T.C.

İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

EĞİTİM VERİLERİNİN ANALİZİNDE REGRESYON VE SINIFLANDIRMA ALGORİTMALARI: ÖĞRENCİ PERFORMANSININ MODELLENMESİ

YÜKSEK LİSANS TEZİ

Erdal Özkul

1009041005

Anabilim Dalı: Matematik ve Bilgisayar Bilimleri

Programı: Matematik ve Bilgisayar Bilimleri

Tez Danışmanı: Dr. Öğr. Üyesi Mehmet Fatih UÇAR

OCAK 2024

T.C.

İSTANBUL KÜLTÜR ÜNİVERSİTESİ

LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

EĞİTİM VERİLERİ ANALİZİNDE REGRESYON VE SINIFLANDIRMA ALGORİTMALARI: ÖĞRENCİ PERFORMANSININ MODELLENMESİ

Yüksek Lisans Tezi

Erdal Özkul

1009041005

Anabilim Dalı: Matematik ve Bilgisayar Bilimleri

Programı: Matematik ve Bilgisayar Bilimleri

Danışman: Dr. Öğr. Üyesi Mehmet Fatih UÇAR

Jüri Üyeleri: Prof.Dr. Ozan Kocadağlı

Dr.Öğr.Üyesi Levent Cuhacı

OCAK 2024

ÖNSÖZ

Tez yolculuğum boyunca desteğini ve motivasyonunu esirgemeyen değerli danışmanım Dr. Mehmet Fatih UÇAR'a şükranlarımı sunuyorum. Onun en önemli katkısı bitmek bilmeyen pozitif yaklaşımı; mesai saati içinde veya dışında beni hiçbir zaman yanıtsız bırakmamasıydı. Özellikle bunaldığım ve şüpheye düştüğüm anlarda, onun bana cesaret vermesi benim tekrardan çalışmama odaklanmamı sağladı.

Ayrıca akademik dünyayla bağlantım olmadan geçen yıllardan sonra bile bana olan sarsılmaz inancını koruyan, tezimin yazım sürecinde bana tecrübesi ile yol gösteren eşim Suna ÖZKUL'a en derin minnettarlığımı ifade etmeliyim. Kızım Zeynep ÖZKUL'dan ise bitmek bilmeyen sorularına; tez yazım sürecinde, tam cevap veremediğim için özür diliyor, beni yaşına rağmen anlayışla karşıladığı için teşekkür ediyorum.

İçindekiler

KISALTMALAR LİSTESİ (LIST OF ABBREVIATIONS)

English Abbreviation	Definition	Tanım
ML	Machine Learning	Makine Öğrenmesi
ANN	Artificial Neural Networks	Yapay Sinir Ağları
MSE	Mean Squared Error	Hata Karelerinin Ortalaması
R2	R-squared (Coefficient of Determination)	Belirleme Katsayısı (R-kare)
RF	Random Forest	Rassal Orman
SVM	Support Vector Machine	Destek Vektör Makineleri
PCA	Principal Component Analysis	Temel Bileşen Analizi
ROC	Receiver Operating Characteristics	ROC eğrisi
AUC	Area Under the Receiver Operating Characteristics Curve	ROC Eğrisi Altındaki Alan
LR	Logistic Regression	Lojistik Regresyon
MLP	Multi-Layer Perceptron	Çok Katmanlı Perceptron
KNN	K-Nearest Neighbors	K-En Yakın Komşular
DT	Decision Tree	Karar Ağacı
VIF	Variance Inflation Factor	Varyans Enflasyon Faktörü
EKK	Least Squares Method	En Küçük Kareler Yöntemi
$\hat{\beta}$	Least Squares Estimator	En Küçük Kareler Tahmincisi
$\hat{\beta}_k$	*Ridge* Estimator	*Ridge* Tahmincisi
LASSO	Least Absolute Shrinkage and Selection Operator	En Küçük Mutlak Küçülme ve Seçim Operatörü

ŞEKİL LİSTESİ

TABLO LİSTESİ

ABSTRACT

The primary aim of this thesis is to model and predict student academic performance using machine learning techniques. The focus is on 'G_avg', a composite measure derived from the average of three periodic academic grades (G1, G2, G3), providing a holistic view of a student's performance over time. The research aims to uncover patterns and factors significantly affecting academic success by predicting 'G_avg'. The insights gained from this study aim to inform educators and policymakers in developing strategies to enhance student performance and identify students who may need additional support. The thesis also aims to compare various machine learning algorithms to determine the most effective approach for such predictive analysis in an educational context.

As time progresses and machine learning algorithms and technological proficiency continue to advance, our analytical capabilities have become more refined, enabling us to develop deeper insights into complex datasets. With over twenty years of experience in teaching mathematics in private and international schools, my inclination towards education has inevitably evolved to overlap with these technological advancements. I believe the analytical prowess of machine learning algorithms can provide a more comprehensive understanding of student data, thereby illuminating their academic achievements in greater detail.

The dataset used in this research was obtained from Kaggle, a well-established online platform providing authentic datasets for a wide range of analytical endeavors. Despite being sourced from a school in Portugal, its inclusion of universal aspects of education makes it meaningful and beneficial for this study. Its extensive features provided a robust foundation for thorough analysis.

The dataset's evaluations consisted of three different exam results (G1, G2, and G3), with the final exam (target value) determined as G3. Contrary to the expected cumulative structure of

G3, it is designed to function as an independent assessment, similar to G1 and G2. This observation led to the development of a new feature, 'G_avg', calculated as the arithmetic mean of these three grades, to provide a more comprehensive target variable for the analysis. A range of regression models, including *Ridge* Regression, *Lasso* Regression, Elastic Net, and Logistic Regression, were incorporated into my methodology. These models were selected based on their ability to effectively capture the complex, non-linear relationships in the data. In the classification part, three success categories were applied to 'G_avg': 'Unsatisfactory', 'Satisfactory', and 'Successful'. In this context, various classification models were applied. The chosen algorithms are widely recognized for their high efficiency and accuracy in classification tasks: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Neural Networks (MLP Classifier).
The purpose of this research is to provide predictive understandings that can guide educators, thereby enhancing students' educational experiences.

Keywords: Machine Learning, Educational Data Analysis, Academic Performance, Regression Models, Classification Algorithms, Predictive Modeling, Data-Driven Education, Student Success Metrics, Elastic Net, *Ridge* Regression, *Lasso* Regression, Logistic Regression, Decision Trees, Random Forest, SVM, Gradient Boosting, Neural Networks, MLP Classifier, Pedagogical Strategies, Kaggle Dataset, Performance Prediction, Success Interval Classification

ÖZET

Bu tezin temel amacı, makine öğrenmesi tekniklerini kullanarak öğrenci akademik performansını modellemek ve tahmin etmektir. Odak noktası, öğrencinin zaman içindeki performansına ilişkin bütünsel bir görünüm sağlayan, üç periyodik akademik notun (G1, G2, G3) ortalamasından türetilen bileşik bir ölçüm olan ' G_avg ' değişkenidir. Araştırma, ' G_avg ' tahminini yaparak akademik başarıyı önemli ölçüde etkileyen kalıpları ve faktörleri ortaya çıkarmayı amaçlıyor. Bu çalışmadan elde edilen bilgiler, eğitimcileri ve politika yapıcıları öğrenci performansını artırmaya ve ek desteğe ihtiyaç duyabilecek öğrencileri belirlemeye yönelik stratejiler geliştirme konusunda bilgilendirmeyi amaçlamaktadır. Ayrıca tez, eğitim bağlamında bu tür tahmine dayalı analiz için en etkili yaklaşımı belirlemek amacıyla çeşitli makine öğrenimi algoritmalarını karşılaştırmayı amaçlamaktadır.

Zaman ilerledikçe ve makine öğrenimi algoritmaları ve teknolojik beceri gelişmeye devam ettikçe, analitik kapasitelerimiz daha da gelişti ve karmaşık veri kümelerine ilişkin derin içgörüler geliştirmemizi sağladı. Özel ve uluslararası okullarda matematik eğitimi verme konusunda yirmi yıldan fazla deneyime sahip biri olarak, eğitime olan eğilimim kaçınılmaz olarak bu teknolojik ilerlemelerle örtüşecek şekilde gelişti. Makine öğrenimi algoritmalarının analitik yeteneklerinin, öğrenci verilerinin daha kapsamlı anlaşılmasını sağlayabileceğine ve böylece akademik başarılarını daha ayrıntılı bir şekilde aydınlatabileceğine düşünüyorum.

Bu araştırmada kullanılan veri seti, çok çeşitli analitik çabalar için özgün veri setleri sağlayan köklü bir çevrimiçi platform olan Kaggle'dan alındı. Veriler Portekiz'deki bir okuldan alınmış olmasına rağmen, eğitimin evrensel yönlerini bünyesinde barındırması nedeniyle bu çalışma için anlamlı ve faydalıdır. Sahip olduğu detaylı öznitelikler kapsamlı bir inceleme için güçlü bir temel oluşturdu.

Bu veri setindeki değerlendirmeler üç farklı sınav sonucundan (G1, G2 ve G3) oluşuyordu ve sonuç değeri (*target value*) G3 olarak belirlenmişti. G3 ise beklenen kümülatif yapısının aksine, yapısı G1 ve G2'ninkine benzeyen ve bağımsız bir değerlendirme işlevini görecek şekilde tasarlanmıştı. Bu gözlem, analize daha kapsamlı bir hedef değişken sağlamak amacıyla bu üç notun aritmetik ortalaması olarak hesaplanan 'G_avg' adlı yeni bir özniteliğin geliştirilmesine yol açtı.

Ridge Regresyonu, *Lasso Regresyonu, Elastik Net* ve Lojistik regresyon gibi bir dizi regresyon modeli metodolojime dahil edildi. Bu modellerin seçimi, verilerde mevcut olan karmaşık, doğrusal olmayan ilişkileri daha etkili bir şekilde yakalama yeteneklerine dayanıyordu.

Sınıflandırma kısmında ise 'G_avg'e üç başarı kategorisi uygulandı: 'Yetersiz', 'Tatmin Edici' ve 'Başarılı'. Bu bağlamda çeşitli sınıflandırma modelleri uygulandı. Seçilen algoritmalar, sınıflandırma görevlerindeki yüksek verimlilikleri ve doğruluklarıyla geniş çapta bilinen algoritmalardır; Lojistik Regresyon, Karar Ağaçları, Rassal Orman, Destek Vektör Makineleri (SVM), Gradyan Arttırma ve Sinir Ağları (MLP Sınıflandırıcı).

Bu araştırmanın amacı, eğitimcilere rehberlik edebilecek, dolayısıyla öğrencilerin eğitim deneyimlerini geliştirebilecek öngörülü anlayışlar sunmaktır.

Anahtar Kelimeler: Makine Öğrenimi, Eğitimsel Veri Analizi, Akademik Performans, Regresyon Modelleri, Sınıflandırma Algoritmaları, Tahmine Dayalı Modelleme, Veriye Dayalı Eğitim, Öğrenci Başarı Metrikleri, Elastik Net, *Ridge* Regresyonu, *Lasso* Regresyonu, Lojistik Regresyon, Karar Ağaçları, Rassal Orman, SVM, Gradyan Artırma, Sinir Ağları, MLP Sınıflandırıcı, Pedagojik Stratejiler, Kaggle Veri Kümesi, Performans Tahmini, Başarı Aralığı Sınıflandırması

1 GİRİŞ

1.1 İçerik ve Uygunluk

Verilerin dünyada hakim olduğu bir dönemde, yapay zeka (AI) ve makine öğreniminin (ML) bu geniş bilgi havuzunu analiz etme ve kullanmadaki önemi önemli ölçüde artmıştır. Bu durum özellikle, geleneksel olarak öğrencilerin ders başına aldıkları notlarla değerlendirilen akademik başarı ölçümünün, sofistike veri analitiği yöntemleri kullanılarak yeniden değerlendirildiği eğitim alanında açıkça görülmektedir. Günümüz dijital çağında ilerledikçe, eğitim verilerini etkili bir şekilde kullanma ve inceleme kapasitesi, ek bir kaynak olmaktan çıkıp, eğitim uygulamalarının ve öğretim yöntemlerinin geleceğini şekillendirmek için çok önemli olan temel bir gereksinim haline gelmiştir.

Yapay zekanın hızlı ilerlemesine ve farklı endüstrilere dahil edilmesine ayak uydurmak hem zor hem de gerekli bir hale geldi. Eğitim alanında ise bu durum, eğitim verilerindeki karmaşık örüntüleri analiz edebilen ve anlayabilen gelişmiş araçlara yönelik artan bir talep anlamına gelmektedir. Bunun ışığında, araştırmamız yeni bir hedef değişken (G_avg) sunarak geleneksel akademik başarı ölçütlerini yeniden tanımlamayı amaçlamaktadır: G1, G2 ve G3 olmak üzere, bu üç notun ortalaması. Bu verilerdeki final notu olan G3, akademik değerlendirmenin önemli bir yönü olmuştur. Ancak bu çalışma, sadece son nota odaklanmanın bir öğrencinin akademik deneyiminin tüm kapsamını yakalamakta başarısız olabileceğini iddia etmektedir. Dolayısıyla G1, G2 ve G3'ün aritmetik ortalamasını hesaplayarak, öğrenci performansının kapsamlı ve doğru bir ölçüsünün belirlenmesi amaçlandı.

Bu yeni düzenleme hem regresyon hem de sınıflandırma amacıyla makine öğrenimi yöntemlerinin kullanılmasını sağlayacaktır. Regresyon, yeni tanımlanan bir hedef değişkenin sonucunu tahmin etmek için kullanılırken, sınıflandırma akademik başarıyı farklı başarı aralıklarına gruplamak için kullanılır. Bu yöntem sadece öğrenci başarısını anlamamızı geliştirmekle kalmaz, aynı zamanda eğitim tekniklerine ve uygulamalarına rehberlik edebilecek ve bunları zenginleştirebilecek ayrıntılı bir bakış açısı sunar.

Tezimiz, yerleşik eğitim değerlendirme metodolojilerini gelişmekte olan yapay zeka alanıyla birleştirmeyi amaçlamakta ve makine öğrenimi algoritmalarının öğrenci performans verilerinden daha derin anlayışlar elde etmek için nasıl kullanılabileceğine dair alternatif bir bakış açısı sunmayı hedeflemektedir. Veri analizinin eğitim çıktılarını ve deneyimlerini iyileştirmek için nasıl kullanılabileceğini araştırdığımız için eğitim ve teknolojinin entegrasyonu çalışmamızın ana odak noktasıdır.

1.2 Tezin Amacı

Bu projenin amacı, özellikle öğrenci akademik başarısını doğru bir şekilde anlamaya ve tahmin etmeye odaklanarak, eğitim verilerini incelemek ve değerlendirmek için gelişmiş makine öğrenimi algoritmalarını kullanmaktır. Bu araştırma, birbirini izleyen üç akademik not verme döneminin (G1, G2 ve G3) ortalamasını ifade eden yeni bir hedef değişken olan 'G_avg' uygulamasına odaklanmaktadır.
Bu araştırmanın temel amaçları şunlardır:

- Farklı makine öğrenimi algoritmalarının 'G_avg' puanını tahmin etmedeki doğruluğunu analiz etmek, böylece geleneksel tek sınava dayalı ölçümlere kıyasla öğrenci performansının daha kapsamlı bir değerlendirmesini sunmak.
- Öğrencileri 'G_avg' puanlarına göre başarı gruplarına ayırmada çeşitli makine öğrenimi modellerinin etkinliğini değerlendirmek.
- Farklı akademik ve akademik olmayan değişkenler ile öğrencilerin genel akademik başarıları arasındaki korelasyonları tanımlamak ve anlamak için bu makine öğrenimi algoritmalarının etkinliğini değerlendirmektir.

Çalışma aşağıdaki temel soruları ele almayı amaçlamaktadır:

- Makine öğrenimi algoritmaları, bir öğrencinin 'G_avg' ile ölçülen toplam akademik başarısını ne ölçüde güvenilir bir şekilde tahmin edebilir?
- Bu özel bağlamda regresyon analizi için en etkili makine öğrenimi yöntemleri hangileridir?
- Sınıflandırma ile ilgili olarak, bu algoritmalar öğrenci başarılarını farklı düzeylerde ne kadar doğru sınıflandırabilir?
- Bu modellerin uygulanmasından akademik başarıyı etkileyen faktörler hakkında ne gibi sonuçlar çıkarılabilir?

Bu tez, eğitim çıktılarının genel olarak daha iyi anlaşılmasını sağlamayı ve eğitimciler ile politika yapıcılara daha etkili eğitim teknikleri ve müdahaleleri geliştirmeleri için pratik bilgiler sunmayı da amaçlamaktadır.

1.3 Araştırmanın Önemi

Bu tezde yürütülen araştırma, iki Portekiz lisesinden öğrenci başarı verilerini kapsayan ve akademik başarıyı potansiyel olarak etkileyen çeşitli faktörlere kapsamlı bir genel bakış sunan bir veri setine dayanmaktadır. Titizlikle derlenen ve temizlenen bu veri seti yalnızca akademik notları değil, aynı zamanda çok çeşitli demografik, sosyal, ebeveyn ve okulla ilgili özellikleri de içermektedir. Böylesine çeşitli değişkenler, eğitim sonuçlarının derinlemesine analizi için verimli bir zemin sağlamaktadır.

Bu çalışmanın önemi çok yönlüdür. İlk olarak, akademik başarıyı ölçmek için yeni bir yaklaşım benimsemekte ve odağı final notundan (G3) daha kapsayıcı bir ölçüt olan ve üç temel dönemdeki notların aritmetik ortalaması olan G_avg'ye çevirmektedir. Bakış açısındaki bu değişiklik, tek bir sonuç değerlendirmesinden ziyade akademik deneyimlerinin tamamını göz önünde bulundurarak öğrenci performansının daha bütünsel bir şekilde anlaşılmasını teşvik etmektedir.

İkinci olarak, gelişmiş makine öğrenimi algoritmalarının bu zengin veri setine uygulanması, öğrenci başarısını etkileyen çeşitli faktörlerin karmaşık etkileşimine ilişkin değerli bilgiler sağlamaktadır. Bu araştırma hem regresyon hem de sınıflandırma tekniklerini kullanarak, yalnızca 'G_avg' puanını önceden belirlemeyi değil, aynı zamanda öğrencileri başarı düzeylerine göre kategorize etmeyi ve böylece akademik başarıya ilişkin ayrıntılı bir görünüm sunmayı amaçlamaktadır

Dahası, bu araştırmanın bulguları eğitimciler ve politika yapıcılar için pratik sonuçlar doğurabilir. Öğrenci performansının dinamiklerini ve bunu etkileyen faktörleri anlamak, eğitim stratejisi ve politika formülasyonunda daha bilinçli kararlar alınmasını sağlayabilir. Örneğin, akademik başarının temel belirleyicilerinin tanımlanması, düşük performans gösterme riski altında olabilecek öğrencileri desteklemek için hedeflenen müdahaleleri bilgilendirebilir.
Özünde bu çalışma, makine öğrenimi merceğinden akademik performansın kapsamlı bir analizini sağlayarak eğitim veri analizi alanına katkıda bulunmaktadır. Ham veriler ile eyleme

dönüştürülebilir içgörüler arasındaki boşluğu doldurarak, eğitim stratejilerinin öğrenci başarısını en üst düzeye çıkarmak için nasıl uyarlanabileceğine dair yeni bir bakış açısı sunuyor.

1.4 Varsayımlar

İki Portekiz lisesinden öğrenci verilerini kapsayan veri setinin, benzer eğitim bağlamlarındaki daha geniş öğrenci nüfusunu temsil ettiği varsayılmaktadır. Dahil edilen demografik, sosyal ve akademik değişkenlerin akademik başarı ile ilgili faktörlerin kapsamlı bir görüntüsünü sağladığı varsayılmaktadır.

Bu araştırma, veri setinde toplanan ve sağlanan verilerin doğru ve tutarlı bir şekilde kaydedildiğini varsaymaktadır. Buna öğrenci notlarının, demografik bilgilerin ve diğer özelliklerin doğruluğu da dahildir.

Öğrenci performansını etkileyen faktörlerin veri setinin kapsadığı dönem boyunca nispeten sabit kaldığı varsayılmaktadır. Bu, sosyo-ekonomik değişiklikler veya eğitim politikalarındaki değişimler gibi veri setinde hesaba katılmayan dış değişkenlerin analiz edilen temel ilişkileri önemli ölçüde etkilemediği anlamına gelmektedir.

Araştırma, seçilen makine öğrenimi modellerinin veri setini analiz etmek için uygun ve etkili olduğunu varsaymaktadır. Bu, algoritmaların farklı değişkenler ile hedef değişken 'G_avg' arasındaki karmaşık ilişkileri yeterince yakalayabileceği varsayımını da içermektedir. Çalışma Portekiz okullarından elde edilen verilere dayanmakla birlikte, elde edilen bulguların ve içgörülerin, belirli bağlamsal faktörlerin farklılık gösterebileceği kabul edilerek, benzer eğitim ortamlarına genellenebileceği varsayılmaktadır.

Araştırma, verilerin etik bir şekilde toplandığını ve kullanıldığını, öğrenci bilgilerinin mahremiyetine ve gizliliğine saygı gösterildiğini varsaymaktadır. Ayrıca veri setinin sonuçları gereksiz yere etkileyebilecek herhangi bir önyargı içermediği de varsayılmaktadır.

1.5 Sınırlamalar

Bu araştırma, yaklaşımı bakımından kapsamlı olmakla birlikte, kabul edilmesi önemli olan bazı sınırlamalara tabidir:

Coğrafi ve Kültürel Özgünlük: Bu çalışmanın birincil sınırlaması, yalnızca Portekiz'deki iki liseden elde edilen bir veri setine dayanmasıdır. Sonuç olarak, bulgular bu coğrafi bölgeye özgü kültürel, eğitimsel ve sosyoekonomik bağlamlardan etkilenmiş olabilir. Bu durum, sonuçların farklı kültürel ve sosyoekonomik arka planlara sahip diğer eğitim ortamlarına genellenebilirliğini sınırlamaktadır.

Veri Kapsamı ve Derinliği: Veri seti, birçok açıdan kapsamlı olmasına rağmen, akademik performansı etkileyen tüm olası faktörleri kapsamayabilir. Psikolojik iyi olma hali, ayrıntılı aile dinamikleri ve özel öğretim metodolojileri gibi değişkenler veri setinde açıkça yer almamaktadır ve dolayısıyla bu çalışmada analiz edilmemiştir.

Zamansal Sınırlamalar: Veriler belirli bir dönemi temsil etmektedir. Zaman içinde eğitim politikalarında, sosyal normlarda veya ekonomik koşullarda meydana gelen değişiklikler, bulguların gelecekteki bağlamlarla ilgisini ve uygulanabilirliğini etkileyebilir.

Algoritmik Kısıtlamalar: Kullanılan makine öğrenimi modellerinin etkinliği, sağlanan verilerin kalitesine ve niteliğine bağlıdır. Bu modellerin doğal sınırlamaları vardır ve veri setindeki çeşitli faktörler arasındaki tüm nüanslı etkileşimleri tam olarak hesaba katamayabilir.

Tahmine Dayalı Sınırlamalar: Bu çalışmada kullanılan makine öğrenimi modelleri örüntüleri tanımlama ve geçmiş verilere dayalı tahminlerde bulunma konusunda becerikli olsa da, gelecekteki eğilimleri tahmin etme veya akademik performansı önemli ölçüde etkileyebilecek benzeri görülmemiş olayları hesaba katma becerileri sınırlıdır.

Etik ve Gizlilikle İlgili Hususlar: Araştırma, etik veri toplama ve kullanım uygulamalarını varsaymaktadır. Veri toplamadaki potansiyel önyargılar veya ele alınmamış gizlilik endişeleri, çalışmanın bulgularının etik sonuçlarını sınırlayabilir.

1.6 Tanımlar

G_avg (Hedef Değişken): 'G_avg' bir öğrenci için üç temel akademik not döneminin (G1, G2 ve G3) aritmetik ortalamasını temsil eden türetilmiş bir değişkendir. Çalışmada birincil hedef değişken olarak görev yapmakta ve öğrenci akademik performansının bütünsel bir ölçüsünü sunmaktadır.

Makine Öğrenimi Algoritmaları: Bunlar, makinelerin verilerden öğrenmesini ve verilere dayalı tahminler veya kararlar almasını sağlayan hesaplama yöntemleridir. Bu tezde, eğitim verilerini analiz etmek amacıyla regresyon ve sınıflandırma görevleri için çeşitli makine öğrenimi algoritmaları kullanılmaktadır.

Regresyon Analizi: Bağımlı (hedef) ve bağımsız (tahmin edici) değişken(ler) arasındaki ilişkiyi araştıran bir tür tahmine dayalı modelleme tekniğidir. Bu bağlamda, regresyon analizi 'G_avg' sürekli sonucunu tahmin etmek için kullanılır.

Bu, verileri önceden tanımlanmış sınıflara veya kategorilere ayırma sürecini ifade eder. Bu çalışmada, öğrencileri 'G_avg' puanlarına göre başarı düzeylerine ayırmak için sınıflandırma analizi kullanılmıştır.

Veri ön işleme, ham verilerin analiz için daha uygun bir formata dönüştürülmesi sürecidir. Bu işlem temizleme, normalleştirme, dönüştürme ve özellik mühendisliğini içerir. Özellik mühendisliği, ham verilerden özellikler (karakteristikler, özellikler ve nitelikler) çıkarmak için alan bilgisini kullanma sürecidir. Bu tezde özellik mühendisliği, mevcut not verilerinden 'G_avg' değişkenini oluşturmak için kullanılmıştır. Bunlar, bir öğrencinin eğitim çalışmalarındaki performansını ölçmek için kullanılan metrikler veya değişkenlerdir. Bu veri setinde bunlar notları, devamsızlıkları ve okulla ilgili diğer özellikleri içerir.

Gizlilik, rıza ve adaleti göz önünde bulundurarak verilerin sorumlu ve saygılı bir şekilde kullanılmasını ifade eder. Bu tez, kullanılan verilerin etik standartlara uygun olduğunu varsaymaktadır.

1.7 Evren Örneği

Bu çalışmanın ana odağı Portekiz'deki iki liseye devam eden öğrencilerin bilgilerini içeren bir veri kümesidir. Bu veri kümesi, demografik, sosyal ve okulla ilgili özelliklerin yanı sıra öğrencinin akademik kayıtlarının ayrıntılı bir görünümünü sunar. Bu araştırmanın daha geniş bağlamı genel olarak lise öğrencilerini kapsamaktadır. Örneklem Portekiz okullarından olmakla birlikte, bu çalışmada kullanılan bulgular ve metodolojiler, çeşitli eğitim ortamlarında lise öğrencilerinin performansını anlamaya yönelik sonuçlar doğurabilir.

Bu araştırmanın metodolojisi birkaç temel bileşeni kapsamaktadır. Çalışma, makine öğrenimi algoritmalarını kullanarak öğrencilerin akademik performansını analiz etmek için regresyon ve sınıflandırma modellerinin bir kombinasyonunu kullanmaktadır.

Örneklem, Portekiz'deki iki okuldan lise öğrencilerinden oluşmakta ve analiz için çeşitli veriler sağlamaktadır. Veri toplama için birincil araç, öğrenci performansı ve geçmişiyle ilgili çok çeşitli değişkenler içeren **student-mat.csv** veri setidir. Veri seti, okul raporları ve anketler kullanılarak derlenmiş ve ilgili eğitim verilerinin kapsamlı bir şekilde toplanması sağlanmıştır.

Analiz, verilerin ön işlemden geçirilmesini, özellik mühendisliğini ('G_avg' değişkeninin oluşturulması) ve akademik performansı anlamak ve tahmin etmek için çeşitli makine öğrenimi algoritmalarının uygulanmasını içermektedir. Çalışma ayrıca, yeni tanımlanan 'G_avg' ile ölçüldüğü üzere, çeşitli faktörler ile genel öğrenci başarısı arasındaki ilişkileri de araştırmaktadır.

2 ARKAPLAN

2.1 Akademik Başarıyı Etkileyen Faktörler

Bu çalışmada, 'G_avg' değişkeni ile ölçülen akademik başarıyı etkileyen faktörler incelenmiştir. Bu analiz hem 'G_avg' ile en yüksek korelasyona sahip belirli özellikleri hem de tüm özelliklerin daha geniş bir kategorizasyonunu içermektedir. Aşağıdaki kategoriler akademik başarıyı etkileyen kilit alanları temsil etmektedir:
Bu, 'G_avg' ile güçlü korelasyonlar gösteren önceki notları (G1, G2 ve G3) içerir. Bu notlar akademik performansın doğrudan göstergeleridir ve genel akademik başarıyı anlamak için kritik öneme sahiptir.

Ebeveynlerin eğitimi (Medu ve Fedu), aile büyüklüğü ve ebeveynlerin iş türleri (Mjob ve Fjob) gibi faktörler akademik başarı ile kayda değer korelasyonlara sahiptir. Bu unsurlar, aile ortamının ve ebeveyn desteğinin öğrenci başarısı üzerindeki etkisini yansıtmaktadır.

Yüksek öğrenim arzusu (higher_yes) ve eğitim desteğinin mevcudiyeti (school_support, family_support) önemli rol oynamaktadır. Bu faktörler, öğrencilerin akademik performanslarını etkileyebilecek isteklerini ve mevcut destek sistemlerini göstermektedir.

Çalışma süresi, internet erişimi, etkinliklere katılım ve sağlık durumu gibi unsurlar, bir öğrencinin akademik başarısını etkileyebilecek alışkanlıkları ve yaşam tarzı hakkında fikir vermektedir. Öğrencinin okulu, okulu seçme nedeni ve okula seyahat süresi (adress_type, reason, travel time) de öğrencinin öğrenim gördüğü daha geniş çevresel bağlamı yansıttığı için önemlidir. Sosyal alışkanlıklar (goout olarak arkadaşlarla dışarı çıkma, Dalc ve Walc olarak alkol tüketimi) ve romantik ilişkiler (romantic_yes) dahil edilerek sosyal yaşam ve eğlence faaliyetlerinin akademik performans üzerindeki etkisi vurgulanmaktadır. Yaş, cinsiyet ve vasi özellikleri gibi demografik faktörler analiz için bağlamsal bir bağlam sağlamaktadır. Bu kategorilerin her biri veri setindeki birden fazla değişkeni kapsamakta ve akademik başarıya katkıda bulunan faktörlere çok yönlü bir bakış açısı sağlamaktadır. Bu tezdeki analiz, örüntüleri ve ilişkileri ortaya çıkarmak için makine öğrenimi modellerini kullanarak bu faktörlerin 'G_avg' puanı üzerindeki etkisini keşfetmeyi amaçlamaktadır.

2.2 Eğitimsel Veri Analizinde Büyük Veri ve Makine Öğrenimi

Gelişmiş bilgi işlemin ortaya çıkması ve çeşitli alanlarda verilerin çoğalmasıyla birlikte, büyük veri ve makine öğreniminin eğitim araştırmalarındaki rolü giderek daha önemli hale gelmiştir. Büyük veri kümelerini işleme ve analiz etme yeteneği, akademik performansı anlamak ve tahmin etmek için yeni yollar açmış ve daha önce erişilemeyen içgörüler sağlamıştır. Eğitimde büyük veri kavramı, öğrenci demografisi, akademik performans, davranış kalıpları ve daha fazlası dahil olmak üzere eğitim sistemleri aracılığıyla toplanan büyük miktarda bilgiyi kapsar. Bu zengin veri derlemesi, eğitim sürecinin çeşitli yönlerini keşfetmek için eşsiz bir fırsat sunmaktadır. Boyd ve Crawford'un vurguladığı gibi, büyük verinin gücü sadece boyutunda değil, aynı zamanda anlamlı içgörüler elde etmek için bu büyük veri kümelerini arama, bir araya getirme ve çapraz referanslama yeteneğinde yatmaktadır.

Bu büyük veri analizinin merkezinde, verilerdeki kalıpları, ilişkileri ve eğilimleri belirlemek için algoritmaları kullanmaya odaklanan yapay zekanın bir alt kümesi olan makine öğrenimi yer almaktadır. Matematiksel ve istatistiksel modellere dayanan bu algoritmalar, akademik performans, öğrenci katılımı ve eğitim müdahalelerinin etkinliği gibi sonuçların tahmin edilmesini sağlar.

Bu çalışma kapsamında, öğrenci performansı veri setindeki 'G_avg' değişkenini analiz etmek için makine öğrenimi algoritmaları kullanılmıştır. Bu, yalnızca akademik sonuçları tahmin etmek için değil, aynı zamanda bu sonuçlara katkıda bulunan altta yatan faktörleri anlamak için çeşitli algoritmaların uygulanmasını içerir. Bu çalışmada kullanılan algoritmalar, temel regresyon modellerinden daha karmaşık sınıflandırıcılara kadar uzanmaktadır ve her biri veri kümesinin belirli özelliklerini ele alma yeteneklerine göre seçilmiştir.

Makine öğreniminin dinamik doğası, onu eğitimsel veri analizi için çok uygun hale getirmektedir. Deneyimlerin ve etkileşimlerin bilgiyi sürekli olarak şekillendirdiği insan öğreniminde olduğu gibi, makine öğrenimi algoritmaları da daha fazla veriye ve farklı senaryolara maruz kaldıkça adapte olur ve gelişir. Bu uyarlanabilirlik, öğrenci verilerinin sürekli geliştiği eğitim ortamlarında çok önemlidir.

2.3 Makine Öğrenimi

Yapay zekanın (AI) kritik bir alt kümesi olan makine öğrenimi (ML), karar vermek için verilerden öğrenen sistemlerin geliştirilmesinde çok önemlidir. Karmaşık kalıpları tanımlamak ve tahminlerde bulunmak veya veri girişlerinden eylemler önermek için algoritmalar kullanır. Büyük veri analizinde önemli bir araç olarak ML algoritmaları, öznitelikle eğitim ortamlarında büyük ve karmaşık veri kümelerinin incelikli yorumlanmasına olanak tanır (Boyd and Crawford)

ML'deki birincil yöntem olan denetimli öğrenme, bilinen sonuçları olan veri kümeleri üzerinde eğitim modellerini içerir. Bu yaklaşım, bu tezin merkezini oluşturan tahmine dayalı modellemede çok önemlidir. Akademik başarıyı etkileyen faktörlerin anlaşılmasına yardımcı olarak girdi özniteliklerinin bilinen hedeflerle eşleştirilmesine olanak tanır (Baker and Inventado).

Regresyon görevleri sürekli sonuçları tahmin etmeye odaklanır. Scikit tarafından sağlanan *Ridge* Regresyonu, *Lasso* Regresyonu, Elastik Net, Rassal Orman ve Gradyan Artırma algoritmaları girdi değişkenleri ile sürekli hedefler arasındaki ilişkileri kurmak için kullanılır. Bu çalışmada, bu tür modeller 'G_avg' puanını tahmin etmek için uygulanmıştır ve bu da bunların eğitimsel veri analizindeki faydasını göstermektedir (Kefalaki et al.).

Ayrık sonuçları tahmin etmek için kullanılan sınıflandırma algoritmaları Logistik Regresyon , Karar Ağaçları, Rassal Orman, Destek Vektör Makineleri (SVM) Gradyan Artırma ve Sinir Ağı

algoritmalarını içerir. Bu yöntemler, bu araştırmada öğrencileri akademik performansa göre sınıflandırmak için kullanılan girdi özniteliklerine dayalı olarak verileri farklı sınıflara ayırır (Kefalaki et al.).

Denetimsiz öğrenme, etiketlenmemiş veriler üzerinde eğitim modellerini içerir. Model, kümeleme ve temel bileşen analizi gibi teknikleri kullanarak, istenen sonuca rehberlik etmeden kalıpları ve ilişkileri ortaya çıkarır. Bu yöntem eğitimsel verilerdeki gizli kalıpların keşfedilmesi açısından önemlidir (Kovac).

Etiketli ve etiketsiz verileri birleştiren Yarı Denetimli Öğrenme, etiketli veri toplamanın pratik olmadığı durumlarda faydalıdır. Deneme yanılmaya dayalı, ödüller ve cezalarla yönlendirilen Takviyeli Öğrenme, eğitim bağlamlarında ilgi çeken başka bir yaklaşımdır.
Bu çalışmada eğitim verilerini analiz etmek için ML'dan yararlanılmıştır. Araştırma, çeşitli denetimli öğrenme algoritmaları uygulayarak akademik başarıyı etkileyen kalıpları belirlemeyi ve gelecekteki sonuçları tahmin etmeyi amaçlamaktadır. ML'nin büyük, çok boyutlu veri kümelerini işleme ve yeni verilere uyum sağlama yeteneği, onu öznitelikle eğitim araştırmaları için uygun kılmaktadır (Waheed et al.).

Makine öğrenimi (ML), yapay zeka kapsamında, açık talimatlar olmadan görevleri genelleştirebilen ve gerçekleştirebilen istatistiksel algoritmalar geliştirmeye odaklanan bir çalışma alanıdır. Verilerden öğrenebilen ve verilere dayalı tahminler veya kararlar alabilen modeller oluşturmayı içerir. Makine öğrenimi algoritmaları, büyük dil modelleri, bilgisayarlı görme, konuşma tanıma ve daha fazlası dahil olmak üzere çeşitli alanlarda uygulanmıştır. Makine öğreniminin matematiksel temelleri matematiksel optimizasyon ve hesaplamalı istatistiklere dayanmaktadır. Makine öğrenimi, denetimsiz öğrenme yoluyla keşfedici veri analizine odaklanan veri madenciliği ile yakından ilgilidir. İş bağlamında makine öğrenimine genellikle tahmine dayalı analitik denir (Viberg et al.; Rebala et al.).

"Makine öğrenimi" terimi 1959'da IBM çalışanı Arthur Samuel tarafından icat edildi. 1960'lardaki ilk makine öğrenimi çabaları arasında, temel pekiştirmeli öğrenmeyi kullanarak verilerdeki kalıpları analiz eden Cybertron gibi deneysel "öğrenme makineleri" vardı. 1980'lere gelindiğinde makine öğrenimi alanındaki araştırmalar, belirli görevleri gerçekleştirmek için verilerden öğrenebilen algoritmaların geliştirilmesine doğru kaydı. Tom M. Mitchell, bir bilgisayar programının, performans ölçümüyle ölçülen görevlerdeki performansının deneyimle birlikte gelişmesi durumunda, bazı görev sınıflarına ve performans ölçümüne ilişkin deneyimlerden öğrendiğinin söylendiğini belirterek resmi bir tanım yaptı (Kreuzberger et al.).

Modern makine öğreniminin iki temel amacı vardır: verileri geliştirilen modellere göre sınıflandırmak ve bu modellere dayanarak gelecekteki sonuçlara ilişkin tahminlerde bulunmak. Örneğin, hisse senedi alım satımına yönelik bir makine öğrenimi algoritması, yatırımcılara gelecekteki potansiyel tahminler hakkında bilgi verebilir. Makine öğreniminin alt kategorileri denetimli öğrenmeyi, denetimsiz öğrenmeyi ve takviyeli öğrenmeyi içerir (Dueben et al.).

Makine öğrenimi uygulamaları, pratik iş çözümlerinden otonom araçlar ve tıbbi teşhis gibi karmaşık görevlere kadar çeşitlilik gösterir. Bununla birlikte, öznitelikle algoritmik karar vermede adaleti sağlama ve *Bias*dan kaçınma konusunda etik hususları da gündeme getirir (Farouni).

Makine öğreniminin tanımı ve kapsamı, başlangıcından bu yana önemli ölçüde gelişti. Örüntü tanıma ve basit öğrenme algoritmalarının ilk günlerinden itibaren makine öğrenimi, çeşitli uygulamalara sahip karmaşık bir alana dönüştü. Teorik araştırmalardan pratik problem çözme yaklaşımlarına geçiş, makine öğrenimini birçok modern teknolojinin ayrılmaz bir parçası haline getirdi. Makine öğreniminin özellikle adalet ve *Bias* açısından etik sonuçları, teknoloji toplumda yaygınlaştıkça giderek daha önemli hale geldi (Raschka).

2.3.1 Denetimli Öğrenme

Denetimli öğrenme, bir modelin etiketli veriler üzerinde eğitildiği makine öğreniminde bir paradigmadır. Eğitim süreci, girdi değişkenlerini (X) ve bir çıktı değişkenini (Y) içerir; model, Y = f(X) olarak ifade edilen, girdiden çıktıya eşleme fonksiyonunu öğrenir. Amaç, modelin yeni, görünmeyen girdi verileri için çıktı değişkenlerini tahmin edebilmesi için bu eşleme fonksiyonuna yaklaşık değer vermektir. Denetimli öğrenmede model, girdi öznitelikleri örnekleri arasındaki eşlemeyi bunlarla ilişkili etiketlerle eşleştirmeyi öğrenir. Bu örneklerle eğitildiğinde model, görünmeyen veriler üzerinde yeni tahminler yapabilir. Çıktı sayısal (regresyon modelleri) veya kategorik (sınıflandırma modelleri) olabilir.

Denetimli öğrenme, basit yaklaşımı ve çeşitli uygulamalardaki etkinliği nedeniyle yaygın olarak kullanılmaktadır. Etiketli verilerin kullanılabilirliği, denetimli öğrenme modellerinin başarısında çok önemli bir rol oynar. Ancak eğitim verilerinin kalitesi ve miktarı, modelin yeni verilere genelleme yeteneğini önemli ölçüde etkileyebilir. Özniteliklerin seçimi ve girdi verilerinin temsili de denetimli öğrenme modellerinin performansında kritik faktörlerdir.

2.3.1.1 Eğitim Seti

Makine öğreniminde eğitim seti çok önemli bir bileşendir. Giriş-çıkış çiftlerinden oluşan bir modeli eğitmek için kullanılan bir veri kümesidir. Eğitim seti, modelin girdiler ve çıktılar arasındaki ilişkileri öğrenmesine yardımcı olur ve esas olarak modele nasıl tahmin veya karar vereceğini öğretir. Eğitim seti, sinir ağlarındaki ağırlıklar gibi modelin parametrelerini uydurmak için kullanılır. Eğitim setinin kalitesi, çeşitliliği ve boyutu, modelin genelleme yapma ve görünmeyen veriler üzerinde doğru performans gösterme yeteneğini önemli ölçüde etkiler. Eğitim seti, makine öğreniminin temelini oluşturur ve modelin öğrendiği birincil kaynak olarak hizmet eder. Eğitim setinin temsil edilebilirliği ve kalitesi, modelin yeni verilere iyi bir şekilde genelleştirilebilmesi için hayati öneme sahiptir. Bununla birlikte, modelin eğitim setine çok yakın uyarlanması durumunda aşırı uyum gibi zorluklar ortaya çıkabilir ve bu da iyi dengelenmiş ve çeşitli bir eğitim veri setine olan ihtiyacın altını çizer.

2.3.1.2 Test Seti

Makine öğrenmesinde test seti, eğitilmiş bir modelin performansını değerlendirmek için kullanılan bir veri setidir. Modelin yeni, görülmemiş verilere genelleme yapma yeteneğinin tarafsız bir değerlendirmesini sağlar. Eğitim setinden farklı olarak test seti, model eğitim aşamasında kullanılmaz ve modelin tahmin gücünün doğru bir şekilde ölçülmesini sağlamak için ayrı tutulur. Test seti, modelin eğitim verilerinden ne kadar iyi öğrendiğini ve daha önce hiç görmediği veriler üzerinde nasıl performans gösterdiğini belirlemeye yardımcı olur.

Test seti, makine öğrenimi iş akışında bir modelin performansının nihai yargıcı olarak görev yapan önemli bir bileşendir. Modelin genelleme yeteneğinin tarafsız bir değerlendirmesini sağlamadaki rolü, modelin etkinliğini ve pratik senaryolarda uygulanabilirliğini anlamak açısından da çok önemlidir. Ancak test setinin kalitesi ve temsil edilebilirliği, modelin karşılaşacağı gerçek dünya koşullarını doğru bir şekilde yansıtmasını sağlamak açısından kritik öneme sahiptir.

2.3.1.3 Doğrulama Veri seti

Makine öğreniminde doğrulama seti, model eğitim sürecinin çok önemli bir parçasıdır. Modelin hiperparametrelerini ayarlarken eğitim veri setine uygun bir modelin tarafsız bir değerlendirmesini sağlamak için kullanılır. Doğrulama seti eğitim setinden ayrıdır ve modeli eğitmek için kullanılmaz. Bunun yerine, modelin performansını değerlendirmek ve model

parametrelerine ince ayar yapmak için kullanılır; bu, aşırı uyumun önlenmesine yardımcı olur ve modelin yeni verilere iyi bir şekilde genelleştirilmesini sağlar.

Doğrulama seti, makine öğrenimi hattında hayati bir rol oynar ve model seçimi ve ayarlama için bir araç görevi görür. *Bias* ve varyans arasında dengeyi sağlayan en iyi modelin ve hiperparametre ayarlarının belirlenmesine yardımcı olur. Doğrulama setinin kullanılması, özellikle sağlık ve biyoloji gibi doğru tahminlerin kritik olduğu alanlarda sağlam ve güvenilir makine öğrenimi modelleri geliştirmek için gereklidir.

2.3.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, algoritmaların etiketlenmiş yanıtlar olmadan veriler üzerinde eğitildiği bir makine öğrenimi türüdür. Amaç, veri içindeki yapıyı ve kalıpları keşfetmektir. Denetimli öğrenmenin aksine, denetimsiz öğrenme tahminler veya sonuçlarla çalışmaz, verilerin kendine özgü özniteliklerine odaklanır. Yaygın uygulamalar arasında kümeleme, boyutluluğun azaltılması ve birliktelik kuralının öğrenilmesi yer alır.

Denetimsiz öğrenme, etiketli verilerin az olduğu veya kullanılamadığı senaryolarda öznitelikle değerlidir. Verilerdeki gizli kalıpları keşfetme konusunda ustadır ve kümeleme ve boyut azaltma gibi görevler için çok önemlidir. Denetimsiz öğrenme, açık geri bildirim sinyallerinin eksikliği nedeniyle zorlayıcı olsa da, *Generative Adversarial Networks*(GAN)'lar ve karşılaştırmalı öğrenme gibi algoritmalardaki gelişmeler, öğrenmenin yeteneklerini önemli ölçüde artırdı.

2.3.3 Yarı Denetimli Öğrenme

Yarı denetimli öğrenme, eğitim sırasında az miktarda etiketli veriyi büyük miktarda etiketsiz veriyle birleştiren bir makine öğrenmesi yaklaşımıdır. Bu yöntem öznitelikle tamamen etiketlenmiş bir veri kümesinin elde edilmesinin pahalı veya pratik olmadığı durumlarda faydalıdır. Yarı denetimli öğrenme genellikle tıbbi görüntüleme veya doğal dil işleme gibi etiketli verilerin elde edilmesinin maliyetli olduğu durumlarda kullanılır.

Yarı denetimli öğrenme, hem etiketli hem de etiketsiz verilerden yararlanarak denetimli ve denetimsiz öğrenme arasındaki boşluğu doldurur. Etiketli verilerin sınırlı olduğu veya elde edilmesinin maliyetli olduğu senaryolarda özellikle değerlidir. *FixMatch* ve *MixMatch* gibi yöntemler, yarı denetimli öğrenmenin tam denetimli yöntemlere yakın performans elde edebildiğini gösterdi ve bu da onu makine öğreniminde güçlü bir araç haline getiriyor.

2.3.4 Takviyeli Öğrenme

Tanım ve Uygulamalar : Takviyeli öğrenme (*Reinforcement Learning*, RL), bir aracının bir hedefe ulaşmak için bir ortamda eylemler gerçekleştirerek karar vermeyi öğrendiği bir makine öğrenimi türüdür. Temsilci, ödüller veya cezalar şeklinde geri bildirim alır ve kümülatif ödülleri en üst düzeye çıkarmayı öğrenir. Denetimli öğrenmenin aksine, RL etiketli giriş/çıkış çiftlerine ihtiyaç duymaz ve optimumun altındaki eylemlerin açık bir şekilde düzeltilmesine ihtiyaç duymaz. Takviyeli öğrenme, etkileşim yoluyla öğrenmeye odaklanması ve karmaşık, dinamik ortamlara uyum sağlama yeteneği ile ayırt edilir. Özellikle karar verme sırasının kritik olduğu ve çevre dinamiklerinin bilindiği veya öğrenilebildiği problemler için çok uygundur. RL'nin çeşitli alanlardaki, özellikle de oyun ve robotikteki başarısı, onun karmaşık gerçek dünya sorunlarını çözme potansiyelinin altını çiziyor.

2.4 Çoklu Doğrusallık ve Öznitelik Seçimi

Makine öğreniminde çoklu bağlantı, bir modeldeki iki veya daha fazla yordayıcı değişkenin yüksek düzeyde ilişkili olduğu ve her değişkenin yanıt değişkeni üzerindeki etkisini izole etmeyi zorlaştırdığı olguyu ifade eder. Öznitelik seçimi, model yapımında kullanılmak üzere ilgili özniteliklerin bir alt kümesinin tanımlanması ve seçilmesi sürecidir. Aşırı uyumun azaltılmasına, model doğruluğunun arttırılmasına ve eğitim süresinin azaltılmasına yardımcı olur.

2.4.1 Regresyon Katsayısı

Metinde açıklandığı gibi regresyon katsayısı kavramı, makine öğrenimi ve istatistikte regresyon analizinin temel bir yönüdür. Sunulan açıklama, doğrusal regresyon modelleri bağlamında regresyon katsayıları, R-kare (R^2) ve varyans şişirme faktörünün (VIF) standart tanımları ve yorumlarıyla uyumludur. İşte bazı önemli noktalar ve bunlara karşılık gelen kaynaklar:

Regresyon Katsayısı ve En Küçük Kareler Yöntemi: Regresyon katsayısı, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkinin gücünü ve yönünü belirlemek için istatistiksel modellemede kullanılan bir ölçüdür. Bahsi geçen en küçük kareler yöntemi, artıkların karelerinin toplamını (gözlenen ve tahmin edilen değerler arasındaki farklar) en aza indirerek regresyon doğrusunu bulmaya yönelik standart bir yaklaşımdır. Bu kavram istatistiksel öğrenmede yaygın olarak tartışılmaktadır ve James, Witten, Hastie ve Tibshirani tarafından yazılan "An Introduction to Statistical Learning" gibi kaynaklarda bulunabilir (Tibshirani et al.).

R-kare (Belirleme Katsayısı): R^2 veya belirleme katsayısı, bağımlı değişkendeki bağımsız değişkenlerden tahmin edilebilen varyansın oranını ölçer. Bir regresyon modelinin uyumunun değerlendirilmesinde önemli bir ölçümdür. R^2'nin model uyumunun bir ölçüsü olarak yorumlanması, Kutner, Nachtsheim , Neter ve Li'nin "Uygulamalı Doğrusal İstatistiksel Modeller" gibi kaynaklarda detaylandırıldığı gibi regresyon analizinde standart bir konudur (Kutner et al.).

Düzeltilmiş R-kare: Düzeltilmiş R^2, modeldeki öngörücülerin sayısını hesaba katmak için çoklu regresyon modelleri bağlamında kullanılır. R^2 değerini değişken sayısına ve örneklem büyüklüğüne göre ayarlayarak birden fazla değişken kullanıldığında model uyumuna ilişkin daha doğru bir ölçüm sağlar (Allison).

Varyans Enflasyon Faktörü (VIF): VIF, regresyon analizindeki çoklu bağlantının boyutunu ölçen bir ölçüdür. Tahmin edicilerin korelasyonu durumunda, tahmin edilen regresyon katsayısının varyansının ne kadar arttığını değerlendirir. VIF'nin yorumlanması ve çoklu doğrusallığın teşhisindeki rolü, Samprit Chatterjee ve Ali S. Hadi öğrenimi alanında standart referanslardır (Chatterjee and Hadi).

2.4.2 Öznitelik Seçim

Öznitelik seçimi, modelin performansını doğrudan etkilediği için makine öğreniminde çok önemlidir. Doğru özniteliklerin seçilmesi daha basit, daha yorumlanabilir ve daha genellenebilir modellere yol açabilir. Bu, Alice Zheng ve Amanda Casari tarafından yazılan "Makine Öğrenimi için Öznitelik Mühendisliği: Veri Bilimcileri için İlkeler ve Teknikler" gibi kaynaklarda tartışılmaktadır (Zheng and Casari).

Metin, aynı veri kümesinde bile farklı öznitelik kombinasyonlarından elde edilen sonuçların çeşitliliğinden bahseder. Bu, özniteliklerin sayısının ve doğasının model karmaşıklığını ve performansını önemli ölçüde etkileyebileceği kavramıyla uyumludur. Öznitelik seçiminin model karmaşıklığı ve aşırı uyum üzerindeki etkisi, Hastie, Tibshirani ve Friedman'ın "İstatistiksel Öğrenmenin Öğeleri"nde görüldüğü gibi makine öğrenimi literatüründe standart bir konudur (Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*)

Doğru öznitelik seçimi, aşırı uyumu azaltmaya ve modelin performansını iyileştirmeye yardımcı olur. Bu, özellikle boyutluluk lanetinin modellerin genelleştirilememesine yol açabileceği yüksek boyutlu verilerde önemlidir. Bu kavram, Christopher Bishop'ın 'Pattern

Recognition and Machine Learning' (Desen Tanıma ve Makine Öğrenimi) adlı eserinde detaylandırılmıştır (Bishop and Nasrabadi).

Metin, bir model için en uygun öznitelikleri belirlemek amacıyla çeşitli öznitelik seçim tekniklerinin kullanımını ima etmektedir. Pang-Ning Tan, Michael Steinbach ve Vipin Kumar'ın "Veri Madenciliğine Giriş" kitabında detaylandırıldığı gibi sarmalayıcı yöntemler, filtre yöntemleri ve gömülü yöntemler gibi teknikler bu amaç için yaygın olarak kullanılmaktadır (Vipin).

Bu kaynaklar, makine öğreniminde öznitelik seçiminin öneminin kapsamlı bir şekilde anlaşılmasını sağlar ve bu alanda standart referanslardır.

2.4.2.1 Temel Veri Temizleme Yöntemleri

Veri temizleme, makine öğrenimi için öznitelik seçiminde temel bir adımdır. Eksik değerler, tutarsızlıklar ve ilgisiz veriler gibi sorunları ele alarak veri kümesini analize hazırlamayı içerir. Etkili veri temizleme, veri kümesinin kalitesini artırarak daha doğru ve güvenilir modellere yol açar.

Temizlik için kullanılan temel teknikler şunlardır:
Eksik Verilerin Ele Alınması: Eksik veriler, yüksek oranda eksik değere sahip özniteliklerin kaldırılması veya istatistiksel yöntemler kullanılarak eksik değerlerin atanması yoluyla giderilebilir. Örneğin, bir anket sorusuna verilen yanıtların önemli bir kısmı eksikse bu öznitelik veri kümesinden çıkarılabilir.

Kategorik Verilerle Başa Çıkmak: Tek tip veya oldukça değişken çıktılara sahip kategorik veriler dikkatle incelenmelidir. Tüm yanıtların aynı (veya tamamen farklı) olduğu öznitelikler, modele anlamlı bir katkı sağlamayabilir ve kaldırılabilir.

Değişken Eliminasyonu ve Seçimi: Bağımsız değişkenler ile hedef değişken arasındaki ilişki çok önemlidir. Bu ilişkilerin değerlendirilmesinde Pearson Korelasyonu ve Ki-kare Testi gibi istatistiksel yöntemler kullanılmaktadır. Pearson Korelasyonu sayısal değişkenler arasındaki doğrusal ilişkinin derecesini ölçerken, kategorik veriler için Ki-kare Testi kullanılır (Kuhn and Johnson).

ANOVA Testi: Varyans Analizi (ANOVA) testi, sayısal bir hedef değişkeni etkileyen kategorik bir değişkenin grupları arasında anlamlı farklılıklar olup olmadığını belirlemek için kullanılır.

Hedef değişkeni önemli ölçüde etkileyen değişkenlerin belirlenmesine yardımcı olur (Provost and Fawcett).

Varyans Enflasyon Faktörü (VIF): VIF, yordayıcı değişkenler arasındaki çoklu bağlantının tespit edilmesi için kullanılır. Yüksek VIF değeri değişkenler arasında güçlü bir korelasyon olduğunu gösterir ve bu da modelin performansını etkileyebilir (Kuhn and Johnson).

Sıralı Öznitelik Seçim Yöntemleri: Sıralı İleri Seçim (SFS) ve Sıralı Geriye Seçim (SBS) gibi teknikler, optimum öznitelik alt kümesini bulmak için değişkenleri yinelemeli olarak eklemek veya kaldırmak için kullanılır. Bu yöntemler modelin karmaşıklığını ve tahmin gücünü dengeler (James et al.).

Öznitelik seçiminde temel temizleme yöntemleri, etkili makine öğrenimi modelleri oluşturmak için kritik öneme sahiptir. Veri setinin yanıltıcı sonuçlara yol açabilecek hatalardan ve ilgisiz bilgilerden arınmış olmasını sağlarlar. Düzgün bir şekilde temizlenen ve seçilen öznitelikler, modelin performansına önemli ölçüde katkıda bulunarak bu adımı makine öğrenimi hattında vazgeçilmez kılar.

2.4.3 Boyut indirgeme

Boyut azaltma, bir veri kümesindeki giriş değişkenlerinin sayısını azaltmak için kullanılan makine öğrenimi ve istatistikteki bir süreçtir. Bazen ayrı bir kavram olarak ele alınsa da sıklıkla öznitelik seçimi tartışmalarına dahil edilir.

Temel Bileşen Analizi (PCA):
PCA, denetimsiz öğrenmede kullanılan doğrusal bir boyut azaltma tekniğidir. Verileri yeni bir koordinat sistemine dönüştürür, en fazla varyansı korurken boyut sayısını azaltır. PCA, yüksek derecede ilişkili değişkenleri, temel bileşenler adı verilen daha küçük, ilişkisiz değişkenler kümesinde birleştirmek için kullanılır. Kümeleme algoritmalarında ve çoklu bağlantı sorunlarının çözümünde yaygın olarak kullanılır.

Doğrusal Diskriminant Analizi (LDA):
LDA, boyutluluğun azaltılması için kullanılan denetimli bir öğrenme yöntemidir.
İki veya daha fazla olay sınıfını en iyi şekilde ayıran özelliklerin doğrusal kombinasyonlarını bulmayı amaçlar. LDA, amacın maksimum sınıf ayrılabilirliğine ulaşmak olduğu sınıflandırma problemlerinde özellikle faydalıdır. Mümkün olduğu kadar sınıf ayrımcılığına neden olan bilgileri korurken boyutları azaltır.

PCA ve LDA gibi boyut azaltma teknikleri, modelleri basitleştirmek ve performanslarını artırmak için görüntü işleme, biyoinformatik ve konuşma tanıma dahil olmak üzere çeşitli alanlarda kullanılmaktadır.

2.5 ALGORİTMALAR

2.5.1 Regresyon Algoritmaları

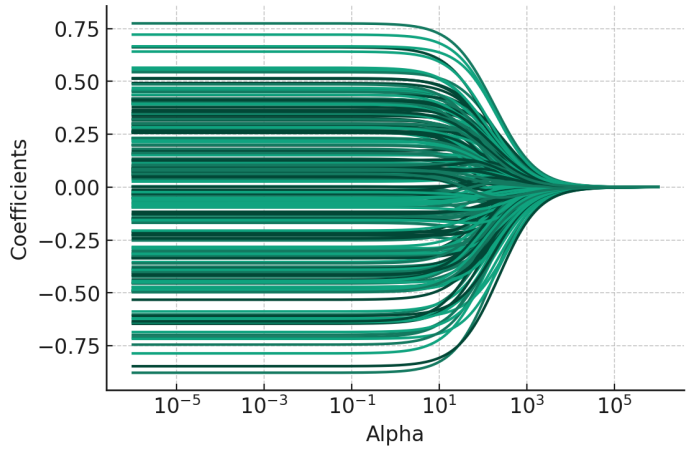
2.5.1.1 *Ridge* Regresyonu

Ridge Regresyonu, çoklu doğrusal bağımlılık sorunu olan çoklu regresyon verilerini analizinde kullanılan bir tekniktir. Çoklu doğrusal bağımlılık olduğunda, en küçük kareler tahminleri *Bias*sızdır, ancak varyansları büyüktür ve bu nedenle gerçek değerden uzak olabilirler. Regresyon tahminlerine bir miktar Bias ekleyerek, *Ridge* regresyonu standart hataları azaltır.

$$\text{Ridge}(\beta) = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Burada, λ ceza teriminin gücünü kontrol eden ayar parametresidir. Bu ceza terimi $\lambda \sum_{j=1}^{p} \beta_j^2$ maliyet fonksiyonuna eklenir ve katsayıları sıfıra doğru küçültür, ancak tam olarak sıfıra indirgemez.

Tikhonov düzenlileştirmesi olarak da bilinen *Ridge* Regresyon, çoklu bağlantı sıkıntısı olan çoklu regresyon verilerinin analizinde kullanılan bir yöntemdir. Çoklu doğrusallık oluştuğunda, en küçük kareler tahminleri tarafsızdır fakat varyansları büyüktür. Dolayısıyla gerçek değerden uzak olmaları mümkündür. Regresyon tahminlerine bir derece yanlılık ekleyerek, *Ridge* regresyonu standart hataları azalma sağlar.



Şekil 1: Regularizasyonun Bir Fonksiyonu Olarak Ridge Katsayıları

Bu görselde *Ridge* regresyon modelinin katsayılarının, düzenlileştirme parametresi olan alfa'nın farklı değerleriyle nasıl değiştiğini göstermektedir. Alfa arttıkça, katsayıların büyüklüğü azalır, bu da *Ridge* regresyonunun küçültme etkisini gösterir.

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.
Ridge regresyonu, hata fonksiyonuna bir ceza terimi (katsayıların büyüklüğünün karesi) ekler. Bu ceza terimi, modelin daha küçük katsayıları tercih edeceği şekilde katsayıları düzenler. *Ridge* regresyonundaki anahtar parametre lambda (λ) veya alfa olarak bilinir. Bu parametre büzülme miktarını kontrol eder: lambda değeri ne kadar büyük olursa, büzülme miktarı da o kadar büyük olur.
Bias-Varyans Dengesi: Tahminlere Bias ekleyerek, *Ridge* regresyonu tahminlerin varyansını önemli ölçüde azaltabilir ve daha iyi uzun vadeli tahmin doğruluğuna yol açabilir.
Uygulamalar:

- *Ridge* regresyonu, bireysel öngörücülerin birbirleriyle yüksek düzeyde korelasyona sahip olduğu verilerde çoklu bağlantı sorunu yaşandığında kullanılır.
- Ayrıca parametre sayısının gözlem sayısını aştığı durumlarda da kullanışlıdır.

Avantajları:

- Modelin karmaşıklığını azaltarak daha iyi tahminler üretebilir.
- Yüksek korelasyonlu bağımsız değişkenlerle uğraşırken özellikle faydalıdır.

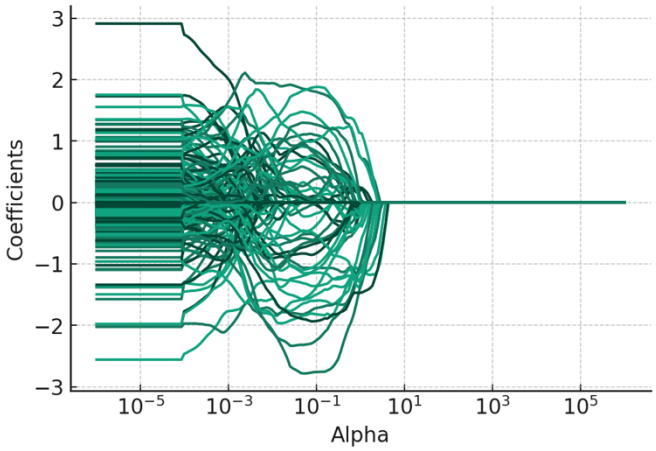
(Hoerl and Kennard; Marquardt and Snee; Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*)

2.5.1.2 *Lasso* Regresyonu

Lasso (Least Absolute Shrinkage and Selection Operator) regresyonu, *Ridge* regresyonuna benzer, ancak mutlak değer cezası kullanır. Bazı katsayıları sıfıra indirgeyebilir, etkin bir şekilde bazı öznitelikleri tamamen dışlayan daha basit bir model seçer.

$$\text{Lasso}(\beta) = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Bu formüldeki ceza $\lambda \sum_{j=1}^{p} |\beta_j|$ terimi olup, model katsayılarında seyreklik sağlar.
Resmi olarak En Az Mutlak Büzülme ve Seçim Operatörü olarak bilinen *Lasso* Regresyonu, büzülmeyi içeren doğrusal bir regresyon tekniğidir. Büzülme, veri değerlerini ortalama gibi merkezi bir noktaya doğru yönlendirir ve özellikle çok sayıda parametreye sahip modellerde etkilidir.



Şekil 2: Regularizasyonun Bir Fonksiyonu Olarak Lasso Katsayıları

Bu görselde *Lasso* Regresyonu katsayılarını gösterir. Dikkat çekici olarak, düzenlileştirme parametresi arttıkça, bazı katsayılar tam olarak sıfıra indirilir, bu da *Lasso*'nun temel özelliklerinden biri olan öznitelik seçimini gösterir.

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.

Düzenlileştirme Tekniği : *Lasso* Regresyonu, regresyon katsayılarının mutlak boyutunu cezalandıran bir düzenleme terimi sunar. Bu yaklaşım bazı özniteliklerin katsayılarını sıfıra indirerek modelden tamamen çıkarılmasına yol açabilir.

Parametre Seçimi : Cezanın gücü, katsayılara uygulanan büzülmenin boyutunu belirleyen ayarlanabilir bir parametre olan lambda (λ) tarafından kontrol edilir.

Öznitelik Seçimi : *Lasso* Regresyonunun en önemli avantajlarından biri, daha az önemli öznitelikleri ortadan kaldırarak modelleri basitleştirerek öznitelik seçimi yapabilme yeteneğidir.

Uygulamalar:

Lasso Regresyonu özellikle çok sayıda özelliğe sahip veri kümelerinde kullanışlıdır; burada öznitelik seçimi etkili tahmin modelleri oluşturmak için çok önemlidir.

Avantajları:

Ürettiği istatistiksel modelin tahmin doğruluğunu ve yorumlanabilirliğini arttırır ve modelin karmaşıklığını sınırlandırarak aşırı uyumun önlenmesinde etkili olabilir.

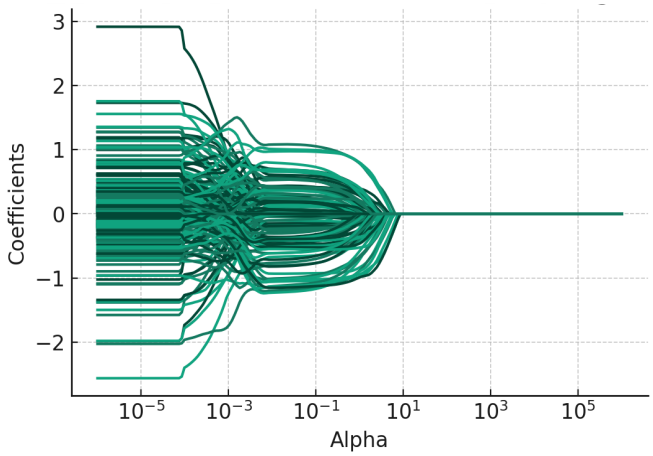
(Tibshirani; Zou and Hastie, “Regularization and Variable Selection via the Elastic Net”; Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*)

2.5.1.3 Elastik Net Regresyonu

Elastik Net, *Ridge* Regresyonu ve *Lasso* Regresyonu arasındaki bir uzlaşmadır. Özellikle birden fazla ilişkili özelliğin bulunduğu durumlarda kullanışlıdır. Elastik Net, regresyon modellerini düzenlemek için *Ridge* Regresyonu ve *Lasso* Regresyonunun cezalarını birleştirir.

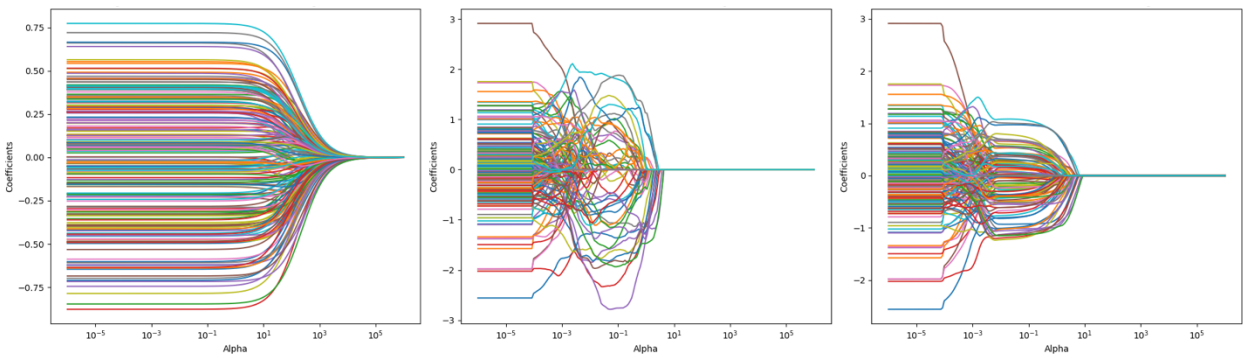
$$\text{Elastic Net}(\beta) = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

Burada, *α Ridge* (L2) ve *Lasso* (L1) cezaları arasındaki karışım parametresidir. $\alpha = 1$ olduğunda Elastik Net, *Lasso*'ya eşdeğerdir ve $\alpha = 0$ olduğunda *Ridge* Regresyonuna eşdeğerdir. Birden fazla ilişkili değişkenin olduğu senaryolarda kullanılır.



Şekil 3: Regularizasyonun Bir Fonksiyonu Olarak Elastik Net Katsayıları

Bu görselde hem *Ridge* hem de *Lasso*'nun özelliklerini birleştiren Elastik Net'i temsil eder. Katsayılar, hem sıfıra doğru küçülür (*Ridge* gibi) hem de bazıları sıfıra ayarlanır (*Lasso* gibi), bu alfa değerine ve L1 ve L2 düzenlileştirmeleri arasındaki denge parametresine bağlıdır.



Şekil 4: Regularizasyonun bir fonksiyonu olarak Ridge, Lasso Regresyonları ve Elastik Net katsayılarının yan yana karşılaştırılması

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.

Lasso ve Ridge Kombinasyonu**:** Elastik Net, modelde hem L1 (*Lasso*) hem de L2 (*Ridge)* cezalarını içerir. Bu yaklaşım, çoklu bağlantı ve öznitelik seçimini etkili bir şekilde ele alarak her iki yöntemin faydalarını devralmasına olanak tanır.

Düzenlileştirme Parametreleri : Yöntem, sırasıyla *Lasso* ve *Ridge* efektlerinin karışımını ve büzülme miktarını kontrol etmek için alfa ve lambda olmak üzere iki parametre kullanır.

Öznitelik Seçimi ve Katsayı Büzülümü : Elastik Net, *Ridge Regresyonu* gibi katsayıları küçültebilir ve *Lasso* gibi otomatik değişken seçimi gerçekleştirebilir.

Uygulamalar:

- Elastik Net, veri kümesinde yüksek düzeyde korelasyona sahip tahmin ediciler olduğunda özellikle kullanışlıdır. Model yorumlanabilirliğinin ve tahmin doğruluğunun çok önemli olduğu alanlarda yaygın olarak kullanılmaktadır.

Avantajları:

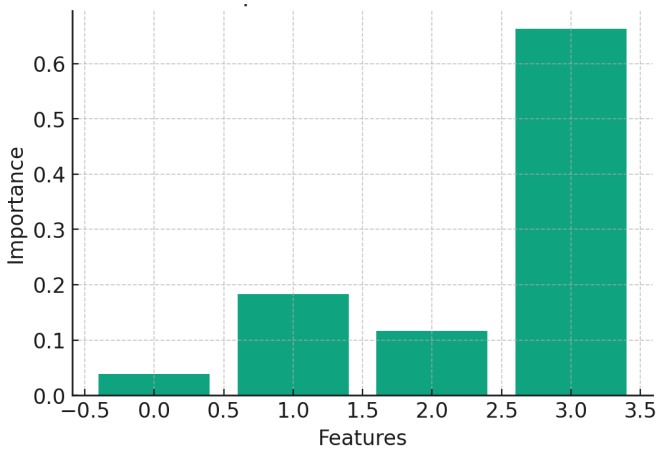
Özellikle tahmin edicilerin gözlemlerden daha fazla olduğu durumlarda veya birkaç tahmin edicinin yüksek düzeyde korelasyona sahip olduğu durumlarda, *Lasso* ve *Ridge* Regresyonunun bazı sınırlamalarına değinir.

(Zou and Zhang; Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Zou and Hastie, "Regularization and Variable Selection via the Elastic Net")

2.5.1.4 Rassal Orman Regresyonu

Rassal Orman Regresyonu, birden fazla karar ağacını birleştirerek çalışan bir topluluk öğrenme yöntemidir. Her ağaç, veri setinin rastgele alt kümelerinden eğitilir ve sonuçlar ortalama alınarak ya da çoğunluk oyuyla birleştirilerek nihai tahmin yapılır. Bu yöntem, aşırı uyum (*overfitting*) sorununu azaltmada etkilidir.

Rassal Orman Regresyonu için belirli bir matematiksel formül yoktur, çünkü bu yöntem birçok karar ağacının sonuçlarının birleştirilmesine dayanır. Ancak, her bir karar ağacı aşağıdaki gibidir: $y=f(x,\Theta k)$, burada Θk rastgele seçilen öznitelikler ve örnekler üzerine kurulu k-inci ağacın parametreleridir.



Şekil 5: Rassal Orman'da Öznitelik Önemleri

Bu görsel, Rassal Orman modelindeki özniteliklerin önem derecelerini göstermektedir. Her sütun, bir özelliğin modelin tahminlerindeki göreceli önemini temsil eder. Yüksek öneme sahip öznitelikler, modelin tahminlerini daha fazla etkiler.

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.

Karar Ağaçları Topluluğu : Rassal Orman Regresyonu, her biri verilerin rastgele bir alt kümesi üzerinde eğitilen karar ağaçlarından oluşan bir 'orman' oluşturur ve tahmin doğruluğunu geliştirmek ve aşırı uyumu kontrol etmek için ortalamayı kullanır.

Doğrusal Olmamayı Ele Alma : Değişkenler arasındaki doğrusal olmayan ilişkilerin ve karmaşık etkileşimlerin ele alınmasında özellikle etkilidir.

Sağlamlık ve Çok Yönlülük : Yöntem, gürültüye karşı dayanıklılığı ve yüksek boyutlu büyük veri kümelerini işleyebilme yeteneği ile bilinir.

Uygulamalar:

- Rassal Orman Regresyonu, hisse senedi fiyat tahmini için finans, konut fiyat tahmini için gayrimenkul ve sıcaklık ve yağış tahmini için çevre bilimi dahil olmak üzere çok çeşitli alanlarda kullanılmaktadır.

Avantajları:

- Tahminlerde yüksek doğruluk sunar ve çok sayıda özelliğe sahip büyük veri kümelerini aşırı uyum olmadan işleyebilir.
- Model aynı zamanda özelliğin önemine ilişkin bilgiler sağlayarak, sonucun tahmin edilmesinde hangi değişkenlerin en etkili olduğunu belirlemeye yardımcı olur.

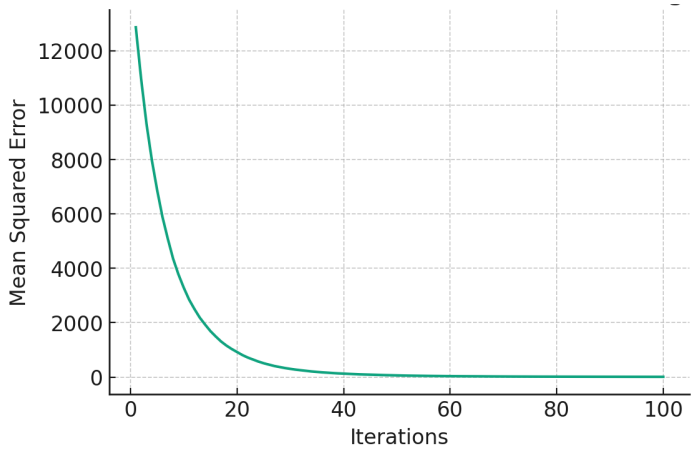
(Cutler et al.; Breiman, “Random Forests”; Liaw and Wiener, “Classification and Regression by RandomForest”)

2.5.1.5 Gradyan Artırma Regresyonu

Gradyan Artırma Regresyonu, zayıf öğrenicileri (genellikle karar ağaçları) ardışık olarak iyileştirerek güçlü bir model oluşturan başka bir topluluk öğrenme tekniğidir. Her bir ardışık öğrenici, önceki öğrenicilerin hatalarını düzeltmeye odaklanır. Güçlü bir tahmin modeli oluşturmak için zayıf tahmin modellerini birleştirir. Gradyan arttırma, hata fonksiyonunun gradyanını azaltarak modeli günceller. Temel formülü şöyledir.

$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

burada $F_m(x)$ m-inci adımdaki model, $h_m(x)$ yeni eklenen öğrenici ve γ_m bu öğrenicinin ağırlığıdır.



Şekil 6: Gradyan Artırma'da İterasyonlar boyunca OKK

Bu görsel, Gradyan Artırma modelinin iterasyonlar boyunca hata azalmasını göstermektedir. Grafik, her iterasyonda modelin performansının nasıl geliştiğini ve hata oranının nasıl azaldığını gösterir. Bu, modelin öğrenme sürecininin ve iyileşmesinin bir göstergesidir.

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.

Sıralı Ağaç Oluşturma : Ağaçları paralel olarak oluşturan Rassal Orman'ın aksine, Gradyan Artırma her seferinde bir ağaç oluşturur. Her yeni ağaç, önceden eğitilmiş ağaçların yaptığı hataların düzeltilmesine yardımcı olur.

Hata Fonksiyonu Optimizasyonu : Yöntem, modelin tahminlerinin gerçek sonuçlardan ne kadar uzakta olduğunun bir ölçüsü olan hata fonksiyonunu en aza indirmeye odaklanır

Öğrenme Hızı : Gradyan Artırma'daki önemli bir parametre, modelin ne kadar hızlı öğrendiğini kontrol eden öğrenme oranıdır. Daha küçük bir öğrenme oranı, modelin yavaş öğrendiği, daha fazla ağaç gerektirdiği ancak daha iyi performansa yol açabileceği anlamına gelir.

Uygulamalar:

- Gradyan Artırma Regresyon, risk modellemesi için finans, iklim değişikliği tahmini için çevre bilimi ve hastalık ilerleme modellemesi için sağlık hizmetleri dahil olmak üzere çeşitli alanlarda kullanılmaktadır.

Avantajları:

- Doğrusal olmayan ilişkiler de dahil olmak üzere çeşitli veri türlerinin işlenmesinde yüksek doğruluğu ve etkinliği ile bilinir.
- Model, eksik verileri işleyebilir ve ölçeklendirme veya normalleştirme gibi veri ön işlemelerini gerektirmez.

(Friedman, "Greedy Function Approximation: A Gradient Boosting Machine"; Natekin and Knoll, "Gradient Boosting Machines, a Tutorial"; Ridgeway)

2.5.2 Çok Doğrusallık Bağlantı Sorunu (Çoklu Doğrusal Bağlantı)

Regresyon analizinde çoklu doğrusallık, iki veya daha fazla bağımsız değişkenin yüksek düzeyde korelasyona sahip olması durumunda ortaya çıkar ve bunların bağımlı değişken üzerindeki bireysel etkilerini ayırt etmeyi zorlaştırır. Bu, regresyon katsayılarının güvenilmez ve istikrarsız tahminlerine yol açabilir.

Çoklu Bağlantının Belirlenmesi - Varyans Enflasyon Faktörü (VIF):

Varyans Enflasyon Faktörü (VIF), çoklu doğrusallığı tespit etmek için yaygın bir yöntemdir. Tahmin edicilerin korelasyonu durumunda, tahmin edilen bir regresyon katsayısının varyansının ne kadar arttığını ölçer.

Bağımsız bir değişkenin VIF'si, diğer tüm bağımsız değişkenlere göre regresyonuyla hesaplanır. VIF'in formülü $\mathrm{VIF}_i = \frac{1}{1 - R^2}$ şeklindedir. Burada R^2 regresyondan elde edilen belirleme katsayısıdır.

VIF değerinin 1 olması, i ninci bağımsız değişken ile geri kalan değişkenler arasında korelasyon olmadığını ve dolayısıyla çoklu doğrusallığın olmadığını gösterir. VIF arttıkça çoklu bağlantının arttığını gösterir. Yaygın bir temel kural, 10'un üzerindeki bir VIF'nin çoklu bağlantı sorununa işaret etmesidir.

Çoklu Bağlantıyı Ele Alma:

Yüksek VIF değerlerine neden olan değişkenlerin belirlenmesi ve kaldırılması, çoklu bağlantının azaltılmasına yardımcı olabilir.

İlişkili değişkenleri birleştirmek veya veri kümesinin hacmini artırmak diğer stratejilerdir. Düzenlileştirmeyi uygulayan *Ridge* ve *Lasso* Regresyonu gibi teknikler de çoklu bağlantının ele alınmasında etkili olabilir. Bu yöntemler, hatayı en aza indirmek için en küçük kareler yaklaşımını değiştirir. Çoklu doğrusallıkla uğraşırken özellikle faydalıdırlar. *Ridge*

Regresyonu, regresyon katsayılarının boyutunu cezalandırırken, *Lasso* Regresyonu bazı katsayıları sıfıra indirerek öznitelik seçimini etkili bir şekilde gerçekleştirebilir.

Makine Öğreniminde *Bias* ve Varyans:

Yüksek *Bias*, bir modelin verilerdeki temel modeli yakalayamadığını (yetersiz uyum) gösterir. Yüksek varyans, modelin çok karmaşık olduğunu ve verilerdeki gürültüyü yakaladığını (aşırı uyum) gösterir. *Bias* ve varyans arasındaki denge, model performansı için çok önemlidir.

Aşırı Uyumu Gidermek İçin Düzenleme:

Düzenlileştirme teknikleri, katsayıların büyüklüğünü cezalandırarak modelin karmaşıklığını ayarlar. Bu yaklaşım, aşırı uyumun azaltılmasına ve model genellemesinin geliştirilmesine yardımcı olur. VIF, *Ridge* ve *Lasso* Regresyonlarının kullanımı da dahil olmak üzere çoklu doğrusallığın ve çözümlerinin kapsamlı bir şekilde anlaşılması için ayrıntılı istatistiksel ve makine öğrenimi kaynaklarına başvurulması tavsiye edilir(Farrar and Glauber).

2.5.3 Sınıflandırma Modelleri

2.5.3.1 Lojistik Regresyon

Lojistik Regresyonu, ikili sınıflandırma için kullanılan istatistiksel bir modeldir. Bir veya daha fazla bağımsız değişkene dayalı olarak kategorik bir bağımlı değişkenin olasılığını tahmin eder. Sürekli bir sonucu öngören doğrusal regresyonun aksine, lojistik regresyon ikili sonuçlar (örn. evet/hayır, doğru/yanlış) için kullanılır.

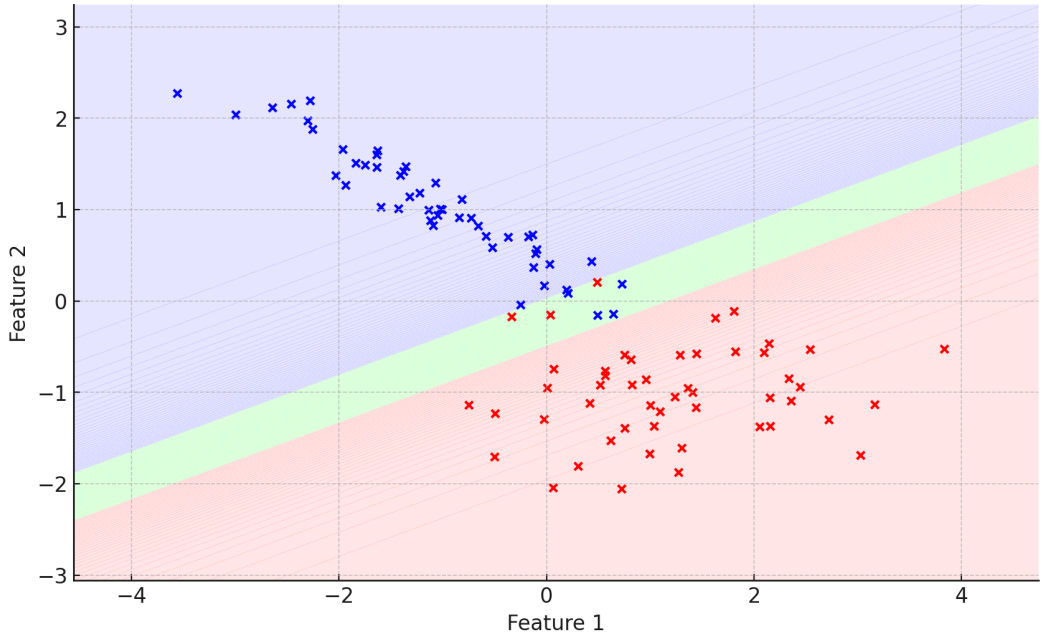
Model Mekaniği: Lojistik fonksiyon (veya sigmoid fonksiyon), lojistik regresyonun merkezinde yer alır. Gerçek değerli herhangi bir sayıyı 0 ile 1 arasında bir değere eşleyerek olasılıkları tahmin etmek için uygun hale getirir.

Lojistik Regresyonu denklemi burada

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- P olayının gerçekleşme olasılığıdır.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ modelin katsayılarıdır.
- $X_1, X_2, \ldots, X_k$ bağımsız değişkenlerdir.

Bu yöntem, verilen örnek verileri gözlemleme olasılığını en üst düzeye çıkarmaya odaklanarak parametreleri tahmin etmek için sıradan en küçük kareler yerine kullanılır.



Şekil 7: Lojistik Regresyon Sınıflandırması

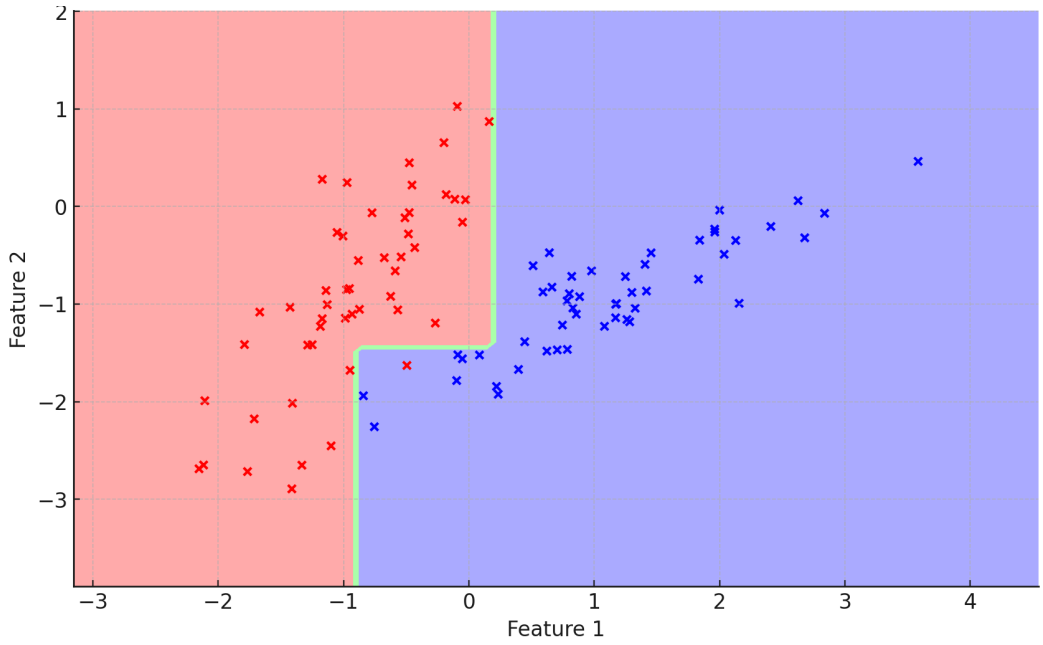
Bu grafik, iki öznitelik kullanılarak oluşturulan basit bir veri setinde Lojistik Regresyon modelinin nasıl çalıştığını göstermektedir. Arka plandaki renk değişimi (mavi ve kırmızı arasındaki geçişler), modelin sınıflandırma kararının olasılığını gösterir. Bu, modelin veri setini nasıl iki farklı sınıfa (mavi ve kırmızı noktalar) ayırdığını belirtir.

Lojistik Regresyon, bir veri noktasının her bir sınıfa ait olma olasılığını hesaplar ve bu olasılığa göre sınıflandırma yapar. Her bir nokta, bir veri örneğini temsil eder ve modelin bu örnekleri nasıl sınıflandırdığı renklerle gösterilir.

Bu görselleştirme, Lojistik Regresyon'un sınıflandırma kararlarını ve veri setindeki olasılık tabanlı sınıf ayrımını nasıl yaptığının temel bir örneğini sunar. Lojistik Regresyon, özellikle iki sınıflı (binary) sınıflandırma problemlerinde yaygın olarak kullanılır ve bu örnekte modelin karar sınırlarını ve olasılık dağılımını göstermektedir.
(Hosmer Jr et al.; Menard)

2.5.3.2 Karar Ağacı

Karar ağaçları, sınıflandırma ve regresyon için kullanılan parametrik olmayan denetimli bir öğrenme metodudur. Şans eseri sonuçları, kaynak maliyetleri ve fayda dahil olmak üzere kararları ve bunların olası sonuçlarını modellerler.



Şekil 8: Karar Ağacı Sınıflandırması

Bu grafik, iki öznitelik (Feature 1 ve Feature 2) kullanılarak oluşturulan basit bir veri setinde Karar Ağacı modelinin nasıl çalıştığını göstermektedir. Arka plandaki renkler (açık kırmızı ve açık yeşil), modelin karar sınırlarını gösterir. Bu sınırlar, modelin veri setini nasıl iki farklı sınıfa (kırmızı ve yeşil noktalar) ayırdığını belirtir.

Her bir nokta, bir veri örneğini temsil eder, ve modelin bu örnekleri nasıl sınıflandırdığı renklerle gösterilir. Model, veri noktalarının özelliklerine göre hangi sınıfa ait olduğuna karar verir ve bu, grafikte farklı renklerde gösterilen bölgelerle temsil edilir.

Bu basit görselleştirme, Karar Ağacı modelinin karmaşık veri setlerinde nasıl ayrımlar yapabileceğinin temel bir örneğini sunar. Karar ağaçları, özniteliklerin değerlerine dayalı olarak, veri noktalarını sınıflara ayıran bir dizi karar kuralları uygular.

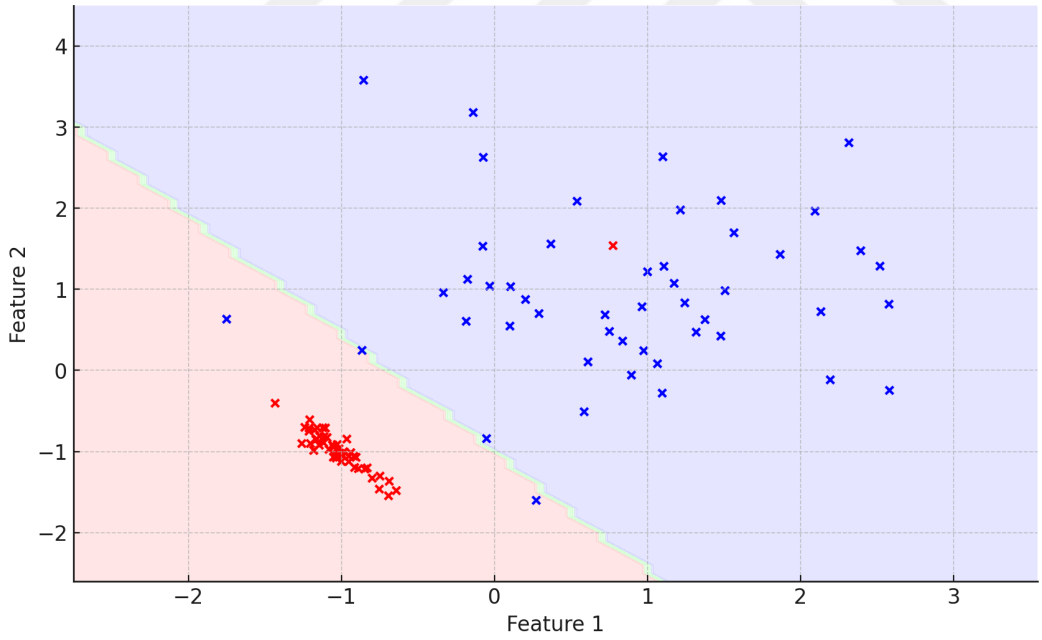
Bir karar ağacı, kök, dallar ve yaprak düğümlerini oluşturan düğümlerden oluşur. Her iç düğüm bir nitelik üzerindeki testi temsil eder, her dal testin sonucunu temsil eder ve her yaprak düğüm bir sınıf etiketini temsil eder. Bunlar bir karar ağacı oluşturmanın temel kavramlarıdır. Entropi, bir grup örnekteki safsızlığı ölçer ve bilgi kazanımı, bir özelliğin verileri sınıflara ne kadar iyi böldüğünü ölçer. Ağacın boyutunu küçültmek ve aşırı uyumun önlenmesi için kullanılan bir tekniktir. Ağacın, örnekleri sınıflandırmak için çok az güç sağlayan bölümlerini kaldırır (Quinlan; Maimon and Rokach).

2.5.3.3 Rassal Orman

Rassal Orman, eğitim zamanında çok sayıda karar ağacı oluşturarak ve sınıfların modu (sınıflandırma) veya bireysel ağaçların ortalama tahmini (regresyon) olan sınıfın çıktısını alarak çalışan bir topluluk öğrenme yöntemidir. Rassal Ormanlar, karar ağaçlarının eğitim setlerine aşırı uyum sağlama alışkanlığını düzeltir. Bunu, bazen eğitim verilerinin farklı alt kümeleriyle oluşturulan çeşitli ağaçları birleştirerek yaparlar. Rastgele bir orman modelinin tahminin yanı sıra temel çıktılarından biri, ağaç ormanında karar vermede en çok kullanılan tahmin edicilere ilişkin bilgiler sağlayan öznitelik önem puanıdır.
(Breiman, “Random Forests”; Liaw and Wiener, “Classification and Regression by RandomForest”)

2.5.3.4 Destek Vektör Makinesi (SVM)

Destek Vektör Makineleri (SVM'ler), sınıflandırma, regresyon ve aykırı değerlerin tespiti için kullanılan bir dizi denetimli öğrenme yöntemidir. Bir SVM'nin birincil amacı, N boyutlu bir uzayda (N - öznitelik sayısı) veri noktalarını belirgin bir şekilde sınıflandıran bir hiperdüzlem bulmaktır.



Şekil 9: Destek Vektör Makinesi (SVM) Sınıflandırması

Bu grafik, iki öznitelik kullanılarak oluşturulan basit bir veri setinde SVM modelinin nasıl çalıştığını göstermektedir. Arka plandaki renkler (açık kırmızı ve açık mavi), modelin karar sınırlarını gösterir. Bu sınırlar, modelin veri setini nasıl iki farklı sınıfa (kırmızı ve mavi noktalar) ayırdığını belirtir.

SVM, sınıflar arasındaki marjı en üst düzeye çıkarmak için en iyi ayırıcı hiperdüzlemi bulur. Her bir nokta, bir veri örneğini temsil eder ve modelin bu örnekleri nasıl sınıflandırdığı renklerle gösterilir. Model, veri noktalarının özelliklerine göre hangi sınıfa ait olduğuna karar verir ve bu, grafikte farklı renklerde gösterilen bölgelerle temsil edilir.

Bu görselleştirme, SVM'nin veri setlerindeki sınıflandırma kararlarını ve sınıf ayrımını nasıl yaptığının temel bir örneğini sunar. SVM, sınıfları ayırmak için lineer veya non-lineer karar sınırları oluşturabilir ve bu örnekte lineer bir sınır kullanılmıştır.

SVM'nin Temel Kavramları:

Hiperdüzlem : SVM'lerde hiperdüzlem, öznitelik uzayındaki farklı sınıfları ayıran bir karar sınırıdır. En iyi hiperdüzlem, iki sınıf arasında en büyük marja sahip olandır.

Destek Vektörleri : Bunlar hiper düzleme en yakın veri noktalarıdır ve marjı tanımladıkları için eğitim setinin kritik unsurlarıdır.

Marj : En yakın sınıf noktalarındaki iki çizgi arasındaki boşluktur. Sınıflandırıcının daha iyi genelleştirilmesi için genellikle daha büyük bir kenar boşluğu tercih edilir.

Çekirdek Hilesi : SVM'ler, verileri daha yüksek boyutlu bir alana dönüştürmek için çekirdek hilesini kullanabilir, böylece ayırıcı bir hiperdüzlem bulmayı kolaylaştırır. Yaygın çekirdekler arasında doğrusal, polinom, radyal temel fonksiyon (RBF) ve sigmoid bulunur.

Denklem ve Parametreler :

Hiperdüzlemi tanımlayan w ağırlık vektörünü ve *Bias* b'yi bulmaya çalışır.; $f(x) = \mathbf{w}^T\mathbf{x} + b$ $f(x)$ karar fonksiyonudur, w ağırlık vektörüdür, x giriş vektörüdür ve b *Bias* terimidir. SVM'deki C parametresi, düşük eğitim hatası ile düşük test hatası (genelleme) arasındaki dengeyi kontrol eden bir düzenleme parametresidir.

Avantajları :

- DVM'ler yüksek boyutlu uzaylarda etkilidir ve öznitelikle yüksek boyutlu uzayda aşırı uyum sağlamaya karşı nispeten dayanıklıdır.
- Açık bir ayrım marjıyla iyi çalışırlar ve uygulanabilecek farklı çekirdek fonksiyonları nedeniyle çok yönlüdürler.

Uygulamalar :

- SVM'ler, görüntü sınıflandırma, metin kategorizasyonu, biyoinformatik (örneğin, protein sınıflandırması için) ve sınıflar arasındaki ayrımın çok önemli olduğu diğer birçok uygulamada yaygın olarak kullanılmaktadır

(Cortes and Vapnik; Schölkopf and Smola; Burges).

2.5.3.5 Gradyan Artırma

Gradyan Artırma, hem regresyon hem de sınıflandırma görevlerinde kullanılan çok yönlü bir makine öğrenme tekniğidir. Sınıflandırmada, her bir ağacın bir öncekinin hatalarını düzeltmeye çalıştığı bir dizi karar ağacını sırayla oluşturur.
Burada kullanılan anahtar kavramlar aşağıdaki gibidir.
Toplu Öğrenme: *Gradyan Artırma*, güçlü bir sınıflandırıcı oluşturmak için birden fazla zayıf öğreniciyi (tipik olarak karar ağaçları) birleştirir. Her ağaç, önceki ağaçların artıkları (hataları) üzerine eğitilir.

Hata Fonksiyonu Optimizasyonu: Algoritma, tahmin edilen ve gerçek sınıf etiketleri arasındaki farkı ölçen bir hata fonksiyonunu en aza indirmeye odaklanır.

Öğrenme Hızı: *Gradyan Artırma*'deki önemli bir parametre, her ağacın nihai modele katkısını etkileyen öğrenme oranıdır. Daha küçük bir öğrenme oranı daha fazla ağaç gerektirir ancak modelin doğruluğunu artırabilir.

Uygulamalar:

- *Gradyan Artırma* sınıflandırıcıları, dolandırıcılık tespiti, müşteri segmentasyonu ve tıbbi teşhis gibi doğru sınıflandırmanın gerekli olduğu çeşitli uygulamalarda yaygın olarak kullanılmaktadır.

Avantajları:

- Yüksek doğruluğu ve etkinliği ile bilinen *Gradyan Artırma*, dengesiz veri kümeleri de dahil olmak üzere farklı veri türlerini işleyebilir.
- Özellikle çok sayıda özelliğin ve eğitim örneklerinin olduğu senaryolarda aşırı uyum sağlamaya karşı dayanıklıdır

(Chen and Guestrin; Natekin and Knoll, “Gradient Boosting Machines, a Tutorial”; Friedman, “Stochastic Gradient Boosting”).

2.5.3.6 Sinir Ağı

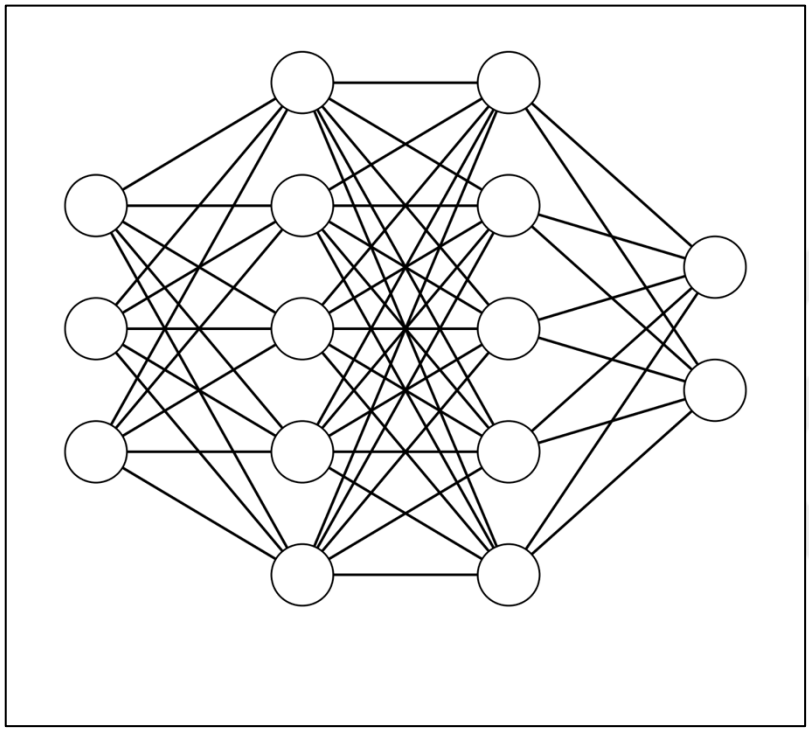
Sinir Ağları, bilgi işleme için beyin yapısından ilham alan yapay zeka ve makine öğrenmesinde kullanılan modellerdir. Temel olarak, birçok basit, birbiriyle bağlantılı işlem birimlerinden (nöronlardan) oluşurlar. Bu modeller, veri üzerinde karmaşık desenleri tanıyabilir ve sınıflandırma, regresyon, resim ve ses tanıma gibi görevlerde kullanılırlar.
Bir yapay sinir ağı, genellikle bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanından oluşur. Her nöron, ağırlıklı girdileri toplayıp bir aktivasyon fonksiyonu uygular.

Matematiksel olarak bir nöron şu şekilde ifade edilir:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

Burada x_i girdiler, w_i ağırlıklar, b *bias* terimi ve f aktivasyon fonksiyonudur.

Aşağıdaki görsel, bir sinir ağının temel yapısını göstermektedir ve gerçek uygulamalarda kullanılan daha karmaşık ağ yapılarından daha basittir.



Şekil 10: Basit Yapay Sinir Ağı Modeli

Bu görsel, tipik bir yapay sinir ağının temel yapısını göstermektedir. Model, üç girdi nöronundan oluşan bir girdi katmanı, iki gizli katman (her biri beş nöron içerir) ve iki nöronlu bir çıktı katmanından oluşmaktadır. Her nöron, önceki katmandaki tüm nöronlarla bağlantılıdır ve bu bağlantılar ağın ağırlıklarını temsil eder. Gizli katmanlardaki her nöron, girdi sinyallerini toplar, bir aktivasyon fonksiyonu uygular ve sonucu bir sonraki katmana iletir. Bu basit görsel, yapay sinir ağlarının temel çalışma prensibini yansıtmaktadır. Gerçek dünya uygulamalarında kullanılan sinir ağları genellikle daha fazla katman ve nöron içerir ve çok daha karmaşık yapılar oluşturabilirler.

Burada kullanılan anahtar kavramlar aşağıdaki gibidir.

Katmanlı Yapı : Tipik bir sinir ağı, bir giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanından oluşur. Her katman, hesaplamaları gerçekleştiren ve bilgiyi bir sonraki katmana aktaran nöronları içerir.

Aktivasyon Fonksiyonları : Her nörona uygulanan bu fonksiyonlar, nöronun aktive edilip edilmeyeceğini belirler. Sınıflandırmaya yönelik yaygın aktivasyon fonksiyonları arasında sigmoid, softmax ve ReLU (Düzeltilmiş Doğrusal Birim) yer alır.

Geriye Yayılım ve Eğitim : Sinir ağları, modelin ağırlıklarını tahminlerinin hatasına göre ayarladığı, geri yayılım adı verilen bir süreç aracılığıyla öğrenir. Gradyan iniş gibi optimizasyon algoritmalarıyla birleştirilen bu süreç, ağın zaman içinde sınıflandırma doğruluğunu geliştirmesini sağlar.

Uygulamalar:

- Sinir ağları, görüntü ve konuşma tanıma, tıbbi teşhis ve duygu analizi gibi çok çeşitli sınıflandırma görevlerinde kullanılır.

Avantajları:

- Verilerdeki karmaşık kalıpları ve ilişkileri öğrenme yeteneğine sahiptirler, bu da onları sınıflandırma görevlerinde oldukça etkili kılar.
- Sinir ağları, çeşitli türdeki giriş verilerine uyum sağlayarak büyük ve yüksek boyutlu veri kümelerini işleyebilir.

Zorluklar ve Dikkat Edilmesi Gerekenler :

- Veri Karmaşıklığı : Kümeleme, özellikle büyük veri kümeleri ve yüksek boyutlu uzay söz konusu olduğunda hesaplama açısından karmaşık olabilir.
- Doğruluk ve Yorumlanabilirlik : Kümelemenin başarısı genellikle özneldir ve yöntemin anlamlı kalıpları ortaya çıkarma yeteneğine bağlıdır.

Denetimli ve Denetimsiz Öğrenmenin Karşılaştırılması :

- Denetimli Öğrenme : Etiketli veri kümeleriyle öğrenmeyi içerir ve genellikle daha basit ve doğrudur.
- Denetimsiz Öğrenme : Etiketlenmemiş verilerle ilgilenir ve daha çok keşifle ilgilidir, genellikle daha az kesin ancak daha anlayışlı bulgulara yol açar .

(Schmidhuber, “Deep Learning in Neural Networks: An Overview”; LeCun et al.; Heaton, “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning: The MIT Press, 2016, 800 Pp, ISBN: 0262035618”)

2.5.4 Derin Öğrenme

2.5.4.1 Yapay Sinir Ağları (YSA)

Yapay Sinir Ağları (YSA), hayvan beyinlerinin biyolojik sinir ağlarından esinlenerek derin öğrenmenin temel taşıdır. Ağ üzerinden bilgi besleyerek girdilerden gelen değerleri hesaplayabilen, birbirine bağlı nöron sistemleridir.

YSA'ların Derinlemesine Açıklaması:

Biyolojik İlham : YSA'lar , özellikle nöronlara, dendritlere ve aksonlara odaklanarak insan beyninin yapısına göre modellenmiştir . Nöronlar, biyolojik nöronların işleyişine benzer şekilde sinyalleri (girdileri) alır, bunları işler ve çıktıları diğer nöronlara iletir.

Ağ mimarisi :

- Giriş Katmanı : Ağın giriş sinyalini alan ilk katmanı.
- Gizli Katmanlar : Hesaplamaların yapıldığı bir veya daha fazla katman. Giriş verilerinden öznitelikler çıkarır ve gösterimleri öğrenirler.
- Çıkış Katmanı : Ağın son çıktısını üretir.

Nöron İşlevselliği : Bir YSA'daki her nöron girdiyi alır ve ona bir ağırlık ve *Bias* uygular. Bu ağırlıklı girdilerin toplamı daha sonra nöronun çıktısını belirleyen bir aktivasyon fonksiyonundan geçirilir.

Öğrenme Süreci : YSA'lar, ağırlıkları ve *Bias*ları ayarlayarak, genellikle geri yayılım adı verilen bir yöntemin gradyan iniş gibi bir optimizasyon tekniğiyle birleştirilmesiyle öğrenir.

Tek Katmanlı ve Çok Katmanlı Ağlar :

- Tek Katmanlı Ağlar : Herhangi bir gizli katman olmaksızın, bir giriş ve bir çıkış katmanından oluşur. Yetenekleri sınırlıdır ve yalnızca doğrusal ilişkileri modelleyebilirler.
- Çok Katmanlı Ağlar : Derin sinir ağları olarak da bilinen bu ağlar, bir veya daha fazla gizli katman içerir ve karmaşık, doğrusal olmayan ilişkileri modellemelerine olanak tanır.

YSA'lar, görüntü ve konuşma tanıma, doğal dil işleme ve tahmine dayalı modelleme dahil olmak üzere çeşitli uygulamalarda kullanılır.

Eksik bilgileri işleyebilirler ve yeni verilere uyum sağlayabilirler, bu da onları çeşitli görevler için çok yönlü hale getirir (Heaton, “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning: The MIT Press, 2016, 800 Pp, ISBN: 0262035618”; Haykin; Schmidhuber, “Deep Learning in Neural Networks: An Overview”).

3 LİTERATÜR İNCELEMESİ

Makine öğreniminin eğitimsel veri analizinde uygulanması, öğrenci başarısını tahmin etmek ve sınıflandırmak için umut verici yollar sunar. Bu tez, 'G_avg' değişkenine odaklanarak hem regresyon hem de sınıflandırma görevleri için bir dizi makine öğrenme algoritması kullanır.
Davranış Analizinde Etkililik: Lorenzen ve arkadaşları, çevrimiçi eğitim sistemlerindeki davranış kalıplarını izlemede makine öğreniminin etkinliğini gösterdi. Sıradan En Küçük Kareler ve *Ridge* Regresyonu gibi regresyon tekniklerine benzer yaklaşımları, öğrenci etkinliklerindeki kalıpları etkili bir şekilde belirleyerek akademik başarıyı etkileyen faktörlerin daha derin anlaşılmasına katkıda bulundu (Lorenzen et al.).

Doğru Performans Tahmini: Tadayon ve Pottie, eğitsel oyunlarda öğrenci performansını tahmin etmek için gizli Markov modelini başarıyla kullanmıştır. Bu, araştırmanızda Karar Ağaçları ve Rassal Orman gibi sınıflandırma algoritmalarının kullanımına uygundur ve akademik sonuçları tahmin etmede makine öğrenimi modellerinin doğruluğunu vurgular (Tadayon and Pottie).

Kapsamlı Veri Analizi: Elrahman ve arkadaşları, e-Ders Kitabı etkileşimlerini kullanarak öğrenci performansını tahmin etmek için bir model önerdi. Bulguları, birden fazla makine öğrenimi yöntemini değerlendirdikleri için hem regresyon hem de sınıflandırma algoritmalarının kullanımını desteklemektedir ve bu modellerin eğitimsel veri analizinde kapsamlı doğasını ortaya koymaktadır (Elrahman et al.).

Model Karmaşıklığı ve Yorumlanabilirliği: Mandalapu ve arkadaşları, dikkatle tasarlanmış özniteliklere sahip lojistik regresyon modellerinin derin modellerden daha iyi performans gösterdiğini buldu. Bu, daha basit modellerin bazen karmaşık olanlardan daha etkili olabileceğini düşündürmektedir ; bu, Elastic-Net veya Destek Vektör Makineleri gibi gelişmiş algoritmaları kullanırken araştırmanız için dikkate alınması gereken bir husustur (Mandalapu et al.).

Modellerin Genelleştirilmesi: Orji ve Vassileva motivasyona dayalı akademik performansı tahmin etmeye yönelik modeller geliştirmiştir. Araştırmaları yüksek doğruluk gösterse de, aynı zamanda bu tür modellerin farklı eğitim bağlamlarında genelleştirilebilirliğine ilişkin soruları da gündeme getiriyor; bu, çalışmanız için potansiyel bir zorluktur (Orji and Vassileva).
Literatür, eğitimsel veri analizinde makine öğrenimi algoritmalarının kullanılmasının hem potansiyelini hem de zorluklarını göstermektedir. Bu modeller kalıpları etkili bir şekilde ortaya

çıkarabilir ve akademik performansı tahmin edebilirken, modelin karmaşıklığı, yorumlanabilirliği ve genellenebilirliğine ilişkin hususlar dikkate alınmalıdır.

4 ANALİZ SÜRECİNE GENEL BAKIŞ VE VERİ KÜMESİ ÖZETİ

Öğrencilere ait öznitelikleri ve matematik performans verilerini kapsayan veri kümesinin ayrıntıları **features_explanations.csv** dosyasında ve aşağıdaki tabloda verilmiştir.

Tablo 1: Veri Setindeki Öznitelikler ve Açıklamaları

No.	Öznitelik	Açıklaması
1	school	Öğrencinin okulu (binary: 'GP' - Gabriel Pereira veya 'MS' - Mousinho da Silveira)
2	sex	Öğrencinin cinsiyeti (binary: 'F' - female or 'M' - male)
3	age	Öğrencinin yaşı (numeric: from 15 to 22)
4	address_type	Öğrencinin ev adresi (binary: 'Urban' or 'Rural')
5	family_size	Ailenin büyüklüğü (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6	parent_status	Ebeveynin beraber yaşama durumu (binary: 'T' - together or 'A' - apart)
7	mother_education	Annenin eğitim durumu (ordinal: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8	father_education	Babanın eğitim durumu (ordinal: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9	mother_job	Annenin işi (nominal: 'teacher', 'health care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	father_job	Babanın işi (nominal: 'teacher', 'health care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	reason	Bu okulu seçme sebebi (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	guardian	Öğrencinin velisi (nominal: 'mother', 'father' or 'other')
13	travel_time	Evden okula gitme süresi (ordinal: '<15 min.', '15 to 30 min.', '30 min. to 1 hour', or '>1 hour')
14	study_time	Haftalık ders çalışma saati (ordinal: '<2 hours', '2 to 5 hours', '5 to 10 hours', or '>10 hours')
15	class_failures	Geçmişteki sınıf başarısızlık sayısı (numeric: n if 1<=n<3, else 4)
16	school_support	Okulda ek eğitim desteği (binary: yes or no)
17	family_support	Aile eğitim desteği (binary: yes or no)
18	extra_paid_classes	Matematik konusunda ek ücretli dersler (binary: yes or no)
19	activities	Müfredat dışı aktiviteler (binary: yes or no)
20	nursery	Kreşe katılım (binary: yes or no)
21	higher_ed	Yüksek öğrenim yapma arzusu (binary: yes or no)
22	internet	Evde İnternet erişimi (binary: yes or no)
23	romantic_relationship	Romantik bir ilişkisi olma (binary: yes or no)
24	family_relationship	Aile ilişkilerinin kalitesi (numeric: from 1 - very bad to 5 - excellent)
25	free_time	Okul sonrası boş zaman (numeric: from 1 - very low to 5 - very high)

26	social	Arkadaşlarla dışarı sıkma sıklığı (numeric: from 1 - very low to 5 - very high)
27	weekday_alcohol	Hafta içi alkol tüketimi (numeric: from 1 - very low to 5 - very high)
28	weekend_alcohol	Hafta sonu alkol tüketimi (numeric: from 1 - very low to 5 - very high)
29	health	Güncel sağlık durumu (numeric: from 1 - very bad to 5 - very good)
30	absences	Okul devamsızlık sayısı (numeric: from 0 to 93)
31	grade_1	İlk dönem notu (numeric: from 0 to 20)
32	grade_2	İkinci dönem notu (numeric: from 0 to 20)
33	grade_3	Son sınav notu (numeric: from 0 to 20)

Bu veri seti, demografik ayrıntılar, aile geçmişleri ve her biri toplam 20 puan üzerinden olmak üzere üç matematik sınavından (G1, G2, G3) alınan puanlar gibi bir dizi özelliği içerir. Bu değerlendirmeler matematik müfredatının çeşitli unsurlarını kapsar.

Bu çalışmanın temel amacı öğrencilerin matematikteki başarılarını etkileyen unsurları belirlemek, anlamak ve bu faktörleri modellemek için farklı regresyon ve sınıflandırma yöntemlerini kullanmaktır. Bu bölümde öncelikle regresyon yöntemleri ele alınacaktır.

4.1 Regresyon Analizi

Veri Hazırlığı, aşağıdakileri içeren çeşitli adımları kapsamaktadır:

1. Veri Yükleme: Veri kümesi Python tabanlı bir analitik ortama aktarılmıştır ve bu durumda Anaconda'daki Jupyter not defteri 7.0.6 kullanılmıştır.
2. Gelişen Öznitelikler: Yeni bir bileşik puan olan G_avg, üç sınav puanının (G1, G2, G3) ortalaması olarak hesaplandı ve öğrencinin genel performansına ilişkin bütünsel bir görünüm sundu . Ancak G1, G2 ve G3'ün de analize bireysel etkisinin göz ardı edilemeyeceği göz önünde bulundurularak veri setinde tutulmasının gerekli olduğuna karar verilmiştir.

```
# Importing the pandas library
import pandas as pd

# Loading the data from 'student-mat.csv' with semicolon as the delimiter
data = pd.read_csv('student-mat.csv', delimiter=';')

# Calculating the average of G1, G2, and G3 and creating a new column 'G_avg'
data['G_avg'] = data[['G1', 'G2', 'G3']].mean(axis=1)

# Displaying the first few lines of the dataset to verify the new column
data.head()
```

Şekil 11: G1,G2 ve G3'ün aritmetik ortalaması olan G_avg isminde yeni bir sütun oluşturuluyor

3. Verilerin Temizlenmesi: Doğru veri türlerinin sağlanması ve eksik veya yanlış verilerin giderilmesi için veri kümesinin hazırlanması ve temizlenmesi yapıldı. Bu hazırlık aynı zamanda kategorik değişkenlerin analizde kullanılabilecek sayısal değişkenlerle

değiştirilmesi için One-Hot Encoder uygulanmasını da içerir. Bunu, sürekli değişkenlere standart ölçeklendirmenin uygulanması takip eder.

```python
# Importing the pandas library
import pandas as pd

# Loading the data from 'student-mat.csv' with semicolon as the delimiter
data = pd.read_csv('student-mat.csv', delimiter=';')

# Calculating the average of G1, G2, and G3 and creating a new column 'G_avg'
data['G_avg'] = data[['G1', 'G2', 'G3']].mean(axis=1)

# Identifying categorical columns (assuming non-numeric columns are categorical)
categorical_cols = data.select_dtypes(include=['object']).columns

# Applying one-hot encoding to categorical columns
one_hot_encoded_data = pd.get_dummies(data, columns=categorical_cols)

# Displaying the first few lines of the new dataframe
one_hot_encoded_data.head()
```

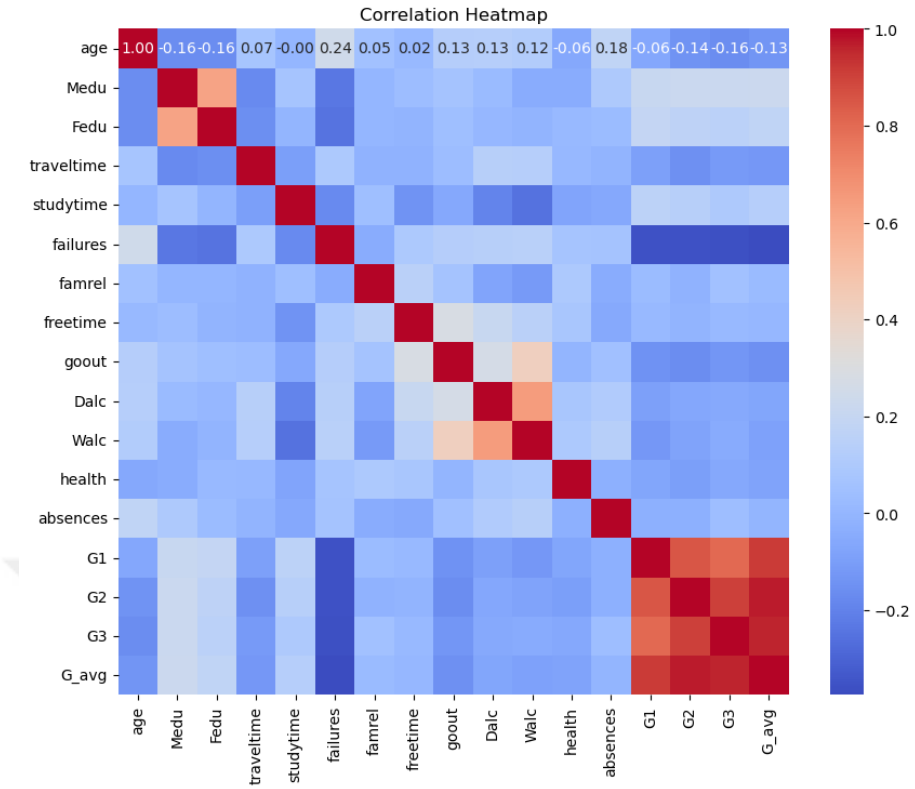
Şekil 12: OneHot Encoder uygulanıyor

```python
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import StandardScaler

# Loading the data from 'student-mat.csv' with semicolon as the delimiter
data = pd.read_csv('student-mat.csv', delimiter=';')

# Calculating the average of G1, G2, and G3 and creating a new column 'G_avg'
data['G_avg'] = data[['G1', 'G2', 'G3']].mean(axis=1)

# Identifying continuous (numeric) columns
continuous_cols = data.select_dtypes(include=['int64', 'float64']).columns

# Applying standard scaling to continuous variables
scaler = StandardScaler()
data[continuous_cols] = scaler.fit_transform(data[continuous_cols])

# Displaying the first few lines to verify the scaling
data.head()
```

s	Medu	Fedu	Mjob	Fjob	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	G_avg
A	1.143856	1.360371	at_home	teacher	...	-0.236010	0.801479	-0.540699	-1.003789	-0.399289	0.036424	-1.782467	-1.254791	-0.964934	-1.357670
T	-1.600009	-1.399970	at_home	other	...	-0.236010	-0.097908	-0.540699	-1.003789	-0.399289	-0.213796	-1.782467	-1.520979	-0.964934	-1.447953
T	-1.600009	-1.399970	at_home	other	...	-0.236010	-0.997295	0.583385	0.551100	-0.399289	0.536865	-1.179147	-0.722415	-0.090739	-0.635408

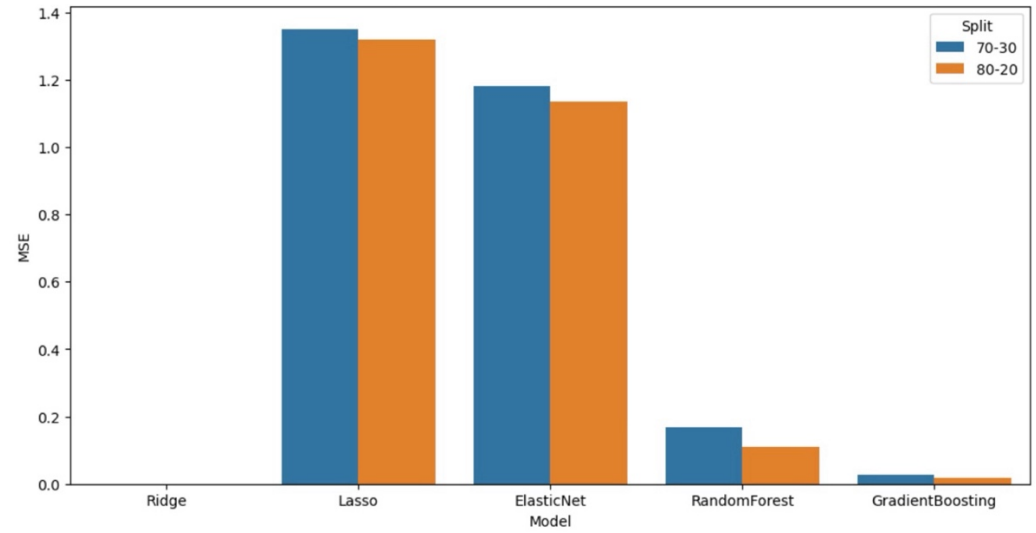
Şekil 13: Standart Ölçeklendirme uygulanıyor

4. Veri Araştırması : Çeşitli niteliklerin dağılımını ve bunların öğrenci performansıyla olası bağlantılarını anlamak için veri setinin ilk incelemesi yapıldı . Korelasyon Isı Haritasının kullanılması, öznitelikleri ve bunların hedef değişkenle (G_avg) korelasyonunu görsel-leştirmek için kullanışlıdır.

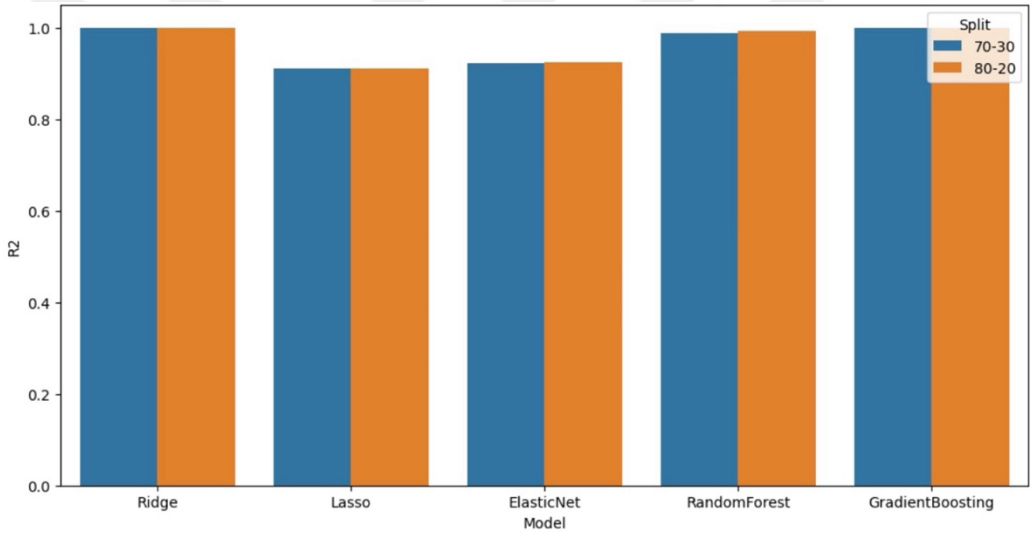


Şekil 14: Korelasyon Isı Haritası

5. Model Seçimi: Analiz için *Ridge* Regresyonu, *Lasso* Regresyonu, Elastik Net, Rassal Orman ve Gradyan Artırma gibi farklı regresyon modelleri kullanıldı.
6. Modellerin değişen koşullar altında etkinliğini test etmek için veriler iki farklı oranda (70-30 ve 80-20) eğitim ve test gruplarına bölündü .Bu boyutta 80-20 daha yaygın olarak kullanılırken test veri boyutunun arttırılması ve bu iki seçenek arasında karşılaştırma yapılması önemlidir.
7. Modellerin Eğitimi ve Değerlendirilmesi: Her model, eğitim veri seti ile eğitime tabi tutuldu ve ardından test veri seti kullanılarak değerlendirildi. Doğruluk değerlendirmesi için Ortalama Kare Hata (MSE) ve R-kare (R2) gibi ölçümler kullanıldı.



Şekil 15: Model Karşılaştırması: Farklı Bölünmeler için Ortalama Kare Hata

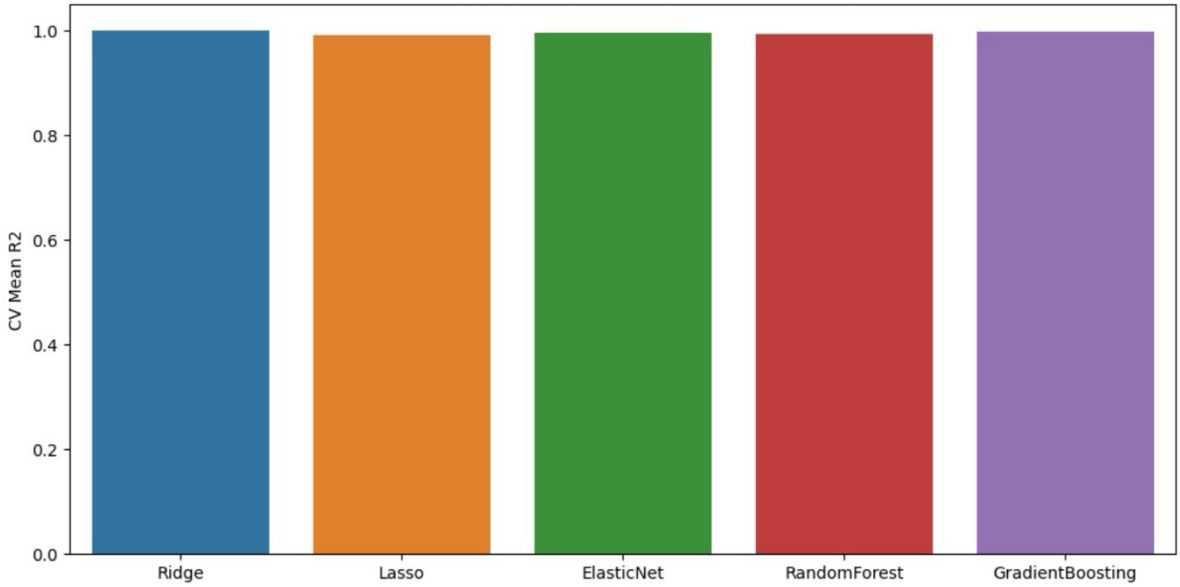


Şekil 16: Model Karşılaştırması: Farklı Bölünmeler için R2 Skoru

8. Çapraz Doğrulama Yoluyla Doğrulama: Modellerin etkinliğinin kapsamlı bir değerlendirmesini sağlamak için çapraz doğrulama yapıldı.

9. Model Etkinliği: *Ridge* Regresyon modeli olağanüstü yüksek performans göstergeleri gösterdi; bu, verilere güçlü bir uyum olduğunu gösteriyor ancak aynı zamanda aşırı

uyum konusunda endişeleri de artırıyor. Gradyan Artırma modeli ise karmaşık veri ilişkilerini anlama kapasitesini yansıtan etkileyici sonuçlar da gösterdi.



Şekil 17: Model Karşılaştırması: Çapraz Doğrulama R2 Skorları

10. Bölünmüş Karşılaştırma: Genellikle, 80-20'lik veri bölünmesi, 70-30'luk bölünmeyle karşılaştırıldığında, muhtemelen eğitim için mevcut verilerin artması nedeniyle, model performansının marjinal olarak iyileşmesine yol açtı.

Tablo 2: Regresyon Analizi Sonuçlarının 70-30 ve 80-20 performans karşılaştırması

Split Results (70-30 vs. 80-20):

Model	Split	MSE (70-30)	R2 (70-30)	MSE (80-20)	R2 (80-20)
Ridge	70-30	4.23e-05	0.999997	3.34e-05	0.999998
Lasso	70-30	1.3508	0.9101	1.3183	0.9110
ElasticNet	70-30	1.1804	0.9214	1.1347	0.9234
RandomForest	70-30	0.1685	0.9888	0.1107	0.9925
GradientBoosting	70-30	0.0265	0.9982	0.0189	0.9987

11. Doğrulama Sonuçları: Çapraz doğrulamadan elde edilen sonuçlar, *Ridge* ve *Gradyan Artırma* modellerinin en iyi performansı gösterdiği test seti sonuçlarıyla uyumludur.

Tablo 3: Çapraz Doğrulama Sonuçları

Cross-Validation Results:

Model	CV Mean R2	CV Std R2
Ridge	0.99999999	1.99e-09
Lasso	0.9922	0.0015
ElasticNet	0.9969	0.0005
RandomForest	0.9938	0.0020
GradientBoosting	0.9977	0.0017

12. Temel Faktörlerin Belirlenmesi: Bu çalışma, öğrenciler arasında matematik performansını etkileyen temel faktörlere ışık tutmakta, eğitimcilere ve politika geliştiricilere müfredatın ve öğrenci destek mekanizmalarının şekillendirilmesinde değerli bilgiler sunmaktadır.
13. Özel Öğrenme Yaklaşımları: Geliştirilen modeller, bireysel öğrenci performanslarını tahmin etmek, daha özelleştirilmiş eğitim yöntemleri sağlamak ve düşük performans riski taşıyan öğrencilere erken destek sağlamak için kullanılabilir.
14. Müfredat Geliştirme: Özellikle özniteliklerin önemiyle ilgili olarak elde edilen bilgiler, müfredatın geliştirilmesini öğrencinin öğrenimi üzerinde en etkili alanlara odaklanacak şekilde yönlendirebilir.
15. Bilgilendirilmiş Politika Geliştirme: Bu sonuçlar, eğitim kurumlarının ve politika yapıcıların matematik eğitiminin kalitesini ve öğrenci başarılarını iyileştirmek için bilinçli, veriye dayalı kararlar almalarına yardımcı olabilir. Matematik eğitimi veri setinin bu regresyon analizi, öğrenci performansını etkileyen kritik faktörleri ortaya koymaktadır. Yerleşik modeller, öğrenci sonuçlarını tahmin etmek için sağlam bir temel sağlar ve eğitimsel taktiklerin ve politika oluşturmanın geliştirilmesinde etkili olabilir. Gelişmiş analitik yöntemlerin uygulanması, matematik eğitimi sonuçlarının arttırılmasında veri merkezli yaklaşımların değerinin altını çizmektedir.

4.2 Sınıflandırma Analizi

Tezin bu bölümünde sınıflandırma algoritmaları kullanılarak matematikteki öğrenci performansının ayrıntılı bir analizi sunulmaktadır. Amaç, öğrencileri çeşitli özniteliklere göre farklı başarı düzeylerine göre kategorize etmek ve eğitim stratejilerinde faydalı olabilecek bilgiler sağlamaktı.

Veri Hazırlama ve Ön İşleme

1. **Veri Yükleme : student-mat.csv** veri kümesi Python'un pandas kütüphanesi kullanılarak yüklendi. Bu veri seti öğrencilerin akademik ve kişisel yaşamlarının yanı sıra matematik notlarına ilişkin çeşitli öznitelikleri içermektedir.
2. **Öznitelik Mühendisliği** :

   (G1, G2, G3) aritmetik ortalaması olarak yeni bir hedef değişken olan **G_avg** oluşturuldu. Veri seti sınıflandırmayı kolaylaştırmak için ön işleme tabi tutuldu. Bu, verileri normalleştirmek için kategorik değişkenlerin tek seferde kodlanmasını ve sürekli değişkenlerin standart ölçeklendirilmesini de içermektedir.

3. **Hedef Değişken Dönüşümü** :

   Sürekli **G_avg değişkeni,** akademik başarının üç düzeyini temsil eden **Başarı_Seviyesi** adlı kategorik bir değişkene dönüştürüldü : Yetersiz (%0-49), Tatmin Edici (%50-74) ve Başarılı (%75-100).

```
# Step 4: Create success classification levels for G_avg
# Defining the classification levels
conditions = [
    (data['G_avg'] < 0.5), # Unsatisfactory: 0-49%
    (data['G_avg'] >= 0.5) & (data['G_avg'] < 0.75), # Satisfactory: 50-74%
    (data['G_avg'] >= 0.75) # Successful: 75-100%
```

Şekil 18: G_avg için başarı sınıflandırma seviyeleri oluşturulması

4. **Model Eğitimi ve Değerlendirme**

   Model Seçimi : Lojistik Regresyon, Karar Ağaçları, Rassal Orman, Destek Vektör Makineleri (SVM), Gradyan Arttırma ve Sinir Ağları dahil olmak üzere çeşitli sınıflandırma modelleri seçildi.

```
# Define the models
models = {
    'LogisticRegression': LogisticRegression(),
    'DecisionTree': DecisionTreeClassifier(),
    'RandomForest': RandomForestClassifier(),
    'SVM': SVC(),
    'GradientBoosting': GradientBoostingClassifier(),
    'NeuralNetwork': MLPClassifier()
}
```

Şekil 19: Sınıflandırma modellerinin tanımlanması

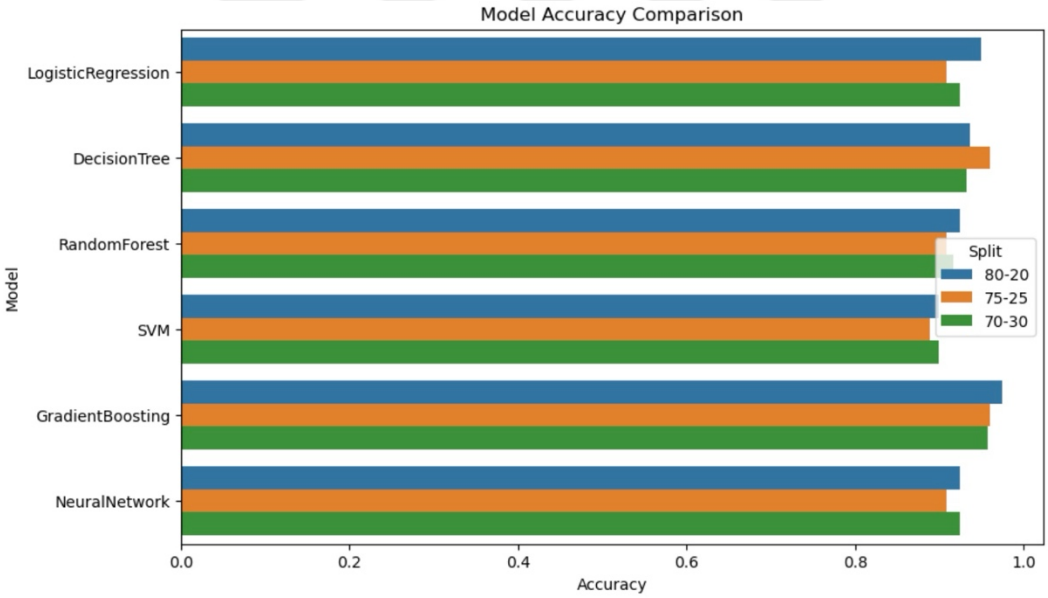
Eğitim-Test Bölünmesi : Veri seti, modelleri değişen koşullar altında değerlendirmek için üç farklı oran (80-20, 75-25, 70-30) kullanılarak eğitim ve test setlerine bölündü.

```
# Define the models
models = {
    'LogisticRegression': LogisticRegression(),
    'DecisionTree': DecisionTreeClassifier(),
    'RandomForest': RandomForestClassifier(),
    'SVM': SVC(),
    'GradientBoosting': GradientBoostingClassifier(),
    'NeuralNetwork': MLPClassifier()
}

# Train-test splits
splits = [0.2, 0.25, 0.3]
```

Şekil 20: 3 Farklı Oranla Eğitim-Test Verileri oluşturuluyor

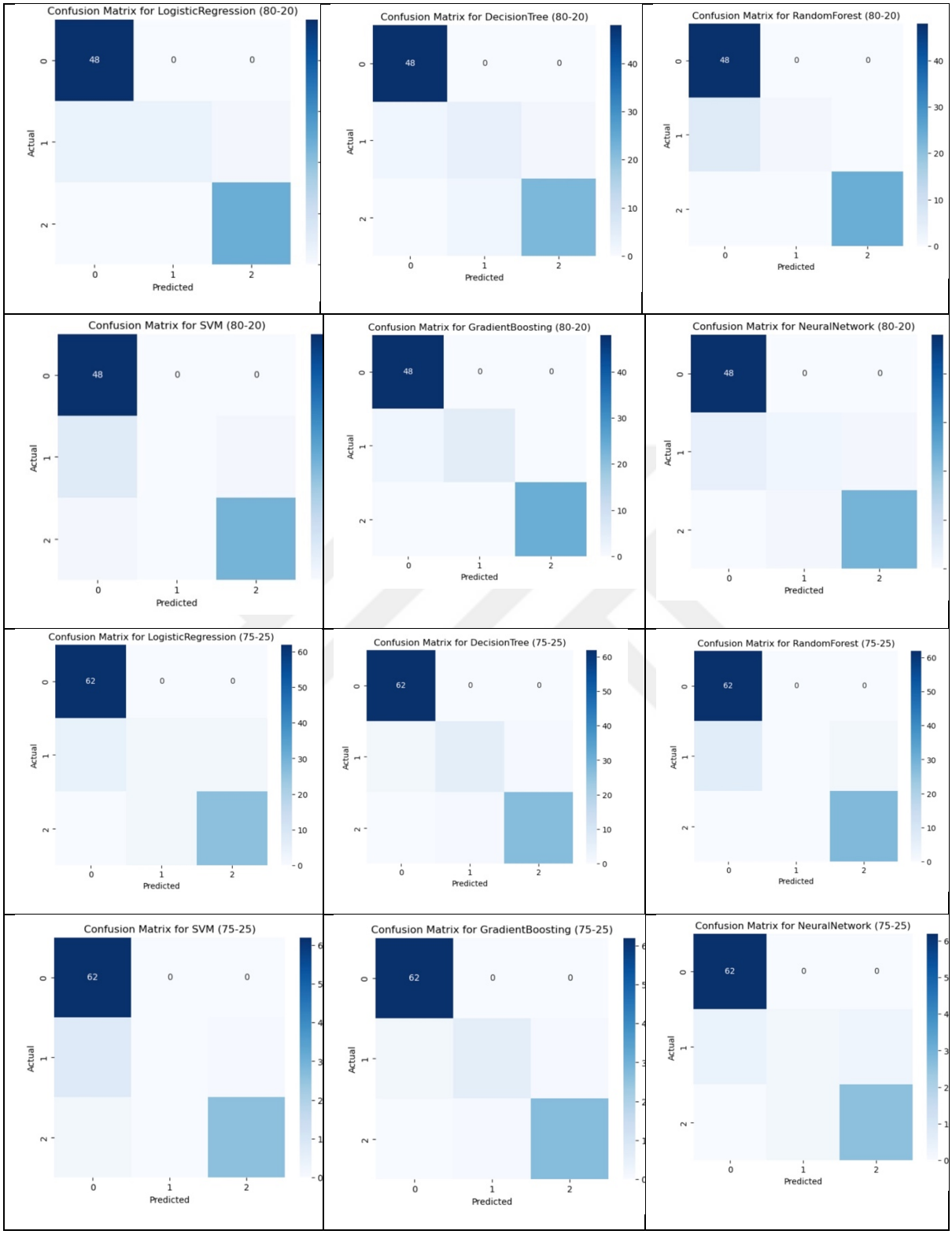
Model Performans Değerlendirmesi :

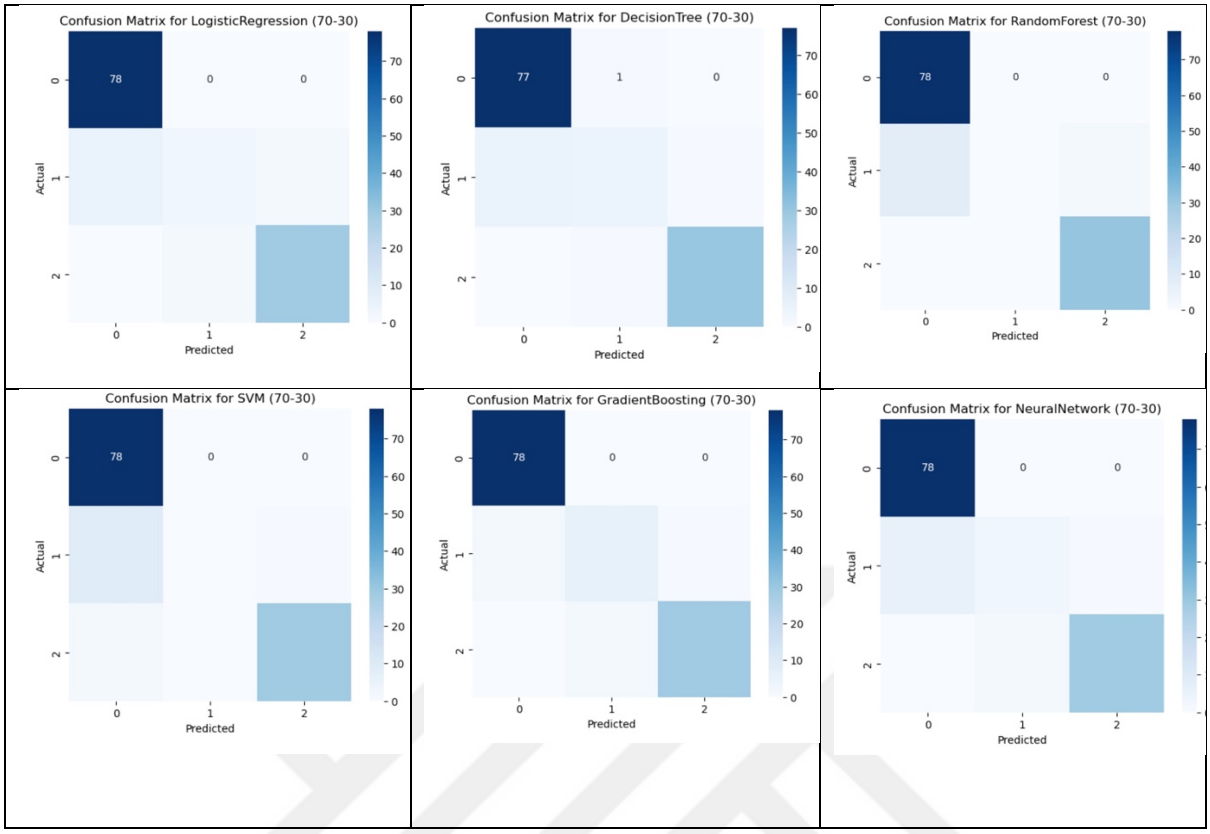
Her modelin performansı doğruluk, kesinlik, geri çağırma ve F1 puanı kullanılarak değerlendirildi.



Şekil 21: Model Doğruluk Karşılaştırması

Farklı sınıflar arasındaki sınıflandırma doğruluğunu görselleştirmek amacıyla her model için karışıklık matrisleri oluşturuldu.


Confusion Matrix for LogisticRegression (80-20)
Confusion Matrix for DecisionTree (80-20)
Confusion Matrix for RandomForest (80-20)
Confusion Matrix for SVM (80-20)
Confusion Matrix for GradientBoosting (80-20)
Confusion Matrix for NeuralNetwork (80-20)
Confusion Matrix for LogisticRegression (75-25)
Confusion Matrix for DecisionTree (75-25)
Confusion Matrix for RandomForest (75-25)
Confusion Matrix for SVM (75-25)
Confusion Matrix for GradientBoosting (75-25)
Confusion Matrix for NeuralNetwork (75-25)
Actual
Predicted

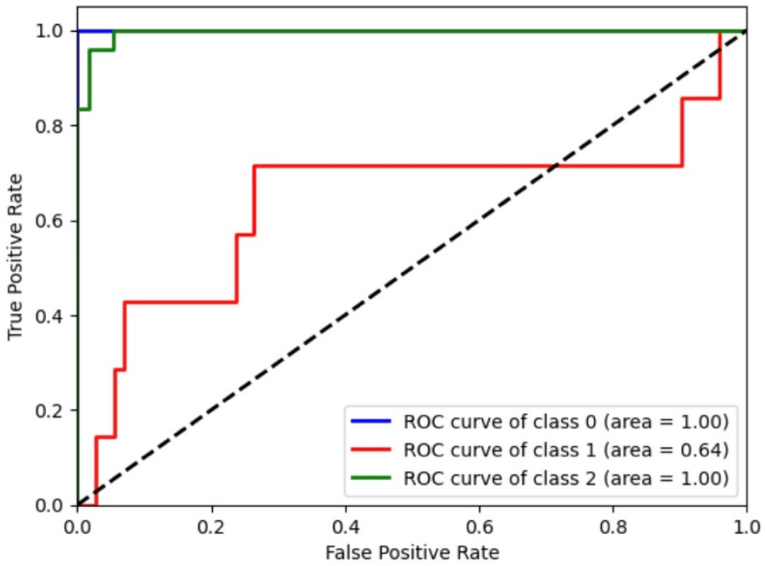


Şekil 22: Her Sınıflandırma Modeli İçin Karışıklık Matrisleri

Modellerin sınıfları ayırt etme yeteneğini değerlendirmek için her sınıf için ROC eğrileri çizildi.

```
# Compute ROC curve and ROC area for each class
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(n_classes):
    fpr[i], tpr[i], _ = roc_curve(y_test[:, i], y_score[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])

# Plotting the ROC curve
colors = cycle(['blue', 'red', 'green'])
for i, color in zip(range(n_classes), colors):
    plt.plot(fpr[i], tpr[i], color=color, lw=2,
             label='ROC curve of class {0} (area = {1:0.2f})'
             ''.format(i, roc_auc[i]))
```

Şekil 23: Her sınıf için ROC eğrisi ve alanının hesaplanması ve görselleştirilmesi kodlanıyor.



Şekil 24: Çok Sınıflı ROC Eğrisi

5 MATEMATİK EĞİTİMİNDE UYGULAMA

Bu sınıflandırma analizinin sonuçlarının matematik eğitimi için çeşitli çıkarımları vardır:

1. **Erken Müdahale** : Düşük performans (Yetersiz) riski taşıyan öğrencilerin belirlenmesi, zamanında müdahale ve desteğe olanak sağlar.
2. **Müfredat Uyarlama** : Farklı başarı düzeylerini etkileyen faktörlerin anlaşılması, çeşitli öğrenme ihtiyaçlarını karşılamak için müfredat düzenlemelerine yol gösterebilir. Farklılaştırılmış eğitim gibi aynı sınıf seviyesindeki farklı başarı düzeyindeki öğrencilerin bir arada çalışmasını daha etkin sağlayacak düzenlemelere zemin oluşturabilir.
3. **Kaynak Tahsisi** : Öğrencilerin başarı düzeylerine göre dağılımını anlamak, eğitim kaynaklarının daha etkili bir şekilde tahsis edilmesine yardımcı olur.

5.1 Genel Analiz: Regresyon ve Sınıflandırma Modelleri

Önemli bulgular

1. **Regresyon Analizi** : Öğrenci performansının önemli belirleyicilerini ortaya çıkardı ve bunların etkilerini ölçtü. *Ridge* Regresyonu ve Gradyan Artırma gibi modeller yüksek tahmin doğruluğu gösterdi.

2. **Sınıflandırma Analizi** : Derecelendirme Arttırma ve Karar Ağaçlarının oldukça iyi performans göstermesiyle öğrencileri etkili bir şekilde farklı başarı düzeylerine ayırdı.

5.2 Eğitimin Önemi

- Regresyon ve sınıflandırma modellerinin birlikte kullanılması, öğrenci performansını etkileyen faktörlerin kapsamlı bir şekilde anlaşılmasını sağlar.
- Bu modeller eğitim politikalarına, öğretim stratejilerine ve bireyselleştirilmiş öğrenci destek mekanizmalarına bilgi sağlayabilir.

5.3 Güçlü ve Zayıf Yönler

- **Güçlü Yönler** : Bu analiz, güçlü öngörüler sunan geniş bir tahmin modeli yelpazesini kapsar. Farklı eğitim-test bölümlerinin kullanılması (70-30, 75-25, 80-20) bulguların güvenilirliğini artırır.
- **Zayıf Yönler** : Çalışma, veri setinin kapsamı ve boyutu nedeniyle sınırlı olabilir. Bazı modeller, daha incelikli ayarlara ihtiyaç duyulduğunu gösteren aşırı uyum belirtileri gösterdi.

5.4 Gelecekteki Araştırmalara Yönelik Öneriler

1. **Veri Zenginleştirme** : Daha çeşitli ve kapsamlı veri kümelerinin dahil edilmesi, analizin genellenebilirliğini artırabilir.
2. **Model Araştırması** : Daha gelişmiş veya yeni makine öğrenimi teknikleriyle denemeler yapmak daha derin içgörüler sağlayabilir.
3. **Nedenselliğe Odaklanma** : Gelecekteki çalışmalar, öğrenci performansının dinamiklerini daha iyi anlamak için sadece korelasyonlar değil, nedensel ilişkiler kurmayı da hedefleyebilir.

5.5 Sonuç

Hem regresyon hem de sınıflandırma yaklaşımlarını kapsayan bu kapsamlı veri analizi, öğrencinin matematikteki performansına ilişkin değerli bilgiler sağlamıştır. Bulgular eğitim stratejileri için önemli çıkarımlara sahiptir ve veriye dayalı yaklaşımların eğitim sonuçlarını iyileştirmedeki potansiyelini vurgulamaktadır. Ancak bu tür analizlerin sürekli olarak

iyileştirilmesi ve genişletilmesi, daha derin ve eyleme dönüştürülebilir içgörüler için çok önemlidir.

6 DENEYLER

Bu araştırmada kullanılan veri seti 395 öğrenciye ilişkin detaylı bilgileri içeren Kaggle'dan alınmıştır. Demografik ayrıntıları, sosyoekonomik durumu, sağlığı, sporu, etkinlikleri, tutumları ve akademik notları kapsayan 33 farklı özelliği içerir. Veri seti 16 sayısal ve 17 kategorik öznitelik ile iyi yapılandırılmıştır ve özellikle eksik veri içermemektedir. Veriler Python programlama dili kullanılarak, Python'un güçlü kütüphanelerinden ve araçlarından yararlanılarak işlendi ve analiz edildi. Kullanılan anahtar kütüphaneler şu şekildedir:

- **Pandas ve NumPy** : Veri işleme ve sayısal hesaplamalar için.
- **Seaborn ve Matplotlib** : Veri görselleştirmesi için.
- **Scikit-learn** : Makine öğrenimi algoritmalarını uygulamak için.
- **Jupyter Notebook** : Anaconda dağıtımı tarafından sağlanan, kodlama için birincil platform olarak hizmet verdi. Bu etkileşimli ortam, Python kodunun yürütülmesini, görselleştirilmesini ve analizini kusursuz bir şekilde kolaylaştırdı.

Hesaplamalar veri analizi görevlerinin verimli bir şekilde işlenmesini sağlayan bir Macbook Pro M2 çipli dizüstü bilgisayarda gerçekleştirildi.

Deneyler, regresyon ve sınıflandırma olmak üzere iki ana kategoriye ayrılmıştır.

1. **Regresyon Modelleri** :
    - **Amaç** : Öğrencilerin ortalama notlarını (**G_avg**) çeşitli özniteliklere göre tahmin etmek.
    - **Kullanılan Modeller** : *Ridge* Regresyonu, *Lasso* Regresyonu ve Elastik Net Regresyonu, Rassal Orman Regresyonu ve Gradyan Artırma Regresyonu.
    - **Bulgular** : Tüm regresyon modelleri karşılaştırılabilir puanlar verdi. Ancak hata marjları açısından en iyi performansı *Ridge* Regresyonu gösterdi.
2. **Sınıflandırma Modelleri** :
    - **Amaç :** Öğrencileri **G_avg'**ye dayalı olarak akademik başarı açısından üç kategoriye ayırmak : Yetersiz, Tatmin Edici ve Başarılı. (*Unsatisfactory, Satisfactory and Successful*)

- **Kullanılan Modeller** : Lojistik Regresyon, Karar Ağaçları, Rastgele Orman, Destek Vektör Makineleri (SVM), Gradyan Artırma ve Yapay Sinir Ağları (MLP Sınıflandırıcı).
- **Bulgular** : Modeller, doğruluk ve ROC eğrisi altındaki alan açısından değerlendirilmiştir. Lojistik Regresyon ve Rastgele Orman en iyi performans gösteren modeller olarak öne çıkmıştır.

3. **Derin Öğrenme Yaklaşımı** :
    - **Kullanılan Model** : Yapay Sinir Ağları.
    - **Sonuç** : Potansiyeli gösteren ancak daha fazla optimizasyon ihtiyacını vurgulayan 0,64'lük bir doğruluk elde edildi.

7 SONUÇ VE GELECEKTEKİ YÖNLENDİRMELER

Bu çalışma, kapsamlı bir veri seti kullanarak, sınıflandırma, regresyon ve kümeleme yöntemlerinin bir karışımını kullanarak, akademik performansın derinlemesine bir analizini başlatmıştır. Araştırma, tahmin algoritmalarından yararlanarak, öğrenciyle ilgili çeşitli öznitelikler ile bunların akademik sonuçları arasındaki karmaşık ilişkileri çözmeyi amaçladı. Önceki bölümlerde ayrıntıları verilen çalışmanın bulguları, akademik performansı etkileyen faktörlere dair önemli bilgiler sunmaktadır.

Önemli Katkılar

1. **Çeşitli Metodolojik Yaklaşım** : Regresyon ve sınıflandırma modelleri de dahil olmak üzere çoklu analitik tekniklerin uygulanması, akademik performansı etkileyen faktörlere ilişkin bütünsel bir bakış açısı sağlamıştır. Bu yaklaşım, öğrenci performansının hem tahmin edilmesine (regresyon) hem de kategorize edilmesine (sınıflandırılmasına) olanak sağladı.
2. **Öznitelik Analizi** : Demografik, sosyoekonomik, tutumsal, sağlık, spor ve akademik yönleri kapsayan çok çeşitli öznitelikler değerlendirildi. Bu kapsamlı analiz, akademik performansın çok yönlü doğasının anlaşılmasına yardımcı oldu.
3. **Model Performansı** : Çalışma, *Ridge Regresyonu* ve *Logistic Regression* gibi modellerin sırasıyla akademik performansı tahmin etme ve sınıflandırmada yüksek etkinlik gösterdiği önemli sonuçlar vermiştir.

7.1 Eğitimsel Etkileri

Araştırmanın sonuçlarının çeşitli bağlamlarda eğitim stratejilerini geliştirmeye yönelik pratik sonuçları vardır. Eğitimciler ve politika yapıcılar, akademik başarının belirleyicilerini anlayarak müdahaleleri ve destek mekanizmalarını daha etkili bir şekilde uyarlayabilirler. Bu araştırmada uygulanan modeller ve algoritmalar, çeşitli eğitim ortamlarından gelen verileri analiz etmek için kullanılabilir ve potansiyel olarak daha geniş eğitim ortamına katkıda bulunabilir.
Gelecekteki yönlendirmeler İleriye bakıldığında, çalışma aşağıdaki yollarla etkisini genişletmeyi amaçlamaktadır:

1. **Veri Setinin Genişletilmesi** : Türkiye'deki farklı bölgelerden, okul türlerinden ve sınıf seviyelerinden verilerin bir araya getirilerek daha kapsamlı ve çeşitli bir veri seti oluşturulması.
2. **Model Uygulamasının Geliştirilmesi** : Bulguları doğrulamak ve iyileştirmek için başarılı model ve algoritmaların bu genişletilmiş veri kümesine uygulanması.
3. **Çeşitlendiren Öznitelik Seti** : Akademik performansın daha incelikli bir şekilde anlaşılmasını sağlamak için duygusal zeka, yaşam memnuniyeti, öğrenme türleri, motivasyon ve sosyal destek ile ilgili ek öznitelikler içerir.
4. **Bağlamlar Arası Analiz** : Bulguların ve modellerin evrensel uygulanabilirliğini keşfetmek için araştırmayı çeşitli eğitim bağlamlarına genişletmek.

7.2 Sonuç

Bu çalışma, akademik performansın anlaşılması ve geliştirilmesinde veriye dayalı yaklaşımların kullanılmasına yönelik önemli bir adımı temsil etmektedir. Makine öğrenimi modellerinin başarılı bir şekilde uygulanması, gelecekteki eğitim araştırmaları ve uygulamaları için umut verici bir yol sunmaktadır. Nihai hedef, farklı eğitim ortamlarında öğrenme deneyimlerini ve sonuçlarını olumlu yönde etkilemek için bu içgörülerden yararlanmaktır.

8 PYTHON CODES

Bu tezde yapılan analizlerin kapsamlı bir şekilde anlaşılması ve tekrarlanabilmesi için, kullanılan kodların hem Jupyter Notebook dosyası hem de bu kodların PDF versiyonu paylaşılmıştır.

Ayrıca, bu tezde kullanılan verileri anlamak için *Kaggle*'dan alınan csv dosyası ve açıklaması da dahil olmak üzere tüm dosyalara bu LINK'ten ulaşılabilir (Cortez and Silva).

Dosya İsmi	**İçerik**
student_math.csv	Kaggle'dan alınan, üzerinde analiz yapılan dosya
Student-mat_data_anlaysis_Erdal_Thesis.ipynb	Jupyter Notebook'ta yazılan kodların tamamını içeren dosya
Student-mat_data_anlaysis_Erdal_Thesis.pdf	Jupyter Notebook'ta yazılan kodların PDF'e dönüştürülmüş kopyası
features_explanations.csv	Öznitelikler ve açıklamaların bulunduğu csv uzantılı tablo

REFERANSLAR:

Allison, Paul D. *Multiple Regression: A Primer*. Pine Forge Press, 1999.

Baker, Ryan S., and Paul Salvador Inventado. *Educational Data Mining and Learning Analytics*. 2018.

Bishop, Christopher M., and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*. no. 4, Springer, 2006.

Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society*, vol. 15, no. 5, 2012, pp. 662–79.

Breiman, Leo. "Random Forests." *Machine Learning*, vol. 45, 2001, pp. 5–32.

---. "Random Forests." *Machine Learning*, vol. 45, 2001, pp. 5–32.

Burges, Christopher J. C. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery*, vol. 2, no. 2, 1998, pp. 121–67.

Chatterjee, Samprit, and Ali S. Hadi. *Regression Analysis by Example*. John Wiley & Sons, 2013.

Chen, Tianqi, and Carlos Guestrin. "Xgboost: A Scalable Tree Boosting System." *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–94.

Cortes, Corinna, and Vladimir Vapnik. "Support-Vector Networks." *Machine Learning*, vol. 20, 1995, pp. 273–97.

Cortez, Paulo, and Alice Maria Gonçalves Silva. *Using Data Mining to Predict Secondary School Student Performance*. 2008.

Cutler, D. Richard, et al. "Random Forests for Classification in Ecology." *Ecology*, vol. 88, no. 11, 2007, pp. 2783–92.

Dueben, Peter D., et al. "Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook." *Artificial Intelligence for the Earth Systems*, vol. 1, no. 3, 2022, p. e210002.

Elrahman, Ahmed Abd, et al. “A Predictive Model for Student Performance in Classrooms Using Student Interactions With an ETextbook.” *ArXiv Preprint ArXiv:2203.03713*, 2022.

Farouni, Rick. “A Contemporary Overview of Probabilistic Latent Variable Models.” *ArXiv Preprint ArXiv:1706.08137*, 2017.

Farrar, Donald E., and Robert R. Glauber. “Multicollinearity in Regression Analysis: The Problem Revisited.” *The Review of Economic and Statistics*, 1967, pp. 92–107.

Friedman, Jerome H. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics*, 2001, pp. 1189–232.

---. “Stochastic Gradient Boosting.” *Computational Statistics & Data Analysis*, vol. 38, no. 4, 2002, pp. 367–78.

Hastie, Trevor, et al. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Haykin, Simon. *Neural Networks and Learning Machines, 3/E*. Pearson Education India, 2009.

Heaton, Jeff. “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning: The MIT Press, 2016, 800 Pp, ISBN: 0262035618.” *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, 2018, pp. 305–07.

---. “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning: The MIT Press, 2016, 800 Pp, ISBN: 0262035618.” *Genetic Programming and Evolvable Machines*, vol. 19, no. 1–2, 2018, pp. 305–07.

Hoerl, Arthur E., and Robert W. Kennard. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*, vol. 12, no. 1, 1970, pp. 55–67.

Hosmer Jr, David W., et al. *Applied Logistic Regression*. John Wiley & Sons, 2013.

James, Gareth, et al. *An Introduction to Statistical Learning*. Springer, 2013.

Kefalaki, Margarita, et al. “Technology and Education. Innovation and Hindrances.” *Journal of Applied Learning & Teaching*, vol. 5, no. S1, 2022.

Kovac, Aljaz. *Sentence Embeddings and Automatic Classification of Menu Items*. 2022.

Kreuzberger, Dominik, et al. “Machine Learning Operations (Mlops): Overview, Definition, and Architecture.” *IEEE Access*, 2023.

Kuhn, Max, and Kjell Johnson. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC, 2019.

Kutner, Michael H., et al. *Applied Linear Statistical Models*. McGraw-hill, 2005.

LeCun, Yann, et al. “Deep Learning.” *Nature*, vol. 521, no. 7553, 2015, pp. 436–44.

Liaw, Andy, and Matthew Wiener. “Classification and Regression by RandomForest.” *R News*, vol. 2, no. 3, 2002, pp. 18–22.

---. “Classification and Regression by RandomForest.” *R News*, vol. 2, no. 3, 2002, pp. 18–22.

Lorenzen, Stephan, et al. “Tracking Behavioral Patterns among Students in an Online Educational System.” *ArXiv Preprint ArXiv:1908.08937*, 2019.

Maimon, Oded Z., and Lior Rokach. *Data Mining with Decision Trees: Theory and Applications*. World scientific, 2014.

Mandalapu, Varun, et al. “Do We Need to Go Deep? Knowledge Tracing with Big Data.” *ArXiv Preprint ArXiv:2101.08349*, 2021.

Marquardt, Donald W., and Ronald D. Snee. “Ridge Regression in Practice.” *The American Statistician*, vol. 29, no. 1, 1975, pp. 3–20.

Menard, Scott. *Applied Logistic Regression Analysis*. no. 106, Sage, 2002.

Natekin, Alexey, and Alois Knoll. “Gradient Boosting Machines, a Tutorial.” *Frontiers in Neurorobotics*, vol. 7, 2013, p. 21.

---. “Gradient Boosting Machines, a Tutorial.” *Frontiers in Neurorobotics*, vol. 7, 2013, p. 21.

Orji, Fidelia A., and Julita Vassileva. “Machine Learning Approach for Predicting Students Academic Performance and Study Strategies Based on Their Motivation.” *ArXiv Preprint ArXiv:2210.08186*, 2022.

Provost, Foster, and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. " O'Reilly Media, Inc.," 2013.

Quinlan, J. Ross. *C4. 5: Programs for Machine Learning*. Elsevier, 2014.

Raschka, Sebastian. "An Overview of General Performance Metrics of Binary Classifier Systems." *ArXiv Preprint ArXiv:1410.5330*, 2014.

Rebala, Gopinath, et al. "Machine Learning Definition and Basics." *An Introduction to Machine Learning*, 2019, pp. 1–17.

Ridgeway, Greg. "Generalized Boosted Models: A Guide to the Gbm Package." *Update*, vol. 1, no. 1, 2007, p. 2007.

Schmidhuber, Jürgen. "Deep Learning in Neural Networks: An Overview." *Neural Networks*, vol. 61, 2015, pp. 85–117.

---. "Deep Learning in Neural Networks: An Overview." *Neural Networks*, vol. 61, 2015, pp. 85–117.

Schölkopf, Bernhard, and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Tadayon, Manie, and Gregory J. Pottie. "Predicting Student Performance in an Educational Game Using a Hidden Markov Model." *IEEE Transactions on Education*, vol. 63, no. 4, 2020, pp. 299–304.

Tibshirani, Hastie Robert, et al. *An Introduction to Statistical Learning*. springer publication, 2017.

Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, 1996, pp. 267–88.

Viberg, Olga, et al. "The Current Landscape of Learning Analytics in Higher Education." *Computers in Human Behavior*, vol. 89, 2018, pp. 98–110.

Vipin, Pang-Ning Tan Michael Steinbach. *Introduction to Data Mining*. 2006.

Waheed, Hajra, et al. "Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models." *Computers in Human Behavior*, vol. 104, 2020, p. 106189.

Zheng, Alice, and Amanda Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. " O'Reilly Media, Inc.," 2018.

Zou, Hui, and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, 2005, pp. 301–20.

---. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, 2005, pp. 301–20.

Zou, Hui, and Hao Helen Zhang. "On the Adaptive Elastic-Net with a Diverging Number of Parameters." *Annals of Statistics*, vol. 37, no. 4, 2009, p. 1733.