	2025	Kaggle'dan alınan, 10.000 gözlem, 6 değişken	Linear Regression, Ridge Regression, LASSO, DVR, Random Forest	DVR ($R^2 = 0.9999$)
Veri seti 5	2025	Kaggle'dan alınan, 2392 gözlem, 15 değişken	Linear Regression, Ridge Regression, LASSO, DVR, Random Forest	Linear Regression ($R^2 = 1.0000$)

Şekil 5.5. Önerilen Yönteme Göre Çalışılan Veri setleri ve Sonuçları

Bu çalışmanın bulguları, daha önceki araştırmalar ile karşılaştırıldığında anlamlı farklılıklar ortaya koymakta ve mevcut literatüre önemli katkılar sunmaktadır. Önceki çalışmalarda genellikle tek bir veri seti üzerinde analizler gerçekleştirilirken, bu çalışmada farklı veri setleri kullanılarak model performansları çok daha kapsamlı bir şekilde değerlendirilmiştir.

Önceki çalışmaların genelinde, sınırlı sayıda veri seti kullanılmış ve algoritma performansları çoğunlukla tek bir model üzerinden değerlendirilmiştir. Ancak, bu çalışmada, 10.000 gözlemlili büyük ölçekli veri setlerini de içeren geniş bir örneklem üzerinde analizler gerçekleştirilmiş, farklı makine öğrenmesi algoritmalarının doğruluk oranları ve performans metrikleri karşılaştırılmıştır. Yapılan analizlerde, Rastgele Orman algoritmasının Veri seti 1 ve Veri seti 3'te en başarılı model olduğu belirlenirken, Destek Vektör Regresyonu (DVR) algoritmasının Veri seti 2 ve Veri seti 4'de en yüksek performansa ulaştığı tespit edilmiştir. Linear Regression algoritması ise Veri seti 5'de ise 1.0000 R^2 değeri ile en başarılı sonucu vermiştir. Bu durum, veri setlerinin özelliklerine bağlı olarak farklı algoritmaların daha başarılı

olabileceğini göstermekte ve model seçiminin başarı tahmini açısından kritik bir faktör olduğunu ortaya koymaktadır.

Benzer şekilde, Ibrahim Baba Suleiman ve arkadaşları (2024) tarafından Kaggle platformundan elde edilen 10.000 öğrenciye ait veri seti kullanılarak gerçekleştirilen çalışmada, akademik performansın tahmini amacıyla doğrusal regresyon modeli uygulanmış ve modelin R^2 değeri %98 olarak hesaplanmıştır. Ancak, söz konusu çalışmada alternatif algoritmalarla karşılaştırma yapılmamış ve hiperparametre optimizasyonu uygulanmamıştır. Buna karşın, bu çalışmada aynı özelliklere sahip bir veri seti kullanılmış ve farklı algoritmaların performansları test edilerek DVR algoritmasının en iyi sonuçları verdiği belirlenmiştir ($R^2 = 0.9469$). Bu durum, algoritma seçiminin model performansı üzerinde doğrudan etkili olduğunu ve farklı modellerin sistematik olarak değerlendirilmesinin tahmin doğruluğunu artırabileceğini göstermektedir.

Ayrıca, Chenrui Pei (2024) tarafından yapılan çalışmada, Kaggle'dan alınan ve 2.392 lise öğrencisine ait veri seti kullanılmış ve çoklu doğrusal regresyon modeli ile akademik performans tahmini yapılmıştır. Modelin R^2 değeri 0.954 olarak hesaplanmış ve devamsızlık sayısının akademik performans üzerinde negatif bir etkisi olduğu belirlenmiştir. Bu çalışmada ise aynı veri seti kullanılmış, ancak çok daha geniş bir model yelpazesıyla analiz gerçekleştirilerek DVR algoritmasının en iyi sonuçları verdiği görülmüştür ($R^2 = 0.9999$). Pei'nin çalışmasında yalnızca çoklu doğrusal regresyon modeli kullanılmışken, bu çalışmada farklı makine öğrenimi algoritmaları denenmiş, hiperparametre optimizasyonu uygulanmış ve model performansları titizlikle karşılaştırılmıştır. Elde edilen bulgular, algoritma seçiminin ve optimizasyon yöntemlerinin model başarısı üzerinde belirleyici bir rol oynadığını göstermektedir.

Önceki çalışmaların önemli bir bölümü, tek bir model veya sınırlı sayıda makine öğrenimi tekniği kullanarak öğrenci başarısını tahmin etmeye odaklanmıştır. Ancak bu çalışmada, geniş veri setleri, farklı algoritmalar ve hiperparametre optimizasyonu gibi unsurlar dikkate alınarak daha kapsamlı bir analiz gerçekleştirilmiştir. Böylece, öğrenci başarı tahmininde model seçim sürecinin önemi vurgulanmış ve veri setinin yapısına bağlı olarak belirli algoritmaların daha üstün performans gösterebileceği ortaya konmuştur.

Bu çalışmada, öğrencilerin sınav puanlarını tahmin etmeye yönelik bir web uygulaması geliştirilmiştir. Uygulama, kullanıcı adı ve şifre ile giriş yapılarak

erişilebilen bir sistem olarak tasarlanmıştır. Kullanıcılar, sisteme giriş yaptıktan sonra CSV formatında bir veri seti yükleyerek öğrencilerin sınav veya final notlarını tahmin edebilmektedir. Model, yüklenen veriler üzerinde tahminleme yaparak çıktı üretmekte ve sonuçlar kullanıcıya sunulmaktadır.

Önceki çalışmalarda, öğrencilerin akademik başarılarını tahmin etmeye yönelik çeşitli modellerin kullanıldığı görülmekle birlikte, kolay bir arayüze sahip olan ve web tabanlı bir sistem üzerinden bu süreci gerçekleştiren bir çalışmaya rastlanmamıştır. Mevcut literatürde genellikle makine öğrenmesi algoritmalarının doğruluğu, farklı veri setleri üzerinde değerlendirilmiş, ancak bu algoritmaların öğretmenler veya eğitimciler tarafından doğrudan kullanılmasını sağlayan bir sistem geliştirilmemiştir. Bu durum, çalışmanın alana sağladığı önemli bir katkı olarak değerlendirilmektedir.

Geliştirilen sistemin kolay bir arayüze sahip olması, öğretmenlerin veya eğitimcilerin herhangi bir teknik bilgiye ihtiyaç duymadan öğrenci performansını analiz edebilmesine olanak tanımaktadır. Ayrıca, modelin doğruluğunun yüksek olması, uygulamanın güvenilirliğini artırmakta ve eğitim süreçlerinde kullanılabilirliğini güçlendirmektedir.

Sonuç olarak, bu çalışmanın bulguları, daha önceki araştırmalar ile büyük ölçüde uyumlu olmakla birlikte, model çeşitliliğinin artırılması, geniş çaplı veri setleri üzerinde analizler yapılması ve hiperparametre optimizasyonunun uygulanması sayesinde daha yüksek başarı oranlarına ulaşılmıştır. Geniş ve çeşitli veri setlerinin kullanımı, modelin genelleştirilebilirliğini artırmış ve öğrenci başarı tahmininde önemli bir avantaj sağlamıştır. Elde edilen sonuçlar, eğitimde makine öğrenimi uygulamalarının etkinliğini ortaya koymakta ve algoritma seçiminin model performansı üzerindeki belirleyici rolünü vurgulamaktadır. Geliştirilen web uygulaması, yalnızca öğrenci notlarını tahmin etmekle kalmayıp, aynı zamanda öğretmenlerin öğrencilerin akademik durumlarını daha yakından takip etmelerine de olanak tanımaktadır. Bu sistem sayesinde, öğretmenler öğrencilerin başarısını etkileyen faktörleri analiz edebilmekte ve bireysel rehberlik hizmetleri sunarak kişiselleştirilmiş eğitim süreçlerine katkı sağlayabilmektedir. Ayrıca, uygulamanın sunduğu tahminler, öğretmenlere öğrencilerin eksik olduğu konular hakkında bilgi verme ve eğitim stratejilerini bu doğrultuda şekillendirme fırsatı sunmaktadır.

Bu yönleriyle ele alındığında, çalışmada geliştirilen web tabanlı sistemin, eğitim alanında veri odaklı karar alma süreçlerine destek olduğu ve öğrenci başarısını

artırmaya yönelik önemli bir araç sunduęu sonucuna ulařılmaktadır.

6. SONUÇLAR

Bu çalışmada, öğrenci akademik başarılarının tahmin edilmesi amacıyla Kaggle platformundan elde edilen beş farklı veri seti üzerinde kapsamlı analizler gerçekleştirilmiş ve çeşitli makine öğrenimi algoritmalarının performansları değerlendirilmiştir. Veri analizi sürecinde öncelikle eksik değer analizi yapılmış ve eksik verilerin veri setine etkisi incelenmiştir. Aykırı değerler Z-skoru yöntemi ile belirlenerek veri setinin güvenilirliği artırılmıştır. Regresyon algoritmalarıyla çalışıldığı için kategorik değişkenler sayısal formatta işlenmek üzere Label Encoding yöntemiyle dönüştürülmüştür. Model performanslarını iyileştirmek amacıyla veri setlerine 10 kat çapraz doğrulama, standartlaştırma ve hiperparametre optimizasyonu gibi işlemler uygulanmış, böylece farklı algoritmaların doğruluk oranları sistematik olarak karşılaştırılmıştır.

Veri seti 1, 6607 gözlem ve 20 değişkenden oluşmaktadır. İlk olarak, 10 kat çapraz doğrulama yöntemi uygulanmış ve Ridge ile Lineer Regresyon algoritmaları %92.69 doğruluk oranına ulaşmıştır. Bu aşamada, ortalama mutlak hata (MAE) 0.5545 ve kök ortalama kare hata (RMSE) 0.8901 olarak hesaplanmıştır. Ardından, veri setine standartlaştırma işlemi uygulanarak model performanslarının iyileştirilmesi sağlanmıştır. Standartlaştırma sonucunda en iyi performans %92.37 doğruluk oranı ile Destek Vektör Regresyonu (DVR) algoritmasından elde edilmiştir. Ancak, standartlaştırma ile MAE 0.8237'ye yükselmiş ve RMSE 0.5944 olarak hesaplanmıştır. Hiperparametre optimizasyonu için GridSearchCV kullanıldığında, en yüksek doğruluk oranı %95.52 ile Random Forest algoritması tarafından elde edilmiştir ve bu aşamada MAE 0.5556, RMSE ise 0.7220 olarak hesaplanmıştır. Son olarak, standartlaştırma ve GridSearchCV birlikte uygulandığında Random Forest algoritmasının doğruluk oranı %95.53'e ulaşmış ve MAE 0.5586, RMSE ise 0.7211 olarak kaydedilmiştir. Bu bulgular, veri setinin karmaşık ve çok sayıda değişken içermesi nedeniyle doğrusal regresyon modellerinin performansının sınırlı kaldığını, doğrusal olmayan ilişkileri daha iyi modelleyebilen Random Forest algoritmasının ise en iyi sonucu verdiğini göstermektedir. Random Forest algoritmasının başarısı, karar ağaçları kullanarak değişkenler arasındaki ilişkileri esnek bir şekilde öğrenebilmesi ve fazla değişkenler arasındaki anlamlı olanları öne çıkarabilmesi ile açıklanmaktadır.

Veri seti 2, 10.000 gözlem ve 15 değişkenden oluşmaktadır. 10 kat çapraz doğrulama Ridge ve Lineer Regresyon algoritmaları ile uygulanmış ve %93.68 doğruluk oranına ulaşmıştır. Bu aşamada, ortalama mutlak hata (MAE) 3.8271 ve

kök ortalama kare hata (RMSE) 4.4291 olarak hesaplanmıştır. Standartlaştırma işlemi uygulandığında, model performansında herhangi bir değişiklik olmamış ve doğruluk oranı yine %93.68 olarak hesaplanmıştır. Bu durumda, MAE 3.8308 ve RMSE 4.4337'ye yükselmiştir. Hiperparametre optimizasyonu için GridSearchCV kullanıldığında, en yüksek doğruluk oranı %93.80 ile Lasso ve Lineer Regresyon algoritmaları tarafından elde edilmiştir. GridSearchCV sonrası MAE 3.7783, RMSE ise 4.3973 olarak hesaplanmıştır. Son olarak, standartlaştırma ve GridSearchCV birlikte uygulandığında, en yüksek doğruluk oranı %94.07 ile Destek Vektör Regresyonu (DVR) tarafından sağlanmıştır ve bu aşamada MAE 3.6525, RMSE ise 4.2676 olarak kaydedilmiştir. Bu sonuçlar, Veri seti 2'de bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin doğrusal modellere uygun olduğunu göstermektedir. Ridge ve Lineer Regresyon modellerinin başarısı, veri setindeki değişkenlerin ağırlıklı olarak doğrusal ilişkiler göstermesiyle açıklanabilir. Ancak, hiperparametre optimizasyonu ile Destek Vektör Regresyonu'nun (DVR) performansının arttığı ve en yüksek doğruluk oranına ulaştığı belirlenmiştir. Bu durum, DVR'nin doğrusal olmayan durumları modelleyebilme yeteneğinin, veri setinin özelliklerine bağlı olarak avantaj sağladığını göstermektedir.

Veri seti 3., 1388 gözlem ve 5 değişkenden oluşmaktadır. 10 kat çapraz doğrulama uygulandıktan sonra Random Forest algoritmasının %97.84 doğruluk oranına ulaştığı belirlenmiştir. Bu aşamada, ortalama mutlak hata (MAE) 1.3689 ve kök ortalama kare hata (RMSE) 0.9734 olarak hesaplanmıştır. Standartlaştırma işleminin ardından, Random Forest algoritmasının doğruluğu %97.92'ye yükselmiştir ve bu durumda MAE 0.9557, RMSE ise 1.3258 olarak belirlenmiştir. Hiperparametre optimizasyonu için GridSearchCV uygulandığında, en iyi sonuç %99.40 doğruluk oranı ile Random Forest algoritmasından elde edilmiştir. GridSearchCV sonrası MAE 0.5378, RMSE ise 0.7358 olarak kaydedilmiştir. Son olarak, standartlaştırma ve GridSearchCV birlikte uygulandığında, doğruluk oranı %97.99 olarak hesaplanmış, bu durumda MAE 0.9171, RMSE ise 1.3184 olarak hesaplanmıştır. Bu sonuçlar, Random Forest algoritmasının veri setindeki bağımsız değişkenlerin etkisini daha iyi modelleyerek genel hataları azaltan ve farklı değişkenler arasındaki ilişkileri yakalayabilen bu algoritmanın, doğruluk oranını önemli ölçüde artırdığı belirlenmiştir.

Veri seti 4, 10.000 gözlem ve 6 değişkenden oluşmaktadır. 10 kat çapraz doğrulama Ridge ve Lineer Regresyon algoritmaları ile uygulanmış ve %98.87 doğruluk oranına ulaşılmıştır. Bu aşamada, ortalama mutlak hata (MAE) 1.6179 ve kök ortalama kare hata (RMSE) 2.0379 olarak hesaplanmıştır. Standartlaştırma

işlemi gerçekleştirildiğinde, Random Forest algoritmasının doğruluk oranı %99.97'ye yükselmiş, bu durumda MAE 0.1678 ve RMSE 0.3187 olarak belirlenmiştir. Hiperparametre optimizasyonu için GridSearchCV uygulandığında, Random Forest algoritmasının doğruluk oranı %99.84 olarak hesaplanmış, bu süreçte MAE 0.5725 ve RMSE 0.7567 olarak kaydedilmiştir. Son olarak, standartlaştırma ve GridSearchCV birlikte uygulandığında, en yüksek doğruluk oranı DVR algoritması %99.99 ile elde edilmiştir. Bu durumda MAE 0.0767 ve RMSE 0.1858 olarak hesaplanmıştır. Bu bulgular, veri setinin karmaşık yapısına rağmen, Random Forest algoritmasının istikrarlı bir şekilde yüksek performans sunduğunu ve veri işleme teknikleri ile model başarısının daha da artırılabilceğini göstermektedir.

Veri seti 5, 2392 gözlem ve 15 değişkenden oluşmaktadır. 10 kat çapraz doğrulama Ridge ve Lineer Regresyon algoritmaları ile uygulandığında, doğruluk oranı %95.07 olarak hesaplanmıştır. Bu aşamada, ortalama mutlak hata (MAE) 0.1623 ve kök ortalama kare hata (RMSE) 0.1996 olarak belirlenmiştir. Standartlaştırma işlemi uygulandığında, Lineer Regresyon modeli %100 doğruluk oranına ulaşmış ve MAE ile RMSE sıfır değerlerini almıştır. Hiperparametre optimizasyonu için GridSearchCV yapıldığında, Lineer Regresyon modeli yine %100 doğruluk oranına ulaşmış ve yine MAE ile RMSE sıfır olarak gözlenmiştir. Son olarak, standartlaştırma ve GridSearchCV birlikte uygulandığında, Lineer Regresyon algoritmasının $R^2 = 1.0000$ skoru ile en iyi sonucu verdiği belirlenmiştir. Bu sonuçlar, Veri seti 5'te bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin tamamen doğrusal olduğunu ve Lineer Regresyon algoritmasının bu tür veri setlerinde en başarılı modeli sunduğunu göstermektedir.

Bu çalışmada elde edilen bulgular doğrultusunda, öğrencilerin akademik başarılarını daha doğru ve etkili bir şekilde tahmin edebilmek amacıyla bir web arayüzü geliştirilmiştir. Bu web arayüzü, makine öğrenimi algoritmalarını kullanarak öğrenciye ait çeşitli özellikleri analiz etmekte ve bu doğrultuda sınav puanını sayısal olarak tahmin edebilmektedir. Öğrencinin çalışma süresi, uyku süresi, devamsızlık bilgisi gibi birçok faktörü göz önünde bulunduran sistem, kişiselleştirilmiş bir değerlendirme sunarak eğitim sürecine rehberlik etmektedir. Böylece, öğrenciler güçlü ve zayıf yönlerini daha iyi analiz edebilirken, eğitimciler de bireysel ihtiyaçlara yönelik daha bilinçli kararlar alabilmektedir. Geliştirilen bu sistem, yalnızca mevcut başarıyı tahmin etmekle kalmayıp, öğrencinin akademik performansını artırmaya yönelik veri odaklı stratejiler geliştirilmesine de katkı sağlamaktadır.

Geliştirilen web arayüzü, öğrencilerin akademik başarılarını tahmin etme yeteneğini değerlendirmek amacıyla Veri seti 3 üzerinde test edilmiştir. Veri setinin orijinalinde bulunan grades değişkeni, modelin tahmin performansını ölçmek için veri setinden çıkarılmış ve sistemden bu notları tahmin etmesi istenmiştir. Web uygulaması, öğrencilerin çeşitli özelliklerini analiz ederek grades değerlerini öngörmüş ve modelin doğruluğu değerlendirilmiştir. Bu süreçte, sistemin öğrenci başarılarını kişiselleştirilmiş bir yaklaşımla tahmin edebildiği ve makine öğrenimi algoritmalarını etkin bir şekilde kullanarak gerçek değerlere yakın sonuçlar üretebildiği gözlemlenmiştir. Şekil 6.1. ve Şekil 6.2' de orijinal veri setinde bulunan gerçek not dağılımı ve web uygulaması tarafından tahmin edilen not dağılımı gösterilmektedir.

Socioeconomic Score	Study Hours	Sleep Hours	Attendance (%)	Grades
0.95822	3.4	8.2	53.0	47.0
0.85566	3.2	5.9	55.0	35.0
0.68025	3.2	9.3	41.0	32.0
0.25936	3.2	8.2	47.0	34.0
0.60447	3.8	10.0	75.0	33.0
0.9832	3.4	9.0	47.0	51.0
0.56648	7.9	8.1	63.0	54.0
0.93487	1.4	8.0	47.0	34.0
0.4666	5.4	8.8	67.0	39.0

Şekil 6.1. Veri seti 3, Orijinal Veri Setinde Bulunan Gerçek Not Dağılımı

Analiz Sonuçları

Başarılı! CSV dosyası analiz edildi ve tahminler oluşturuldu.

socioeconomic score	study hours	sleep hours	attendance (%)	grades	Predicted_Grades
0.95822	3.4	8.2	53.0	nan	47.28
0.85566	3.2	5.9	55.0	nan	34.91
0.68025	3.2	9.3	41.0	nan	32.68
0.25936	3.2	8.2	47.0	nan	34.36
0.60447	3.8	10.0	75.0	nan	33.17
0.9832	3.4	9.0	47.0	nan	50.74
0.56648	7.9	8.1	63.0	nan	54.03
0.93487	1.4	8.0	47.0	nan	35.23
0.4666	5.4	8.8	67.0	nan	39.11

Şekil 6.2. Web Uygulaması Tarafından Tahmin Edilen Not Dağılımı

Uygulamanın doğruluk oranını değerlendirebilmek amacıyla, gerçek sınav notları ile model tarafından tahmin edilen sınav notları karşılaştırılmış ve elde edilen sonuçlar görselleştirilmiştir.

Modelin performansını ölçmek için R^2 (Determinasyon Katsayısı) kullanılmıştır. Yapılan hesaplamalar sonucunda, geliştirilen model için $R^2 = 0.9950$ olarak bulunmuştur. Ayrıca, tahminleme işlemi için Random Forest algoritması uygulanmış ve bu yöntemde elde edilen R^2 değeri 0.9940 olarak hesaplanmıştır.

Bu iki değer birbirine oldukça yakın olması, modelin yüksek doğrulukla tahmin yapabildiğini göstermektedir. Modelin sınav notlarını gerçeğe çok yakın bir şekilde tahmin edebilmesi, sistemin güvenilir olduğunu ve eğitim süreçlerinde etkili bir şekilde kullanılabileceğini kanıtlamaktadır. Bunun sonucunda, geliştirilen web uygulamasının başarılı bir şekilde çalıştığı ve öğrenci notlarını yüksek bir doğruluk oranıyla tahmin edebildiği ortaya konmuştur.

Sonuç olarak, bu çalışmada uygulanan veri ön işleme adımları, çapraz doğrulama, standartlaştırma ve hiperparametre optimizasyonunun model performansları üzerinde önemli etkileri olduğu belirlenmiştir. Doğru algoritma seçiminin ve veri işleme tekniklerinin, model başarısını doğrudan etkilediği görülmüştür. Random Forest algoritmasının, doğrusal olmayan ilişkiler içeren ve çok

değişkenli veri setlerinde en başarılı model olduğu belirlenmiştir. Buna karşılık, Destek Vektör Regresyonu'nun (DVR), veri setine bağlı olarak yüksek doğruluk oranlarına ulaşabildiği ve hiperparametre optimizasyonu ile performansının arttığı gözlemlenmiştir. Lineer Regresyon modelinin ise, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkinin doğrusal olduğu durumlarda en başarılı model olduğu tespit edilmiştir. Bu çalışma, eğitim verilerinde makine öğrenimi modellerinin etkinliğini ortaya koymakta ve algoritma seçim sürecinin model performansında belirleyici olduğunu göstermektedir. Geliştirilen web uygulaması ile gerçek verilerle yapılan kıyaslamalar sonucunda, sistemin sınav notlarını büyük oranda doğru tahmin edebildiği ve eğitim süreçlerinde öğretmenler için faydalı bir karar destek mekanizması olarak kullanılabileceği ortaya konmuştur. Bu çalışma, veri ön işleme, model seçimi ve optimizasyon süreçlerinin bir makine öğrenimi modelinin doğruluğunu nasıl etkilediğini ortaya koymakta ve bu süreçlerin sistem performansını artırmada kritik bir rol oynadığını göstermektedir.

Ayrıca, uygulanan yöntemin eğitim sistemine çeşitli açılardan katkı sağlayabileceği belirlenmiştir. Öğrenci başarısının önceden tahmin edilmesi sayesinde, düşük akademik performans gösterme riski taşıyan öğrencilerin erken tespit edilebileceği ve bu doğrultuda kişiselleştirilmiş destek mekanizmalarının oluşturulabileceği ortaya konmuştur. Kullanılan modellerin sunduğu öngörüler, öğretmenler ve eğitim yöneticileri tarafından değerlendirilerek öğrencilerin öğrenme süreçlerinin iyileştirilmesine katkı sağlayabilir.

Elde edilen bulgular, eğitim süreçlerinde veri odaklı karar alma mekanizmalarının uygulanabilirliğini göstermektedir. Modellerin sunduğu tahminler doğrultusunda, öğrencilerin akademik başarılarını artırmaya yönelik kişiselleştirilmiş öğrenme stratejilerinin geliştirilmesi mümkün hale gelmiştir. Ayrıca, eğitim politikalarının belirlenmesi sürecinde, model çıktılarının eğitim kaynaklarının daha etkin dağıtılması açısından yol gösterici bir unsur olabileceği değerlendirilmektedir.

Uygulanan yöntemlerin, öğretmenler ve eğitim yöneticileri için bir karar destek mekanizması olarak kullanılabileceği ortaya konmuştur. Modellerin sunduğu tahminler doğrultusunda, öğrencilerin akademik gelişimlerinin takip edilebileceği ve gerekli müdahalelerin zamanında gerçekleştirilebileceği belirlenmiştir. Bu durumun, uzun vadede öğrenci başarısının artırılmasına katkı sağladığı gibi, eğitimde fırsat eşitliğinin sağlanmasına da destek olabileceği düşünülmektedir.

Sonuç olarak, uygulanan makine öğrenimi yöntemlerinin, öğrenci başarısının

erken dönemde tahmin edilerek eğitim süreçlerine entegre edilmesiyle, akademik performansın artırılmasına ve bireysel öğrenme ihtiyaçlarının daha iyi karşılanmasına olanak tanıdığı belirlenmiştir. Bu tür yaklaşımların geniş ölçekli uygulamalara entegre edilmesi durumunda, eğitimde veri madenciliği ve makine öğrenimi tekniklerinin kullanımına yönelik önemli bir katkı sağlayabileceği öngörülmektedir.

7. ÖNERİLER

Bu araştırmada kullanılan tekniklere ilave olarak, Gradient Boosting algoritmaları gibi daha gelişmiş makine öğrenimi yöntemlerinin denenmesi önerilmektedir. Örneğin, XGBoost, LightGBM ya da CatBoost algoritmalarının performanslarının değerlendirilmesi ve bu tekniklerin doğruluk oranlarını artırıp artırmadığının araştırılması faydalı olacaktır. Ayrıca, bu algoritmalar için yapılan hiperparametre ayarlamalarına daha fazla yoğunlaşarak model performansının daha üst seviyelere çıkarılması sağlanabilir.

Bunun yanı sıra, toplu öğrenme (ensemble) yöntemlerinin kullanılması, tahmin doğruluğunu artırabilir. Özellikle, bagging ya da boosting tabanlı yaklaşımlar bir araya getirilerek hibrit bir yapı oluşturulabilir ve bu sayede daha dengeli ve genellenebilir sonuçlar elde edilebilir. Birden fazla modelin birlikte çalıştığı bir toplu öğrenme modeli, tekil modellerin kısıtlamalarını aşma noktasında etkili olabilir.

Özellik mühendisliği bağlamında, veri setine yeni değişkenler eklemek ya da mevcut değişkenleri dönüştürmek modellerin performansını iyileştirebilir. Örneğin, değişkenlerin logaritmik dönüşümleri, polinomik genişlemeleri ya da etkileşim terimlerinin eklenmesi gibi yöntemlerle modelin hedef değişkenle olan ilişkisi daha iyi anlaşılabilir. Bununla birlikte, otomatik özellik seçimi ve türetme araçlarının kullanılması, bu süreçleri daha verimli hale getirebilir ve manuel işlemleri minimuma indirebilir.

Araştırmanın gerçek zamanlı uygulamalardaki etkilerini incelemek de değerlendirilebilir. Örneğin, öğrenci başarısını tahmin eden bir sistemin eğitim kurumlarında uygulanabilirliği test edilebilir. Bu tür bir gerçek zamanlı öğrenme sistemi, öğrencilere özel öneriler sunarak akademik başarılarını yükseltebilir ve modellerin pratikte ne kadar etkili olduğunu anlamak için önemli bir fırsat sunabilir.

Son olarak, çalışmada kullanılan veri setlerinin ötesine geçerek farklı veri setleri üzerinde analizler yapılması önerilmektedir. Bu, geliştirilmiş modellerin genellenebilirlik düzeyini ve farklı koşullardaki performansını ölçmek açısından faydalı olacaktır. Özellikle, öğrenci başarısını etkileyen daha çeşitli ve kapsamlı veri setleri üzerinde çalışarak, bu faktörlerin tahmin doğruluğuna olan katkıları incelenebilir. Böyle bir yaklaşım, modellerin daha kapsayıcı ve tutarlı sonuçlar üretmesini sağlayacaktır.

Bunlara ek olarak, kişiselleştirilmiş eğitim anlayışına daha fazla katkı sağlayabilecek, kapsamlı ve gelişmiş bir web uygulamasının geliştirilmesi önerilmektedir. Mevcut sistem, öğrenci başarı tahminine yönelik önemli bir adım olmakla birlikte, gelecekte bu sistemin daha etkileşimli ve çok yönlü bir hale getirilmesi, eğitim süreçlerinde daha geniş bir etki alanı yaratabilir.

Örneğin, geliştirilecek web uygulaması, yalnızca sınav notlarını tahmin etmekle kalmayıp, öğrencinin eksik olduğu konuların belirlenmesi ve bu eksikliklerin giderilmesine yönelik kişiselleştirilmiş öneriler sunan bir yapıya dönüştürülebilir. Bu doğrultuda, sistemin öğrenciye özel çalışma planları oluşturması, eksik kazanımlara yönelik ek materyaller sunması ve öğretmenlerin rehberlik süreçlerini destekleyen analizler sağlaması sağlanabilir.

Bunun yanı sıra, öğrenci performansına dayalı olarak dinamik geri bildirim mekanizmalarının entegre edilmesi önerilmektedir. Örneğin, öğrencinin geçmiş başarıları ve öğrenme eğilimleri dikkate alınarak, hangi öğrenme yöntemlerinin daha etkili olduğu tespit edilebilir ve buna uygun çalışma stratejileri geliştirilebilir. Böylece, her öğrencinin öğrenme süreci daha verimli hale getirilebilir.

Ayrıca, öğretmenlerin daha kapsamlı analizler yapabilmesi için öğrenci gelişim raporları, performans analizleri ve görselleştirilmiş veriler sunan bir gösterge paneli (dashboard) eklenmesi faydalı olacaktır. Bu sayede öğretmenler, öğrencilerin hangi alanlarda zorlandığını daha kolay belirleyerek, bireysel veya grup bazlı müdahalelerde bulunabilirler.

Bununla birlikte, uygulamanın öğrencilere yalnızca sınav tahminleri sunan bir sistem olmaktan çıkarılıp, aynı zamanda öğretmenlerin öğrenciye rehberlik edebileceği bir platforma dönüşmesi önerilmektedir. Örneğin, öğretmenler, sistem üzerinden öğrencilerle bireysel geri bildirim paylaşabilir, rehberlik hizmetleri sunabilir ve öğrencinin akademik ilerlemesine yönelik önerilerde bulunabilirler.

Son olarak, geliştirilecek uygulamanın yapay zeka destekli öneri sistemleri ile daha da güçlendirilmesi önerilmektedir. Öğrencinin öğrenme tarzına ve başarı geçmişine uygun olarak video dersler, interaktif testler veya okuma materyalleri sunan bir sistem, bireysel öğrenme süreçlerini destekleyerek eğitimin daha verimli hale getirilmesine katkı sağlayabilir.

Bu tür geliştirmeler, kişiselleştirilmiş eğitimi daha etkin hale getirecek,

öğretmen ve öğrenciler arasındaki etkileşimi artıracak ve eğitim süreçlerinde veri odaklı kararların daha sağlıklı bir şekilde alınmasına olanak tanıyacaktır. Bu doğrultuda, geliştirilecek kapsamlı ve çok yönlü bir web uygulamasının, öğrenci başarısını artırma ve eğitimde dijitalleşmeyi daha etkili bir şekilde kullanma konusunda önemli bir katkı sağlayacağı düşünülmektedir.

KAYNAKLAR

- Abbas, A. (2024). Student performance prediction: EDA & ML. Kaggle. Retrieved December 20, 2024, from <https://www.kaggle.com/code/ahmedabbas757/student-performance-prediction-eda-ml> (Eriřim Tarihi: 15.11.2024, Veri seti 4)
- Alaa, M. (tarih yok). *Data analysis - student performance factors*. Kaggle. (Eriřim tarihi: 2025, řubat, Veri seti 1) <https://www.kaggle.com/code/muhammadalaa-sar/data-analysis-studentperformancefactors>
- Alpaydin, E. (2020). *Introduction to machine learning* (4th ed.). MIT Press.
- (Alfajr & Defiyanti, 2024; Dai vd., 2023; Kumar vd., 2024; Mrg & Hasibuan, 2024; Salman vd., 2024)
- Adak, M. F., & Duralioęlu, Ö. (2023). Makine öğrenmesi yöntemleri kullanılarak öğrencilerin kazanım bilgileri ile sınavlardaki başarı durumunun tahmini. *Journal of Intelligent Systems: Theory and Applications*, 6(1), 43-51.
- Aghalarova, S., & Keser, S. B. (2021). Önerilen Yapay Sinir Aęı Algoritması ile Ortaokul Öğrencilerin Akademik Performansının Tahmini. *Veri Bilimi*, 4(2), 19-32.
- Alkan, A. (2024). Öğrencilerin sınavlardaki performansının makine öğrenmesi teknikleriyle tahminlenmesi. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 7(3), 1116-1128.
- Alpar, R. (2010). Basit doğrusal regresyon çözümlemesi. In Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik (ss. 285-304). Detay Yayıncılık.
- Atalay, M., & Çelik, E. (2017). Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamaları. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 9(22), 155-172.
- Atalay, M., & Çelik, E. (2017). Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamaları. Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 9(22), 155-172.
- Bakan, Z., & Kanbay, F. (2024). Makine öğrenmesi yöntemleri ile eğitim başarısına etki eden faktörlerin modellenmesi. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 23(45), 27-41.
- Bilgin, M. (2017). Gerçek veri setlerinde klasik makine öğrenmesi yöntemlerinin performans analizi. *Breast*, 2(9), 683.
- Bölükbaş, M. A. (2013). Çalışan memnuniyetinin veri madencilięi ile incelenmesi (Master's thesis, Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Bilimleri Enstitüsü).
- Börekçi, M. (2000). Yabancı dille öğretim. *Kazım Karabekir Eğitim Fakültesi Dergisi, Dil Bilimi ve Dil Öğretimi Özel Sayı*, 204-218.
- Can, M. B., Eren, Ç., Koru, M., Özkan, Ö., & Rzayeva, Z. (2012). Veri

kümelerinden bilgi keşfi: Veri madenciliği. Başkent Üniversitesi Tıp Fakültesi XIV. Öğrenci Sempozyumu, Ankara.

- Can, Ş., Özdil, T., & Yılmaz, C. (2018). Üniversite öğrencilerinin ders başarısını etkileyen faktörlerin lojistik regresyon analizi ile tahminlenmesi. *International Review of Economics and Management*, 6(1), 28-49.
- Chen, C. M., Wang, J. Y., & Yu, C. M. (2017). Assessing the attention levels of students by using a novel attention aware system based on brainwave signals. *British Journal of Educational Technology*, 48(2), 348-369. <https://doi.org/10.1111/bjet.12333>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Proceedings (pp. 785–794). ACM.
- Coşkun, S., Kartal, M., Coşkun, A., & Bircan, H. (2004). Lojistik regresyon analizinin incelenmesi ve dış hekimliğinde bir uygulama. Cumhuriyet Üniversitesi Dış Hekimliği Fakültesi Dergisi, 7(1), 41-50.
- Çakmak, Ö. (2008). Eğitimin ekonomiye ve kalkınmaya etkisi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 11, 33-41.
- Çetin, V., & Yıldız, O. (2021). A comprehensive review on data preprocessing techniques in data analysis. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 28(2), 299-312.
- Deepika, K., & Sathyanarayana, N. (2018). Comparison of student academic performance on different educational datasets using different data mining techniques. *International Journal of Computational Engineering Research (IJCER)*, 8(9), 28-38.
- Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018, October). Unsupervised learning based on artificial neural network: A review. In 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS) (pp. 322-327). IEEE.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155-161.
- Dubey, K. P., Gupta, N. K., & Sharma, A. (2024). Performance prediction data mining system for disabled students using machine learning. *Fusion: Practice & Applications*, 15(2).
- El Kharoua, R. (2024). Students performance dataset. Kaggle. Retrieved December 20, 2024, from <https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset> (Erişim Tarihi: 15.11.2024, Veri seti 5)
- Emhan, Ö., & Akın, M. (2019). Filtreleme tabanlı öznelik seçme yöntemlerinin anomali tabanlı ağ saldırısı tespit sistemlerine etkisi. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 10(2), 549–559.

- Er, M. B., Balıkçı, H. C., Alpsülün, M., Bozdağ, Z., Yıldız, A., Tezcan, A., & Yıldırım, N. T. (2024). Yapay zeka uyumlu algoritmalarla öğrencilerin LGS puanı tahmini ve modellenmesi. *Necmettin Erbakan Üniversitesi Ereğli Eğitim Fakültesi Dergisi*, 5(2), 377-402.
- Erten, H. (2015). *Veri madenciliği teknikleri ile organ nakli için uygun donör oranının hesaplanması* (Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü, Ankara.
- Ertürk, S. (1972). *Eğitimde program geliştirme*. Yelken Tepe Yayınları.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-37.
- Gökçe, A. T. (2008). Küreselleşme sürecinde uzaktan eğitim. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi*, 11, 1-12.
- Güvenç, E., Sakal, M., Çetin, G., & Özkaraca, O. (2022). Öğrencilerin dersteki niteliklerinin makine öğrenmesi teknikleri kullanılarak sınıflandırılması. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 10(3), 1359-1371.
- Irmak, S. (2009). Veri madenciliği yöntemleri ve sağlık sektörü veritabanlarında bilgi keşfi: Tanımlayıcı ve kestirimci model uygulamaları.
- Jingbo, C., & Yujie, C. (2024, April). Design of an academic early warning model for university students based on data mining techniques. In *2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB)* (pp. 358-362). IEEE. <https://doi.org/10.1109/ICEIB52918.2024.0100>
- Jordan, M., Kleinberg, J., & Scholkopf, B. (2008). Support vector machines. *Information Science and Statistics*. New York: Springer.
- Kalikov, A. (2006). *Veri madenciliği ve bir e-ticaret uygulaması* (Yüksek Lisans Tezi). Gazi Üniversitesi Fen Bilimleri Enstitüsü.
- Kaynak: (https://erdincuzun.com/makine_ogrenmesi/...) (2020)
- Khemka, A. (2003). A collaborative predictive data mining model (Unpublished master's thesis). University of Missouri-Kansas City, Missouri.
- Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M., & Özçift, A. (2016). KNN algoritması ve R dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. *Marmara Fen Bilimleri Dergisi*, 28(3), 89-94.
- Kumral, C. D., Topal, A., Ersoy, M., Çolak, R., & Yiğit, T. (2022). Random forest algoritmasının FPGA üzerinde gerçekleştirilerek performans analizinin yapılması. *El-Cezeri*, 9(4), 1315-1327.
- Magdin, M., & Turcáni, M. (2015). Personalization of students in course management systems on the basis using method of data mining. *Turkish Online Journal of Educational Technology-TOJET*, 14(1), 58-67.
- Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29(1), 3-20.

- Microsoft. (n.d.). Veritabanı yönetim sistemi (DBMS) ve SQL sunucusu için temel yedekleme ve geri yükleme konseptleri. Microsoft. [https://technet.microsoft.com/trtr/library/ms175382\(v=sql.105\).aspx](https://technet.microsoft.com/trtr/library/ms175382(v=sql.105).aspx)
- Mining, W. I. D. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.
- Oğuzlar, A. (2003). Veri ön işleme. Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, (21).
- Orhan, H., & Vergili, M. (2022). Genomik veri setlerinin LASSO ve elastik net regresyon yöntemleri ile analizi. SDU Journal of Health Science Institute/SDÜ Sağlık Bilimleri Enstitüsü Dergisi, 13(3).
- Pei, C. (2024). Research on the influencing factors of student performance. *Theoretical and Natural Science*, 41, 43-50.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. NeurIPS - 32nd Conference on Neural Information Processing Systems, Montreal, 2018.
- Rabelo, A., Rodrigues, M. W., Nobre, C., Isotani, S., & Zárata, L. (2024). Educational data mining and learning analytics: A review of educational management in e-learning. *Information Discovery and Delivery*, 52(2), 149-163. <https://doi.org/10.1108/IDD-12-2023-0114>
- Ratner, B. (2000). A comparison of two popular machine learning methods. *Mine Tech*.
- Salal, Y. K., Abdullaev, S. M., & Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Technology*, 8(4C), 54-59.
- Sammut, C., & Webb, G. I. (2010a). Mean absolute error. Springer. https://doi.org/10.1007/978-1-4419-1698-8_652
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210-229.
- Shamim, A.. *Student performance EDA & ML*. Kaggle.(Erişim tarihi: 2025, Şubat, Veri seti 2) <https://www.kaggle.com/code/adilshamim8/student-performance-eda-ml>
- Staneviciene, E., Gudoniene, D., Punys, V., & Kukstys, A. (2024). A case study on the data mining-based prediction of students' performance for effective and sustainable e-learning. *Sustainability*, 16(23), 1-15. <https://doi.org/10.3390/su16239802>
- Stealth Technologies. *Predict student performance dataset*. Kaggle. (Erişim tarihi: 2025, Şubat, Veri seti 3) <https://www.kaggle.com/datasets/stealthtechnologies/predict-student-performance-dataset>
- Suleiman, I. B., Okunade, O. A., Dada, E. G., & Ezeanya, U. C. (2024). Key factors influencing students' academic performance. *Journal of Electrical Systems*

and Information Technology, 11(1), 41.

- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
- Tang, Z., & MacLennan, J. (2005). Data mining with SQL Server 2005. Wiley.
- Tunç, A., & Ülger, İ. (2016). Veri madenciliği uygulamalarında özellik seçimi için finansal değerlere binning ve five number summary metotları ile normalizasyon işleminin uygulanması. XVIII. Akademik Bilişim Konferansı, Adnan Menderes Üniversitesi, 30, 3.
- Türk Dil Kurumu. (2020, Temmuz 10). Eğitim. Erişim tarihi 16 Ağustos 2024, <https://sozluk.gov.tr> adresinden alındı.
- Üstüner, M., Abdikan, S., Bilgin, G., & Şanlı, F. B. (2020). Hafif gradyan artırma makineleri ile tarımsal ürünlerin sınıflandırılması. Türk Uzaktan Algılama ve CBS Dergisi, 1(2), 97-105.
- Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. Annals of Statistics, 14(4), 1261-1290.
- Yağcı, M. (2024). Akademik başarının veri madenciliği yöntemleriyle tahmin edilmesi. *Mühendislik Bilimleri ve Tasarım Dergisi*, 12(2), 443-454.
- Yılmaz, Ş. Ş., Özer, İ., & Gökçen, H. (2021). Türkçe metinlerde derin öğrenme yöntemleri kullanılarak duygu analizi. International Symposium of Scientific Research and Innovative Studies, 22-25 Şubat 2021, 971-982.
- Zilyas, D., & Yılmaz, A. (2023). Makine öğrenmesi yöntemleri ile eğitim başarısının tahmini modeli. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 14(3), 437-447.