

Öğrencilerin Sınavlardaki Performansının Makine Öğrenmesi Teknikleriyle Tahminlenmesi

Ayşe ALKAN^{1*}

¹Samsun Bilim ve Sanat Merkezi, 55070, Samsun

¹<https://orcid.org/0000-0002-9125-1408>

*Sorumlu yazar: ayse.alkan55@gmail.com

Araştırma Makalesi

Makale Tarihiçesi:

Geliş tarihi: 16.01.2024

Kabul tarihi: 14.02.2024

Online Yayınlanma: 25.06.2024

Anahtar Kelimeler:

Makine öğrenmesi

Bilişim

Eğitim

Sınıflandırma

Tahminleme

ÖZ

Yapay zekânın bir alt dalı olan makine öğrenmesi eğitim alanında hızla popülerlik kazanan bir bilgisayar bilimi dalıdır. Eğitimde öğrenci gelişiminin, eğitim programlarının etkisinin, öğretim yöntemlerinin etkililiğinin belirlenmesi adımı olan değerlendirme aşamasının önemi büyüktür. Öğrencilere ait veri yığınlarından anlamlı sonuçlar üretilmesini sağlayan makine öğrenmesi, değerlendirme aşamasında yön gösterici olmaktadır. Bu çalışmada kamuya açık olarak paylaşılan, “Students Performance in Exams” veri seti kullanılmıştır. Öğrencilere ait bilgilerin bulunduğu 8 öznitelikten oluşan veri seti üzerinde öğrenci başarısının tahminlenmesine yönelik makine öğrenmesi temelli sınıflandırmalar gerçekleştirilmiştir. K-en Yakın Komşu (KNN), Karar Ağacı (KA), Naive Bayes (NB), Rastgele Orman (RO), Destek vektör makinesi (DVM), Lojistik regresyon (LR), Lineer Diskriminant Analizi (LDA) ile toplamda yedi ayrı yöntemle ve 5 kat çapraz doğrulamayla yapılan sınıflandırma işlemleri sonucunda en yüksek doğruluk 0,67 ile DVM algoritması ile elde edilmiştir.

Predicting Students' Performance in Exams with Machine Learning Techniques

Research Article

Article History:

Received: 16.01.2024

Accepted: 14.02.2024

Published online: 25.06.2024

Keywords:

Machine learning

Informatics

Education

Classification

Estimation

ABSTRACT

Machine learning, a sub-branch of artificial intelligence, is a computer science branch that is rapidly gaining popularity in the field of education. The evaluation phase, which is the step of determining student development in education, the effect of educational programs, and the effectiveness of teaching methods, is of great importance. Machine learning, which enables the production of meaningful results from data stacks belonging to students, is guiding in the evaluation phase. In this study, the publicly shared “Students Performance in Exams” data set was used. Machine learning-based classifications were carried out to predict student success on the data set consisting of 8 attributes containing information about students. K-nearest neighbor (KNN), Decision tree (KA), Naive Bayes (NB), Random forest (RO), Support vector machine (DVM), Logistic regression (LR), Linear Discriminant Analysis (LDA) The highest accuracy was obtained with the DVM algorithm as a result of the classification processes made with the method and 5-fold cross-validation.

To Cite: Alkan A. Öğrencilerin Sınavlardaki Performansının Makine Öğrenmesi Teknikleriyle Tahminlenmesi. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2024; 7(3): 1116-1128.

1. Giriş

Bilim ve teknolojideki hızlı ilerleme ve gelişimler hayatımızın her alanına da aynı hızla girmektedir. Bireysel olarak yapılması çok uzun zaman ve maliyet gerektiren işlemler teknolojik araç gereçlerle kısa zamanda ve düşük maliyette gerçekleştirilebilmektedir. Günümüzün popüler gelişmelerinden birisi de yapay zekânın bir dalı olan makine öğrenmesidir. Makine öğrenmesinin yapay zekâ alanındaki konumu Şekil 1’de gösterilmektedir (Machine, 2017). Tarih sürecince bilim insanları “makinelere düşünebilir mi?” sorusuna cevap aramışlardır (Turing, 1950). Makine öğrenmesi ile bilgisayarların geçmiş verilerdeki deneyimlerinden faydalanılarak geleceğe yönelik tahminlerde bulunulurken istatistiksel modeller kullanılmaktadır (Alpaydin, 2020; Koza ve ark., 1996).



Şekil 1. Yapay zekâ ve makine öğrenmesi ilişkisi

Farklı disiplinlerde kullanılabilme imkânı sunan bu teknoloji, verilerden öğrenme yoluyla bilgi çıkarabilme ve tahminler yapabilme yeteneğine sahiptir. Makine öğrenmesi, veri analitiği ve karar verme süreçlerinde büyük veri setlerini işleyerek değerli bilgiler çıkarabilmekte ve stratejik planlamalara yardımcı olabilmektedir. Ayrıca kişiselleştirilmiş deneyimler sunma, sağlık hizmetlerini geliştirme, iş süreçlerini otomatikleştirme, doğal dil işleme, ulaşım ve finans sektöründe verimliliği artırma gibi birçok alanda önemli katkıları bulunmaktadır. Son zamanlarda eğitim alanında büyük veri yığınlarının bulunması makine öğrenmesinin eğitim alanında kullanılması ihtiyacını ortaya çıkarmıştır. Takvim yaşına göre aynı sınıf ortamında bulunan öğrencilerin sosyo-demografik özellikleri ve öğrenme yetenekleri birbirinden farklılık gösterebilmektedir. Olası problemlere karşı önceden alınabilecek önlemlerle öğrenci başarılarının artması sağlanırken, özel yetenekli öğrenciler önceden belirlenerek bireylere özel programların hazırlanmasında makine öğrenmesi avantajlar sunmaktadır. Eğitim alanında, yapay zekâ ve veri bilimindeki hızlı gelişmelerle birlikte, doğru ve zengin öğrenme verileri toplanabilmekte ve özel öğrenme ihtiyaçlarını ortaya çıkarabilmektedir. Herkese uyan tek bir yaklaşımdan hassas eğitime geçiş sağlanabilmektedir (Lu ve ark., 2018; Tsai ve ark., 2020). Ayrıca; Xie ve ark., (2019), şimdiye kadar, kişiselleştirilmiş sistemlerin büyük bir kısmının geleneksel bilgisayarları veya cihazları kullandığını, buna karşılık giyilebilir cihazların, akıllı telefonların ve tablet bilgisayarlar gibi akıllı cihazların daha az tercih edildiğini belirtirken yapay zekânın ise; bu kişiselleştirilmiş öğrenme sistemleri üzerinde önemli etki yarattığını belirtmektedirler.

Eđitimde makine öğrenmesinin kişiselleştirilmiş öğrenme, öğrenci takibi, erken müdahale, verimlilik ve tasarruf sağlaması, öğretmenlere destek olması, sanal gerçeklik, artırılmış gerçeklik gibi yeni teknolojilerle birleştirilerek eğitimde yenilikçi yaklaşımların geliştirilmesine olanak tanınması sebebi ile bu çalışmada da öğrenci başarılarının tahminlenmesine yönelik makine öğrenmesi temelli sınıflandırmalar gerçekleştirilerek karşılaştırmalı bir şekilde incelenmiştir.

2. Yapılmış Çalışmalar

Djulovic ve Li (2013) üniversite birinci sınıf öğrencilerinin öğrenimlerine devam durumunu tahminlemek için farklı algoritmalarını kullanarak modeller geliştirmişlerdir. %86,27 ile en yüksek genel doğruluğa elde etmişlerdir. Iam-On ve Boongoen (2017) üniversite öğrencilerinin öğrenimi bırakma eğilimlerini belirlemek için çalışmışlar ve araştırma sonucunda başarılı bir akademik geçmişi olan öğrencilerin devam durumlarının yüksek olduğunu belirtmişlerdir. Lise öğrencilerinin okulu terk etme durumlarının tahmini üzerine çalışan Chung ve Lee (2019) çalışmalarında %95 doğruluk oranı ile sonuçlar üretmiştir. Iatrellis, Savvasi, Fitsilis ve Gerogiannis (2021) yükseköğrenim programlarındaki öğrencilerin sonuçlarını tahmin etmek için çalışmalarında iki aşamalı bir makine öğrenimi yaklaşımı ile yüksek doğrulukta tahminler üretildiğini belirtmişlerdir. Badal ve Sungkur (2022) öğrencilerin performansını tahmin etmek ve çevrimiçi öğrenme platformunun özelliklerini analiz etmek amacıyla yapmış oldukları çalışmada öğrenci profili ve etkileşimle ilgili özniteliklerle not ve katılım tahmini için %85 ve %83 ile başarılı sonuçlar elde etmişlerdir. Çakıt ve Dağdeviren (2022) üniversitenin akademik itibarı, üniversitenin bulunduğu şehrin imkânları, üniversitenin tesisleri ve kültürel imkânları ile ilgili verilere dayalı olarak öğrenci yerleştirme yüzdesini tahmin etmeye yönelik yapmış oldukları çalışmada farklı farklı makine öğrenimi yaklaşımlarının başarılarını karşılaştırmışlardır. Guleria ve Sood (2023) öğrencilerin kariyer danışmanlığı için bir çerçeve önerdikleri çalışmalarında farklı makine öğrenimi yaklaşımlarının performanslarını incelemişlerdir. Chen ve Zhai (2023) çalışmalarında farklı uygulama senaryolarında makine öğrenimi yöntemlerinin performansını araştırmak için üç farklı türde göreve yönelik eğitim verisi kullanılmışlardır. Çalışmaları sonucunda deneysel sonuçlarda makine öğrenimi algoritmalarının başarılarını incelemişlerdir. Seveli ve Başer (2020) akademik başarıyı tahmin ettikleri çalışmalarında %86,4 doğrulukla rastgele orman algoritması en yüksek performansı sergilemiştir.

3. Materyal ve Metot

Bu çalışmanın amacı, öğrencilerin sınavlardaki performansının farklı sınıflandırıcılar kullanılarak tahmin edilmesini sağlamaktır. Çalışmada kullanılan “Students Performance in Exams” veri setine, Kamuya açık bir veri deposundan erişilmiştir (Web1). Veri setinde öğrencilerinin sosyo-demografik bilgileri ile matematik, okuma ve yazma derslerine ait notları bulunmaktadır. Çalışma, veri setinde farklı makine öğrenmesi algoritmaları kullanarak öğrencilerin sınavlardaki performansının tahminine yönelik sınıflandırma çalışmalarını içermektedir. Sınıflandırma işlemleri için K-en Yakın Yomşu

(KNN), Karar Ağacı (KA), Naive Bayes (NB), Rastgele Orman (RO), Destek Vektör Makinesi (DVM), Lojistik Regresyon (LR), Lineer Diskriminant Analizi (LDA) yöntemleri kullanılmış ve performansları değerlendirilmiştir.

3.1. Veri Seti

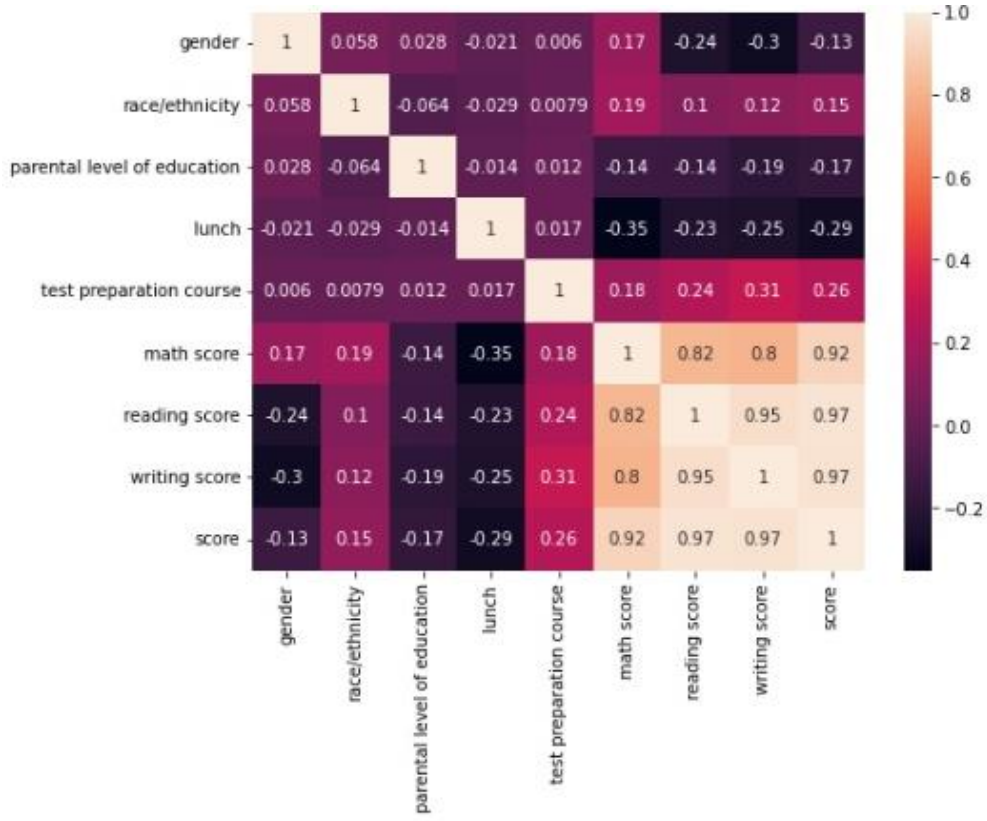
Veri seti 1000 örnekten ve 8 tane öznelikten oluşmaktadır. Veri setinde bulunan özneliklerin ayrıntıları Tablo 1’de verilmektedir. Veri setinde bulunan özneliklerin istatistiki bilgileri de Tablo 2’de yer almaktadır. İki değişken arasındaki ilişkinin gücünü ve yönünü ölçmek amacıyla veri seti özneliklerinin korelasyon ilişkileri Şekil 2’de verilmiştir. Korelasyon katsayısı değer aralığı -1 ile 1 arasında olmaktadır. 1: Mükemmel pozitif korelasyonu, 0: İlişkisizliği ve -1: Mükemmel negatif korelasyonu göstermektedir. Şekil 2 incelendiğinde, writing score ve reading score arasında yüksek oranda ilişki olduğu görülmektedir. Bu durumda bir öğrencinin iyi okuması durumunda iyi yazdığı söylenebilir. Math_score’un çok yüksek bir korelasyonu bulunmamaktadır, bu sebeple bir öğrencinin matematikte iyi performans göstermesi durumunda okuma ve yazma da iyi performans göstermesinin gerekli olmadığı söylenebilir. Math score, writing score ve reading score arasındaki korelasyon ilişkileri ayrıntılı olarak Şekil 3’te gösterilmektedir.

Tablo 1. Veri setinde bulunan öznelikler

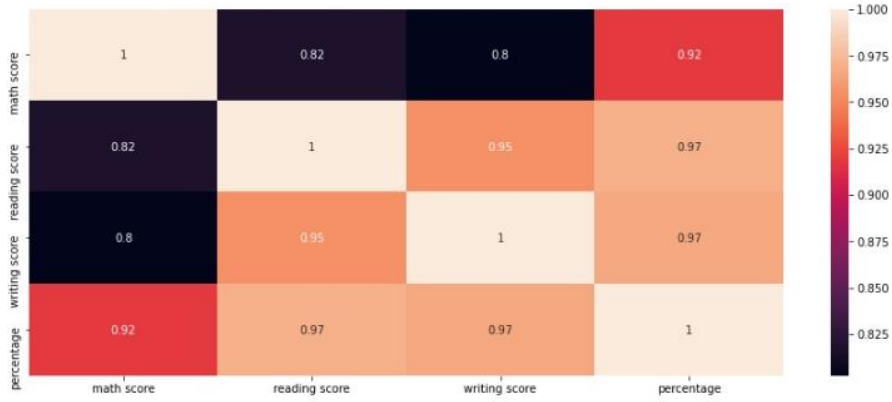
| Öznelik Adı | Açıklama |
|-----------------------------|-----------------------|
| gender | [erkek/ kadın] |
| race/ethnicity | [A’dan E’ye gruplar] |
| parental level of education | [ebeveyn eğitimi] |
| lunch | [okula yerleşim türü] |
| test preparation course | [test hazırlık kursu] |
| math score | [matematik skoru] |
| reading score | [okuma skoru] |
| writing score | [yazma skoru] |

Tablo 2. Veri setindeki özneliklerin tanımlayıcı istatistikleri

| Öznelik | Ortalama | Standart Sapma | En Küçük | En Büyük |
|-----------------------------|----------|----------------|----------|----------|
| gender | 0,482 | 0,499 | 0,0 | 1,0 |
| race/ethnicity | 2,174 | 1,157 | 0,0 | 4,0 |
| parental level of education | 2,486 | 1,829 | 0,0 | 5,0 |
| lunch | 0,645 | 0,478 | 0,0 | 1,0 |
| test preparation course | 0,642 | 0,479 | 0,0 | 1,0 |
| math score | 46,145 | 14,980 | 0,0 | 80,0 |
| reading score | 41,212 | 14,382 | 0,0 | 71,0 |
| writing score | 44,104 | 15,043 | 0,0 | 76,0 |



Şekil 2. Öznitelikler arasındaki korelasyon



Şekil 3. Math score, writing score ve reading score arasındaki korelasyon

3.2. Sınıflandırma Yöntemleri

Bu bölümde çalışma kapsamında kullanılan sınıflandırma modelleri açıklanmıştır.

K-En Yakın Komşu (K-Nearest Neighbors, KNN) Algoritması, sınıflandırma ve regresyon problemlerini çözmek için kullanılan basit ve temel bir makine öğrenmesi algoritmasıdır. Veri noktalarını uzayda birbirine olan yakınlıklarına göre gruplamak ve tahmin yapmak için kullanılır. Öklide göre mesafenin belirlenmesinde Denklem 1'deki formül kullanılmaktadır (Kilinç ve ark., 2016).

$$d(i,y) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{yk})^2} \quad (1)$$

Denklemdaki sembollerin açıklamaları şu şekildedir:

d: İki nokta arasındaki uzaklık.

i: Birinci noktanın indis değeri.

y: İkinci noktanın indis değeri.

p: Özellik vektörlerinin boyutu (özellik sayısı).

x_{ik}: Birinci noktanın "i" indisli özelliğinin değeri.

x_{yk}: İkinci noktanın "k" indisli özelliğinin değeri.

Karar Ağacı (Decision Tree) algoritması, veri kümesindeki özelliklerin değerlerine göre yapılan basit kararlarla verileri sınıflandıran ve regresyon yapabilen bir makine öğrenmesi algoritmasıdır. Karar ağaçları, hem sınıflandırma (classification) hem de regresyon (regression) problemleri için kullanılabilir. Karar ağacı karar düğümleri, dallar ve yapraklardan oluşmaktadır (Han ve Kamber, 2006).

Naive Bayes algoritması, makine öğrenmesi ve istatistiksel sınıflandırma için kullanılan olasılık temelli bir algoritmadır. Sınıflandırma problemlerini çözmek için yaygın olarak kullanılır ve genellikle doğal dil işlemede, spam filtrelemede, duygu analizinde, medikal teşhislerde ve diğer sınıflandırma görevlerinde etkili sonuçlar verir. Algoritma Bayes teoremini kullanır ve sınıf değişkeninin değeri dikkate alındığında tüm değişkenlerin bağımsız olduğunu varsayar (Dimitoglou ve ark., 2012). Bayes teoreminde kullanılan matematiksel formül Denklem 2'de açıklanmıştır.

$$P(A \setminus B) = (P(A \setminus B) * P(A)) / P(B) \quad (2)$$

Denklemdaki sembollerin açıklamaları şu şekildedir:

P(A|B): A olayının, B olayı gerçekleştiğindeki koşullu olasılığı.

P(B|A): B olayının, A olayı gerçekleştiğindeki koşullu olasılığı.

P(A): A olayının önceden bilinen marjinal olasılığı.

P(B): B olayının önceden bilinen marjinal olasılığı.

Rastgele Orman (Random Forest) algoritması, makine öğrenmesi alanında yaygın olarak kullanılan ve sınıflandırma, regresyon ve kümeleme gibi çeşitli görevler için etkili sonuçlar veren bir ensemble

(topluluk) öğrenme yöntemidir. Rastgele Orman, birçok karar ağacının (decision tree) bir araya gelmesiyle oluşturulur ve bu ağaçların tahminlerinin ortalaması veya en çok oy alan sınıf veya değerle sonuç alınarak sınıflandırma veya regresyon yapar. Rastgele Ormanlar, tahmin ortalamasının temel gerçeğe (sınıflandırma) veya gerçek değere (regresyon) yaklaşacağı şekilde, her bir karar ağacının sahip olduğu tahmin varyansının üstesinden gelebilir (Liang ve ark., 2020).

Destek Vektör Makinesi (Support Vector Machine - SVM), makine öğrenmesi alanında sınıflandırma ve regresyon problemlerini çözmek için kullanılan güçlü bir algoritmadır. SVM, veri noktalarını uzayda sınıflara ayıran bir hiper düzlem (hyperplane) bulmaya odaklanır ve bu düzlem, sınıflar arasındaki en büyük marjı (maksimum marj) elde etmek için optimize edilir. Sınıflar arasındaki marjın maksimize edilmesi, SVM'nin yüksek genelleme performansına ve iyi bir ayırt edici güce sahip olmasına olanak tanır.

Lojistik Regresyon (Logistic Regression), istatistik ve makine öğrenmesi alanlarında sınıflandırma problemlerini çözmek için kullanılan bir istatistiksel modeldir. Temel olarak, bir girdi verisi alır ve bu veriyi bir veya daha fazla kategoriye (sınıfa) atar. Bu nedenle, genellikle iki sınıflı (binary) sınıflandırma problemleri için kullanılır, ancak çoklu sınıflı sınıflandırmaya da genişletilebilir.

Linear Diskriminant Analizi (Linear Discriminant Analysis - LDA), istatistik ve makine öğrenmesi alanlarında, sınıflandırma problemlerini çözmek için kullanılan bir örüntü tanıma ve boyut azaltma yöntemidir. LDA, çok boyutlu veri setlerini daha düşük boyutlu bir alt uzaya dönüştürerek sınıflar arasındaki farklılıkları vurgulayarak sınıflandırma yapar.

2.2. Performans Değerlendirme

Çalışma kapsamında performans değerlendirilmesinde karmaşıklık matrisi ve çapraz doğrulama uygulanmıştır. Sınıflandırma işlemlerinin performansını ölçmek için Tablo 3'te gösterilen karmaşıklık matrisi (confusion matrix) kullanılmıştır. Karmaşıklık matrisi, bir algoritmanın performansını değerlendirmek için kullanılan bir metrik tablodur. Özellikle sınıflandırma problemlerinde, algoritmanın gerçek değerler ile tahmin ettiği değerler arasındaki farkı anlamak ve modelin performansını anlamak için kullanılmaktadır. Karmaşıklık matrisindeki metrikler arasında doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1-puanı (F1-score) gibi önemli değerler bulunmaktadır. Bu değerler, algoritmanın hangi durumlarda ne kadar başarılı olduğunu ve hangi sınıfları daha iyi veya daha kötü sınıflandırdığını anlamaya yardımcı olur.

Tablo 3. Karmaşıklık matrisi

| TAHMİN | GERÇEK | |
|---------|---------|---------|
| | Pozitif | Negatif |
| | Pozitif | DP |
| Negatif | YN | DN |

Karmaşıklık matrisi, 4 farklı değeri gösteren bir tablodur:

Doğru Pozitif (DP): Gerçek sınıfı pozitif olarak doğru bir şekilde sınıflandırılan örnek sayısı.

Yanlış Pozitif (YP): Gerçek sınıfı negatifken yanlış bir şekilde pozitif olarak sınıflandırılan örnek sayısı.

Doğru Negatif (DN): Gerçek sınıfı negatif olarak doğru bir şekilde sınıflandırılan örnek sayısı.

Yanlış Negatif (YN): Gerçek sınıfı pozitifken yanlış bir şekilde negatif olarak sınıflandırılan örnek sayısı.

Doğruluk (Accuracy) : Bir sınıflandırma modelinin doğru tahminlerinin toplam veri sayısına oranını ifade eden bir performans metriğidir. Denklem 3'te verilen matematiksel ifade ile hesaplanmaktadır.

$$\text{Doğruluk} = \frac{DP + DN}{DP + YP + DN + YN} \quad (3)$$

Kesinlik (Precision): Sınıflandırma modelinin doğruluk oranını ifade eden metriktir. Denklem 4'te verilen matematiksel ifade ile hesaplanmaktadır.

$$\text{Kesinlik} = \frac{DP}{DP + YP} \quad (4)$$

Duyarlılık (Recall): Sınıflandırma modelinin gerçek pozitiflerin tespit edilme oranını veren metriktir. Denklem 5'te verilen matematiksel ifade ile hesaplanmaktadır.

$$\text{Duyarlılık} = \frac{DP}{DP + YN} \quad (5)$$

F1-puanı (F1-score): Sınıflandırma modelinin doğruluğunu ve hassasiyetini aynı anda değerlendiren bir performans metriğidir. F1 puanı kesinlik ve duyarlılık arasındaki dengeyi ifade etmektedir (Canibey ve Sevli, 2022). Denklem 6'da verilen matematiksel ifade ile hesaplanmaktadır.

$$\text{F1 - Puanı} = 2 * \frac{\text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \quad (6)$$

Çapraz doğrulama ise; modelin güvenilir bir şekilde değerlendirilmesini sağlamak için yaygın olarak kullanılan bir tekniktir. Çapraz doğrulama yönteminde veri kümesi parçalara ayrılarak, bu veriler farklı eğitim ve farklı test veri kümelerini oluşturmaktadır. Bu çalışmada da 5 kat çapraz doğrulama uygulanmıştır.

3.3. Uygulama Adımları

Toplam 1000 adet örnekten oluşan veri seti üzerinde 8 öznitelik incelenerek, veri dağılımları analiz edilmiş ve öznitelikler sayısal değere dönüştürülerek deneysel işleme hazır hale getirilmiştir. Her öğrenci için math score, reading score ve writing score özniteliklerinin ortalaması alınarak 40 ve üzeri

not alması durumu geçer not olarak kabul edilmiştir. Öğrencinin geçip kalma durumu için result öz niteliği oluşturulmuş ve yeni oluşturulan tabloya eklenmiştir. Oluşturulan yeni tablo Tablo 4'te gösterilmektedir.

Tablo 4. Tahminleme işlemi için kullanılacak öz nitelikler

| Öznitelik Adı | Açıklama |
|-----------------------------|-----------------------------|
| gender | [erkek/ kadın] |
| race/ethnicity | [A'dan E'ye gruplar] |
| parental level of education | [ebeveyn eğitimi] |
| lunch | [okula yerleşim türü] |
| test preparation course | [test hazırlık kursu] |
| result | [başarı durumu geçti/kaldı] |

Öğrencilere ait olan başarı durumu (result) tahminlenmeye çalışılmıştır. Veri seti üzerinde KNN, KA, NB, RO, DVM, LR, LDA sınıflandırıcıları kullanılmıştır. Bütün sınıflandırma işlemlerinde 5 kat çapraz doğrulama uygulanmıştır. Sınıflandırıcıların performansları doğruluk, duyarlılık, kesinlik ve F1-puanı metrikleri açısından incelenmiştir. Sınıflandırıcılar için kullanılan parametreler ve değerleri Tablo 5'te gösterilmektedir.

Tablo 5. Sınıflandırıcılar için kullanılan parametreler ve değerleri

| Sınıflandırıcı | Kullanılan parametre ve değerler |
|----------------|----------------------------------|
| DVM | kernel='rbf', C=2 |
| LR | max_iter=250 |
| KNN | n_neighbors=13 |
| RO | n_estimators =100 |
| NB | Default |
| KA | Default |
| LDA | Default |

4. Bulgular ve Tartışma

Öğrencilerin sosyo-demografik bilgileri ile matematik, okuma ve yazma derslerine ait notları bulunduran veri seti K-en Yakın Komşu (KNN), Karar Ağacı (KA), Naive Bayes (NB), Rastgele Orman (RO), Destek Vektör Makinesi (DVM), Lojistik Regresyon (LR), Lineer Diskriminant Analizi (LDA) modelleri ile tek tek sınıflandırılmıştır. Her sınıflandırma işlemi için 5 kat çapraz doğrulama yapılmıştır. Modellerin doğruluk, duyarlılık, kesinlik ve F1-puanı değerleri belirlenmiş ve Tablo 6'da gösterilmiştir.

Tablo 6. Modellerin sınıflandırma başarıları

| Sınıflandırıcı | Doğruluk | Kesinlik | Duyarlılık | F1-puanı |
|----------------|----------|----------|------------|----------|
| KNN | 0,61 | 0,59 | 0,53 | 0,50 |
| KA | 0,59 | 0,62 | 0,58 | 0,58 |
| NB | 0,66 | 0,64 | 0,62 | 0,62 |
| RO | 0,61 | 0,62 | 0,59 | 0,59 |
| DVM | 0,67 | 0,64 | 0,61 | 0,61 |
| LR | 0,65 | 0,65 | 0,63 | 0,63 |
| LDA | 0,65 | 0,65 | 0,62 | 0,62 |

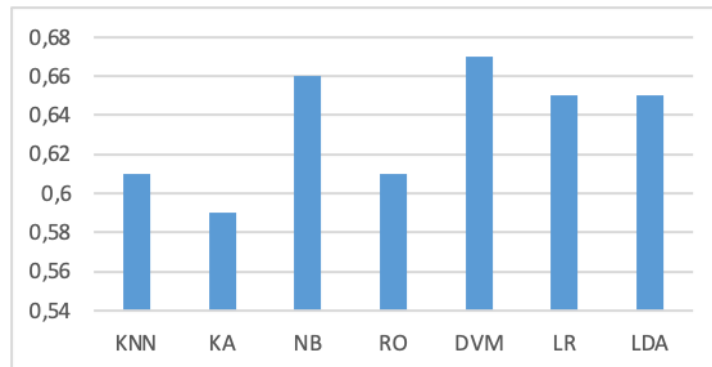
Tablo 6’da verilen sonuçlar incelendiğinde en yüksek doğruluk %67 ile DVM modelinden elde edilmiştir. KA, %59 ile en düşük doğruluğu sağlamıştır. Ölçülen doğruluk, modellerin genel başarısını temsil etmektedir. Modellerin genel başarı sıralaması şu şekilde yapılabilir: DVM>NB>LR, LDA>RO, KNN>KA.

Modelin doğru bir şekilde tahmin ettiği örneklerin oranını ifade eden kesinlik değeri LR ve LDA’da %65 ile en yüksek puana sahiptir. NB ve DVM %64 ile ikinci sırada bulunmaktadır. KA ve RO %62 ile üçüncü sırada yer alırken, KNN modeli, dördüncü sırada yer almaktadır.

Hassasiyet veya gerçek pozitif oranı olarak adlandırılan Duyarlılık değeri bakımından %63 ile LR modeline aittir. Bu modeli %62 ile NB ve LDA, %61 ile DVM, %59 ile RO, %58 ile KA ve %53 ile KNN takip etmektedir.

Modelin doğru pozitifleri ve yanlış pozitifleri ele alırken ki başarısını gösteren F1-puanı açısından en yüksek başarı %63 ile LR modeline aittir. Bu modeli %62 ile LDA ve NB, %61 ile DVM, %59 ile RO, %58 ile KA ve %47 ile KNN takip etmektedir.

Çalışma kapsamında kullanılan modellerin doğruluk başarı durumları Şekil 5’te gösterilmektedir. Sınıflandırma sonucunda modellerin doğruluk metriği bakımından en başarılı model DVM olarak belirlenmiştir.



Şekil 5. Modellerin Doğruluk (Accuracy) başarı değerleri

Veri setindeki deęişkenler arasındaki ölçek farklarını gidererek, algoritmaların daha iyi performans göstermesine yardımcı olabilmek amacıyla normalizasyon teknięi kullanılarak modellerin doęruluk (Accuracy) başarı deęerleri karşılaştırılmalı olarak Tablo 7’de gösterilmiştir.

Tablo 7. Modellerin normalizasyon öncesi ve sonrası doęruluk başarı deęerleri

| Sınıflandırıcı | Normalizasyon Öncesi | Normalizasyon Sonrası |
|----------------|----------------------|-----------------------|
| KNN | 0,61 | 0,61 |
| KA | 0,59 | 0,59 |
| NB | 0,66 | 0,66 |
| RO | 0,61 | 0,63 |
| DVM | 0,67 | 0,68 |
| LR | 0,65 | 0,65 |
| LDA | 0,65 | 0,65 |

Tablo 7’de verilen sonuçlar incelendiğinde normalizasyon öncesi ve sonrası doęruluk başarı deęerleri incelendiğinde RO’nun doęruluk başarı deęeri %61’den %63’e yükselirken, DVM’nin doęruluk başarı deęeri de %67’den %68’e yükselmiştir.

5. Sonuç ve Öneriler

Makine öğrenmesinin kullanımı birçok alanda çözümler sunmaktadır. Özellikle büyük veri miktarlarının arttığı, karmaşık problemlerin ele alındığı ve daha akıllı sistemlerin ihtiyaç duyulduğu günümüz dünyasında, makine öğrenmesinin farklı disiplinlerde kullanımı ve bilimsel araştırmalara konu olması bakımından daha da artmaktadır. Öğrencilere ait büyük veri yığınlarının bulunması makine öğrenmesinin eğitim alanında da kullanımını mümkün kılmaktadır. Öğrencilerin performansının tahmin edilebilmesi ve öğrenme hedeflerine ulaşmadaki potansiyel başarısızlığa erken önlem alınabilmesi için kullanılacak makine öğrenmesi eğitimin tüm paydaşlarına avantajlar sağlayacaktır. Bu çalışmada öğrencilerin sosyo-demografik bilgileri ile matematik, okuma ve yazma derslerine ait notları bulduran 1000 adet örnekten oluşan veri seti yedi ayrı makine öğrenmesi ile sınıflandırılmış ve farklı metrikler ile incelenmiştir.

K-en yakın komşu (KNN), Karar ağacı (KA), Naive Bayes (NB), Rastgele orman (RO), Destek vektör makinesi (DVM), Lojistik regresyon (LR), Lineer Diskriminant Analizi (LDA) modelleri ile yapılan Sınıflandırma sonucunda modellerin doęruluk metrięi bakımından en başarılı model en başarılı model DVM olmuştur. Normalizasyon teknięi ile modellerin doęruluk başarı durumları incelendiğinde ise RO ve DVM’nin deęerlerinde az da olsa yükselme olmuştur. Öğrencilere ait daha fazla özneliğin bulunması, farklı sınıflama çalışmalarının yapılması ve incelenmesinin gelecekte yapılacak çalışmalarda yer alması önerilebilir.

Araştırmacıların Katkı Oranı Beyan Özeti

Yazar, makaleye %100 oranında katkı sağlamış olduğunu beyan eder.

Çıkar Çatışması Beyanı

Makale yazarı herhangi bir çıkar çatışması olmadığını beyan eder.

Kaynakça

- Alpaydin E. Introduction to machine learning, fourth edition. MIT Press. 2020.
- Badal YT., Sungkur RK. Predictive modelling and analytics of students' grades using machine learning algorithms. *Educ Inf Technol* 2023; 28, 3027–3057.
- Canibey ST., Seveli O. Bireylerin gelir dağılım seviyelerinin makine öğrenmesi teknikleri ile belirlenmesi. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi* 2022; 5(2): 753-766.
- Chen Y., Zhai L. A comparative study on student performance prediction using machine learning. *Educ Inf Technol* 2023; 28, 12039–12057.
- Çakıt E., Dağdeviren M. Predicting the percentage of student placement: A comparative study of machine learning algorithms. *Education and Information Technologies* 2022; 27(1): 997-1022.
- Chung JY., Lee S. Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review* 2019; 96, 346-353.
- Dimitoglou G., Adams JA., Jim CM. Comparison of the C4.5 and a naive bayes classifier for the prediction of lung cancer survivability (arXiv:1206.1121). *arXiv*. 2012; <https://doi.org/10.48550/arXiv.1206.1121>.
- Djulovic A., Li D. Towards freshman retention prediction: A comparative study. *International Journal of Information and Education Technology* 2013; 3(5): 494-500.
- Guleria P., Sood M. Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies* 2023; 28(1): 1081-1116.
- Han J., Kamber M. *Data mining: Concepts and techniques (Second edition) (2nd bs)*. Kaufmann Publisher, 2006.
- Iam-On N., Boongoen T. Generating descriptive model for student dropout: A review of clustering approach. *Human-centric Computing and Information Sciences* 2017; 7(1): 1-24.
- Iatrellis O., Savvas IK., Fitsilis P., Gerogiannis VC. A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies* 2021; 26, 69-88.
- Kilinç D., Borandağ E., Yücalar F., Tunali V., Şimşek M., Özçift A. KNN algoritması ve r dili ile metin madenciliği kullanılarak bilimsel makale tasnifi. *Marmara Fen Bilimleri Dergisi* 2016; 28(3): Article 3. <https://doi.org/10.7240/mufbed.69674>.
- Koza JR., Bennett FH., Andre D., Keane MA. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. İçinde J. S. Gero & F. Sudweeks (Ed.), *Artificial Intelligence in Design '96, 1996*; 151-170. Springer Netherlands. https://doi.org/10.1007/978-94-009-0279-4_9

- Liang YC., Maimury Y., Chen AHL., Juarez JRC. Machine learning-based prediction of air quality. *Applied Sciences* 2020; 10(24): Article 24. <https://doi.org/10.3390/app10249151>.
- Lu O., Huang A., Huang J., Lin A., Ogata H., Yang SJH. Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society* 2018; 21(2): 220-232.
- Machine A. The artificial use of the term "Intelligence". *Medium*. 2017. <https://becominghuman.ai/the-artificial-use-of-the-term-intelligence-c82878447dfd>.
- Sevli O., Başer VG. Prediction of student academic success in distance education using machine learning. Conference: Future-Learning 2020 - 8th International Conference on Future Learning and Informatics: Data Revolution, Muğla, Turkey.
- Tsai SC., Chen CH., Shiao YT., Ciou JS., Wu TN. Precision education with statistical learning and deep learning: A Case study in Taiwan. *International Journal of Educational Technology in Higher Education* 2020; 17(1): 1-13.
- Turing A. Computing machinery and intelligence. *Mind* 1950; 49(236): 433-460.
- Xie H., Chu HC., Hwang GJ., Wang CC. Trends and development in technology-enhanced adaptive/personalized learning: A Systematic review of journal publications from 2007 to 2017. *Computers & Education* 2019; 140, 103599. doi:10.1016/j.compedu.2019.103599.
- Web 1: Students performance in exams. <https://www.kaggle.com/datasets/spscientist/students-performance-in-exams>. Erişim tarihi, Kasım,2023.