İSTANBUL TİCARET ÜNİVERSİTESİ FEN BİLİMLERİ DERGİSİ

İstanbul Commerce University Journal of Science

http://dergipark.org.tr/ticaretfbd

Araştırma Makalesi / Research Article

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE EĞİTİM BAŞARISINA ETKİ EDEN FAKTÖRLERİN MODELLENMESİ*

MODELLING OF THE EFFECTS ON EDUCATIONAL SUCCESS BY MACHINE LEARNING ALGORITHMS

Zeynep BAKAN[1] **Filiz KANBAY[2]**

https://doi.org/10.55071/ticaretfbd.1442084

Sorumlu Yazar / Corresponding Author
zeyno.cuk@hotmail.com.

Geliş Tarihi / Received
27.02.2024

Kabul Tarihi / Accepted
06.05.2024

Öz

Sağlık, medya, bankacılık ve finans alanında sınıflandırma, kümeleme ve tahmin amacıyla kullanılan makine öğrenmesi günümüzde eğitim alanında da kullanılmaktadır. Bu çalışmada eğitim öğretim kurumlarının belirleyecekleri stratejilerde veya alacakları önlemlerde yol gösterici olması ve hatta daha büyük ana kütle, daha farklı okul türü ya da farklı kademelerde, farklı sektörlerde uygulanarak sonuçların genelleştirilmesine fayda sağlaması amacıyla makine öğrenmesi yöntemlerinden K-en yakın komşu, naive bayes, rastgele orman, destek vektör makineleri, karar ağaçları, boosting makine öğrenmesi sınıflandırma algoritmaları ile kurulan matematiksel modellemeler ile öğrencilerin akademik başarılarını etkileyen faktörler araştırılmıştır. Kurulan matematiksel modelin başarısına etki eden hiperparametreler ızgara taraması yöntemi ile belirlenerek maksimum model başarısı sağlanmıştır. Matematiksel modellemelerde akademik başarı ölçütü çıktı olarak belirlenerek; kurulan matematiksel modellerde çıktı ve girdi sayılarına ait model başarılarının değişimi incelenmiş; çıktıların ve girdilerin sayısının çeşitli yöntemlerle (denetimli ve denetimsiz yöntemlerle) azaltılması işlemlerinin matematiksel model başarısına etkileri gözlenmiştir. Sonuç olarak, en yüksek model başarılarının iki sınıf etiketli veri setine ait olduğu görülmüştür. K-en yakın komşu, naive bayes, rastgele orman, destek vektör makineleri, karar ağaçları, boosting model başarıları sırasıyla 0,62, 0,61, 0,96, 0,72, 0,86, 0,79 olarak elde edilmiştir.

Anahtar Kelimeler: Destek vektör makineleri, eğitim-öğretim, makine öğrenmesi, k-en yakın komşu, naive bayes, rastgele orman.

Abstract

Machine learning, which is used for classification, clustering and prediction in the fields of health, media, banking and finance, is also used in the field of education today. In this study, by using the mathematical models established with machine learning classification methods such as K-nearest neighbour, naive bayes, random forest, support vector machines, decision trees and boosting; the factors affecting students' academic success were investigated to guide educational institution the strategies , to determine the measures to be taken, and even to benefit the generalization of the results by applying them to a larger population, different types of schools or at different levels, in different sectors. Maximum model success was achieved by determining the hyperparameters that affected the success of the established mathematical model by the grid scanning method. In mathematical modelling, the academic success criterion is determined as the output; The changes in the model success of the output and input numbers in the established mathematical models were examined; The effects of reducing the number of outputs and inputs by various methods (supervised and unsupervised methods) on the success of the mathematical model have been observed. Finally the best accuracy scores were obtained from the data set with two class labels. The accuracy scores of the algorithms (K-nearest neighbour, naive bayes, random forest, support vector machines, decision trees and boosting) respectively were 0,62, 0,61, 0,96, 0,72, 0,86, 0,79.

Keywords: Education-training, machine learning, k-nearest neighbours, naive bayes, random forest, support vector machines.

**Bu yayın Zeynep BAKAN isimli öğrencinin Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, Matematik Programındaki Yüksek Lisans tezinden üretilmiştir.*

[1]Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Matematik Anabilim Dalı, İstanbul, Türkiye.
zeyno.cuk@hotmail.com, Orcid.org/0000-0001-6742-8376.

[2]Yıldız Teknik Üniversitesi, Fen Edebiyat Fakültesi, Matematik Bölümü, İstanbul, Türkiye.
fkanbay@yildiz.edu.tr, Orcid.org/0000-0003-3889-6351.

1. GİRİŞ

Sınıflandırma, kümeleme ve tahmin amacıyla günümüzde sağlık, medya, bankacılık ve finans alanında kullanılan makine öğrenmesi eğitim alanında da kullanılmaktadır. Eğitimde makine öğrenmesi yöntemleri ile eğitim alanında karşılaşılan büyük verinin analizi yapılarak eğitim hedeflerinin tahmini, eğitim alanında erken önlemlerin tespiti, öğrencilerin öğrenme süreçlerinin ve içeriklerinin belirlenmesi, ölçme değerlendirme sürecindeki analiz gibi birçok önemli konu ele alınabilir (Tosunoğlu ve ark.,2021). E Tosunoğlu ve ark. 2015-2020 yılları arasında yapılmış 5 yıllık süreci kapsayan makine öğrenmesi ile ilgili 201 makalenin %35,3 oranıyla en fazla lisans öğrencilerine ait olduğunu; %22,5 karar ağaçları, %17,8 destek vektör makineleri ve %14,2 navie bayes sınıflandırma yöntemlerinin kullanıldığını belirtmiştir. P. Kaur ve ark. eğitimde veri madenciliğinde öğrencilerin performansı, öğrenme ve akılda tutma gibi kabiliyetlerini iyileştirmek için WEKA yazılımı kullanarak Çok Katmanlı Algılayıcı, Naïve Bayes, SMO, J48, REPTree sınıflandırma algoritmaları ile yavaş öğrenen öğrencilerin tahminini yapmıştır (Kaur ve ark.,2015). M. Sweeny ve ark., yüksek öğretim kurumlarında bulunan öğrencilerin % 50 başarı düzeyine etki eden faktörlerin tahmini için geliştirdikleri sistem ile Faktoring Makineleri (FM), Rastgele Orman (RM) ve Kişileştirilmiş Doğrusal Regresyon algoritmaları kullanarak eğitmen özellikleri ve öğrenci başarısı arasında güçlü bir ilişki olduğunu saptamıştır (Sweeny ve ark.,2016). E. Fernandes ve ark. eğitimin daha verimli hale getirilmesi ve yıl sonu başarısızlığının en aza indirgenmesi amacıyla model güçlendirici bir algoritma olan Gradient Boosting'e dayalı bir sınıflandırma modeli ile okul öncesi ve sonrası öğrenci akademik verilerini kullanarak 2015-2016 yıllarında lise öğrencilerinin sene sonundaki akademik başarılarını tahmin etmiş, devamsızlık ve notlar arasında bir ilişki bulmuştur (Fernandes ve ark.,2019). M. Yıldız ve C. Börekçi 9. sınıf öğrencilerinden toplanan eğitsel durumları ve aileleri ile ilgili verileri kullanarak sınav sonucunu denetimli sınıflandırma algoritmalarından Neural Network algoritması ile %98,6 başarı oranı ile öğrencilerin derslere yönelik tutum ve yargıları ile çalışma düzenlerinin sınıflandırmayı etkilediğini ve bu değişkenler üzerinde durulduğu taktirde öğrencilerin akademik başarılarının olumlu yönde gelişebileceği çıkarımına varmıştır (Yıldız & Börekçi,2020). C. Marquez-Vera ve ark. Meksika'da 419 lise öğrencisinden toplanan verilerle öğrencilerin okulu bırakmasının erken tahmini için WEKA yazılımı aracılığıyla ICMR2 algoritmasını kullanmış ve çalışma sonucunda elde edilen veriler ile okulu erken bırakma riski taşıyan öğrencileri belirleyen EWS sistemini geliştirmiştir (Marquez ve ark.,2015). C. Romero, P Espejo ve ark. üniversite öğrencilerinin bir derse ait final notlarını karar ağaçları, sinir ağları sınıflandırma algoritmaları, istatistiksel yöntemler ve bulanık kural tümavarım yöntemleri ile tahmin ederek, öğretim üyelerinin karar verme süreçlerinde kullanabileceği moodle madenciliği aracı geliştirmiştir (Romero ve ark.,2013). S. Kayalı ve S. Buyrukoğlu Portekiz'de iki farklı ortaokul öğrencisine ait günlük teknolojik cihaz kullanma saati, okula gitmeden kahvaltı yapma durumu, stress gibi önemli bilgilerin olduğu 33 öznitelik ve 944 örneklemden oluşan veri setine Karar Ağacı, Rastgele Orman ve Destek Vektör Makinası algoritmaları uygulamış ve öğrencilerin başarısını tahmin etmede en etkili algoritmanın Rastgele orman algoritması olduğu sonucuna varmıştır (Kayalı & Buyrukoğlu, 2022). M. Gök 24 soruluk bir anket ile elde edilen veri setini en iyi temsil eden öznitelikleri KÖAK öznitelik seçim yöntemi kullanarak Anne-Baba öğrenim durumu, gelir, kardeş sayısı, uyku, haftalık ders çalışma süresi ve ailenin ders konusundaki tutumu olarak sadeleştirmiş ve 6. , 7., 8. sınıf öğrencilerinin sene sonu Türkçe ve matematik notlarını Lojistik, Naive Bayes, K-NN, Doğrusal DVM, RTF DVM ve Rastgele Orman algoritmaları ile inceleyerek Lojistik algoritması en başarılı sonuç verdiğini tespit etmiştir. (Gök,2017). B. Abbasoğlu Yalova ilindeki dört resmi ortaokulundaki 1395 ortaokul öğrencilerinin 2019-2020 eğitim öğretim yılındaki sosyo-ekonomik durumları ve demografik özelliklerini içeren 27 bağımsız değişkenden en önemli değişkenlerin öğrenci yaşı, devamsızlık durumu, ebeveynlerin birlikte/ayrı olup olmama durumu, gelir durumu, anne/baba eğitim durumu, kendine ait oda ve gelir durumu olarak belirlemiş ve bunların yıl sonu ortalamalarına etkilerini en yüksek tahmin başarısına sahip olan lojistik algoritma ile analiz etmiştir (Abbasoğlu, 2020). Nedeva ve Pehlivanova, Stara Zagora'da Trakia Universitesinde yürütülen çalışmanın sonuçlarını

kullanarak öğrenci akademik başarısına etki eden faktörleri Weka ile incelemiştir (Nedeva & Pehlivanova, 2021). N. Yılmaz ve B. Şekeroğlu Yakın Doğu Üniversitesinin Eğitim ve Mühendislik Fakültelerindeki 101 öğrenciye kişisel, aile ve eğitim tercihleri ile ilgili 30 soruluk üç bölümden oluşan bir anket uygulayarak soru türlerinin öğrenci performansı üzerindeki etkilerini incelemiş ve bu amaçla Geriye Yayılımlı Sinir Ağı(BPNN), Radyal Tabanlı Fonksiyonlu Sinir Ağı (RBFNN), Karar Ağaçları(DT) ve Logistik regresyon(LOGR) algoritmalarını kullanmış ve en yüksek başarının %70-%88 doğruluk oranı ile Radyal Temelli Fonksiyonlu Sinir ağına ait olduğunu tespit etmiş; bunun yanı sıra eğitimin geleneksel eğitimden ziyade uygulamaya yönelik olması, öğrencilerin aktif katılımları ve öğrenci motivasyonunun artması için öğretim sürecine başlamadan önce ders amaçlarının açıklanması, içeriğin soyut olmaması, günlük hayatta kullanılabilir olması, konunun niteliğine göre yöntem ve teknikler uygulanması, öğrencilerin laboratuvar çalışmalarına aktif katılımı gibi konular önermiştir (Yılmaz & Şekeroğlu,2020). Bezek, IG, GR, SU, CB, Relief-F, One R Measure gibi özellik seçimi yöntemleri kullanarak bu veri setinin özniteliklerini azalan önem sırasına göre: cinsiyet, projelerin/faaliyetlerin akademik başarıya etkisi, mezuniyet sonrası beklenen not ortalaması, haftalık çalışma saatleri, alınan burs türü, okuma saatleri olarak belirlemiştir. (Bezek, 2023). Phatai ve Luangrungruang, öğrenci performans sınıflandırması için NN with Artificial Bee Colony (NNABC), NN with Harmony Search (NN-HS), NN with Teaching Learning Based Optimization (NN-TLBO), and NN with Student Psychology Based Optimization (NN-SPBO) yöntemlerini kullanmıştır (Phatai & Luangrungruang, 2023). Jabardi, veri setinin özelliklerini aile, eğitim ve öğrenci kişisel bilgilerini içeren üç kümeye ayırmış ve yaptığı çalışma sonucunda aile ve eğitime ait özelliklerin öğrenci başarısı üzerinde, öğrenci kişisel özellerinden daha çok etkiye sahip olduğunu bulmuştur (Jabardi, 2022). Hengpraprohm ve arkadaşları, eniyi özellik seçimi tekniği olarak bilgi kazanımı ve bilgi kazanım oranı olduğunu belirterek her iki teknik ile öğrencilere ait cinsiyet, son dönem not ortalaması, beklenen genel not ortalaması, okuma sıklığı, ebeveyn durumu gibi özniteliklerin bulunduğu 16 özniteliğin seçildiğini ifade etmiştir (Hengpraprohm ve ark 2022). Pallathadka ve ark. veri seti için kurulan Naive Bayes, ID3, CD4.5, SVM gibi makine öğrenmesi algoritmalarını doğruluk, hata oranı gibi parameterelere göre değerlendirmiştir (Pallathadka ve ark 2023). Şekeroğlu ve ark. ham veri üzerinde bir data ön işleme veya özellik seçimi kullanmadan makine öğrenmesi algoritmaları kullanarak çeşitli eğitim verileri üzerinde tahmin ve sınıflandırma yöntemlerini çalışmışlardır (Şekeroğlu ve ark., 2019).

Öğrencilerin akademik başarılarını etkileyen faktörlerin belirlenmesi için Makine Öğrenmesi sınıflandırma yöntemleri ile matematiksel modellemeler kurulmasını amaçlayan bu çalışmada elde edilen matematiksel modellerin model doğruluk ölçütlerini yükseltecek parametreler araştırılmış, öğrenci başarısını etkileyen öznitelikler sıralanmıştır. İkinci bölümde makine öğrenmesi tanımı, algoritmaları, model değerlendirme ölçütleri ve boyut azaltma yöntemleri gibi temel kavramlara kısaca değinilmiştir. Üçüncü bölümde veri seti ve uygulanacak işlemlere dair akış diyagramı verilmiştir. Dördüncü bölüm veri seti üzerinde matematiksel modellemelerin geliştirildiği ve değerlendirildiği uygulama aşamasını açıklamaktadır. Çalışmaya ait son bölüm öğrenci başarısını etkileyen faktörleri belirlemek için kurulan matematiksel modellemeler ve bu modellerin başarısını arttıracak belirlenmiş parametrelerin değerlendirildiği ve görselleştirildiği grafik ve tabloları içermektedir.

2. MAKİNE ÖĞRENMESİ VE ALGORİTMALARI

2.1. Makine Öğrenmesi

Veri madenciliği ile yakından ilgili olan makine öğrenmesi çok sayıda veriyi anlamlı bilgiye bilinçli eyleme dönüştürecek bilgisayar algoritmalarının geliştirilmesiyle ilgilenen bir alan olarak tanımlanabilir (Lantz,2013). Makine öğrenmesi verilerden anlamlı sonuçlar çıkarıp bu sonuçları

geliştirip iyileştirir ve modeller oluşturur, veri madenciliği eldeki verilerden çıkan sonuçları yorumlar (Bilgin,2018).

2.1.1. Gözetimli öğrenme

Danışmanlı öğrenme olarak da bilinen gözetimli öğrenme türünde, sisteme girdilerle birlikte sonuçlar da verilerek etiketli veriler üzerinden girdi ve çıktı arasındaki etkileşimi ortaya çıkararak algoritmanın etiketsiz girdiler için yeni sonuçları üretmesi beklenir.

Gözetimli öğrenme sınıflandırma ve regresyon olmak üzere ikiye ayrılır:

1. Sınıflandırma: Örüntü tanıma olarak da isimlendirilen bu makine öğrenmesi türünde, çıktıların sonuçlarının kategorik verilere ayrılması ile algoritmanın mevcut verileri sınıflandırması beklenir (Bozinovski,1981).
2. Regresyon: Çıktıları sürekli sayısal değerlerden oluşan ve değişkenler arasında bir ilişki kuran modellemeler ile tahminler yapan bu sınıflandırma türü sürekli değerler olarak adlandırılan sayısal değerli tahminler bulmak için kullanılır (Amasyalı,2008).

2.1.2. Gözetimsiz öğrenme

Danışmansız öğrenme olarak da bilinen bu öğrenme türünde sisteme sadece girdiler başka bir ifadeyle etiketsiz veriler verilerek, algoritmanın bu girdilere dair sonuçlar (kümeleme gibi) üretmesi beklenir.

2.1.3. Pekiştirmeli öğrenme

Pekiştirmeli takviyeli öğrenme olarak da bilinen bu öğrenme türünde makinelere ödül ceza sistemine dayanan geri dönüt verilir. Bir takım deneme yanılma eyleminden sonra makine daha fazla ödülü almak için en iyi cevabı öğrenir. Pekiştirmeli öğrenmede bir öğretici ile makineye ne yapması gerektiğini önceden öğreten gözetimli öğrenmeden farklı olarak, eleştirmen ile makineye ne kadar iyi yaptığı konusunda bilgi verilir. Böylece makine neden-sonuç ilişkisini kendi kurar, eleştirmen önceden bilgi vermez. Bu öğrenme türüne örnek olarak yapay zeka robotları veya oyun oynama vb. verilebilir (Alpaydın, 2010).

2.2. Sınıflandırma Algoritmaları

2.2.1. K-En yakın komşu algoritması

Denetimli bir algoritma türü olan bu algoritma sınıfı belirlenmek istenen etiketsiz verinin etiketini genellikle 1, 3, 5 … gibi tek sayılardan tercih edilen k parametresi ve Tablo 1'de verilen Manhattan, Öklid, Minkowski metrikleri ile tayin edilmiş en yakın k tane etiketli komşularının çoğunluk etiket sayısına göre belirler (Xu ve ark., 2013), Nayak ve ark (2022).

Tablo 1. En Çok Kullanılan Uzaklık Ölçütleri

Manhattan	**Öklid**	**Minkowski**
$d = \sum_{i=1}^{k} \lvert x_i - y_i \rvert$	$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$	$d = (\sum_{i=1}^{k} (\lvert x_i - y_i \rvert)^p)^{1/p}$

2.2.2. Naive Bayes algoritması

Çok sayıda girdi verisinin olduğu problemlerde olasılığı tahmin etmede başarılı olan Naive Bayes algoritması Thomas Bayes'in teoremi üzerine temellenmiş koşullu olasılık hesaplamasına dayanan tahmin edici bir sınıflandırma algoritmasıdır. Etiketsiz verilerin tüm sınıflar üzerindeki koşullu olasılıklarını hesaplayarak sınıfını belirlemeye yönelik çalışan bir istatiksel gözetimli sınıflandırma türüdür (Lantz, 2013).

Naive Bayes algoritması bağımlı olayların olasılık hesaplanmasında X ve Y bağımsız iki olay olmak üzere, Y olayı için X olayının olasılığını hesaplayan $P(X|Y) = \frac{P(Y|X).P(X)}{P(Y)}$ denklemi ile bilinen koşullu olasılık formülünü kullanır. Buna göre bilinen $x_1, x_2, \dots, x_n$'ler için y olayının olasılığı $P(y|x_1, x_2, \dots, x_n) = \frac{P(x_1, x_2, \dots, x_n|y).P(y)}{P(x_1, x_2, \dots, x_n)}$ şeklindedir. Bilinen y değerine göre ayrı ayrı $x_1, x_2, \dots, x_n$'lerin olasılıkları hesaplanarak $P(y|x_1, x_2, \dots, x_n) = \frac{P(y).\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \dots, x_n)}$ formülüyle ifade edilir. Burada $P(x_1, x_2, \dots, x_n)$ sabit olduğu için hesaplamada $P(y).\prod_{i=1}^{n} P(x_i|y)$ çarpımına bakılır ve elde edilen maksimum olasılık değerine göre y nin tahmin sınıfı bulunur (Zhang, 2004).

2.2.3. Karar ağaçları

Kök düğümler, ara düğümler, dallar ve yapraklardan oluşan bir denetimli öğrenme türü olan karar ağaçları benzer grupları küçük ve daha küçük alt kümelere bölen sınıflandırma problemlerinde kullanılan basit ve hızlı bir algoritma türüdür. Akış şemasına benzeyen ve hiyerarşik yapıda olan bu algoritma ağaç şeklinde bir modele benzer; hem sayısal hem de kategorik verileri modelleyebilir, anlaşılması ve yorumlanması daha basittir (Mitchell, 1997).

2.2.4. Rastgele orman

Birden fazla karar ağacı modeline ait çıktıların birleştirilmesiyle oluşturulmuş yüksek boyutlu bir denetimli öğrenme algoritmasıdır. Birden fazla ağacın bir araya gelerek oluşturulduğu ağaç toplulukları rastgele özellik seçimiyle birleştirildiği için rastgele orman ismini almıştır. Bu algoritma her bir karar ağacı modeline ait sınıf değeri tahminleri alınarak oylamaya tabii tutar ve sonucunda en yüksek oyu alan sınıfı seçer (Breiman, 2001; Lantz, 2013; Şahin, 2021).

2.2.5. Destek vektör makineleri

Destek vektör makinelerinin çalışma prensibi sınıfları birbirinden ayıracak maksimum marjinli hiper düzlemi seçmeye dayanır. Öncelikle verileri sınıflara ayıran en uygun hiper düzlem seçilir ve destek vektörleri belirlenir. Hiper düzleme en yakın noktalar destek vektörleri olarak adlandırılır. Ardından destek vektörlerinin arasındaki uzaklık olan marjinin en fazla olduğu hiper düzlem seçilerek sınıflandırma tamamlanır (Harrington,2012).

Marjin genişliği $2/w$ ile ifade edilmektedir w arttıkça destek vektörleri arasındaki uzaklık azalacaktır. Hiperdüzlem $w.x + b = 0$, destek vektörleri $w.x + b = \pm 1$ dir. Amaç w yi maksimize etmektir (Cortes & Vapnik,1995; Akarsu,2016).

2.2.6. Boosting

Birden fazla sınıflandırıcının kullanılması ile elde edilen sonuçların birleştirilmesine dayanan bir metot olup tahmin ediciler yani algoritmalar sıralı bir şekilde çalışır. Her model bir önceki modelin yanlış sınıflandırdığı verilere odaklanarak, her model bir önceki modelde yapılmış hataları

düzeltmeye çalışır. Böylece model sayısı arttıkça sonucun güvenirliliği artar (Han ve ark.,2011; Bilgin, 2018).

2.3. Model Değerlendirme Ölçütleri

Makine öğrenmesi algoritmaları ile belirlenen modelin ne düzeyde başarılı olduğunu ölçmek için kullanılan tekniklerdir.

Model değerlendirme ölçütleri, Tablo 2'de verilen hata matrisi (confusion matrix) tablosundan yararlanılarak hesaplanılır; Doğruluk (accuracy), kesinlik(precision), duyarlılık(recall), f-1 skor ve ROC-AUC skor en çok kullanılan model değerlendirme ölçütleridir (Sokolova & Lapalme, 2009; Han ve ark., 2011):

Tablo 2. Hata Matrisi (Confusion Matrix)

Karışım Matrisi		Tahmin Değerler	
		Pozitif	Negatif
Gerçek Değerler	Pozitif	TP	FN
	Negatif	FP	TN

TP: Gerçek değeri pozitif olup model tarafından pozitif tahmin edildiği durumdur.
TN: Gerçek değeri negatif olup model tarafından negatif tahmin edildiği durumdur.
FP: Gerçek değeri pozitif olup model tarafından negatif tahmin edildiği durumdur.
FN: Gerçek değeri negatif olup model tarafından pozitif tahmin edildiği durumdur.

$$Doğruluk(Accuracy) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Kesinlik(Precision) = \frac{TP}{TP+FP} \quad (2)$$

$$Duyarlılık(Recall) = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Skor = \frac{2*Kesinlik*Duyarlılık}{Kesinlik+Duyarlılık} \quad (4)$$

2.4. Boyut Azaltma Yöntemleri

2.4.1 Temel bileşen analizi (PCA)

PCA verilerin boyutunu azaltma, veri sıkıştırma, örüntü tanıma, öznitelik çıkarımı ve veri görselleştirmek için kullanılan bir denetimsiz öğrenme algoritmasıdır. Çok büyük veri kümelerini daha düşük boyutlu uzaylara taşıyan popüler bir yöntemdir (Tipping & Bishop,1999).

PCA ile veriyi k boyutlu bir uzaya indirgemek için denklem 5, 6 ve 7 kullanılarak normalize edilmiş verinin kovaryans matrisinden elde edilen büyükten küçüğe doğru sıralanmış özdeğerlere karşılık gelen özvektörlerin k tanesi seçilir ve W projeksiyon matrisi elde edilir. Böylece orijinal veri kümesi projeksiy on matrisini kullanarak k-boyutlu yeni uzaya dönüşmüş olur (Çalışkan & Talu, 2020).

$$\bar{X} = \frac{1}{m} \sum_{i=1}^{m} X_i \quad (5)$$

$$C = \sum_{i=1}^{n}(X - \bar{X})(X - \bar{X})^T \quad (6)$$

$$\det(\lambda I - C) = 0,\ (\lambda_k I - C) \times V_k = 0 \quad (7)$$

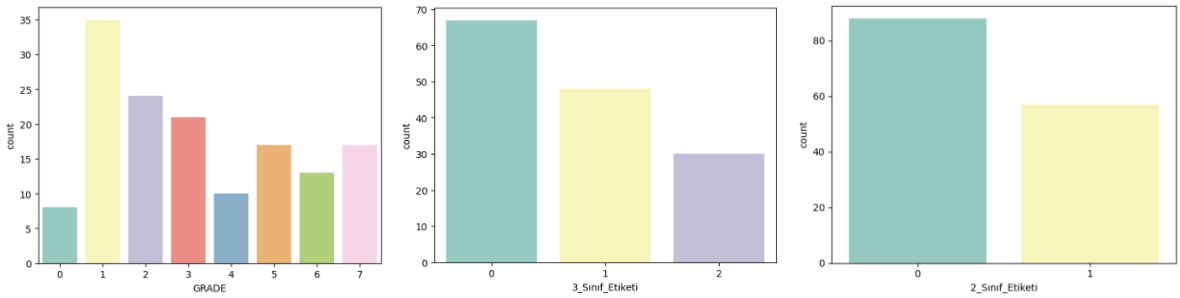
2.4.2 Komşuluk bileşen analizi (NCA)

En yakın komşu tabanlı özellik seçim yöntemi olarak bilinen NCA, en doğru sınıflandırmayı gerçekleştirmek için pozitif ağırlıklar yöntemi kullanır. Yüksek boyutlu verilerden en iyi tahmin sonucuna ulaşmak, zaman ve iş maaliyeti azaltmak için çok sayıdaki özellikten gerekli olan özellikler alt kümesini seçer ve veri boyutunu azaltır (Goldberger ve ark., 2005), (Yang ve ark., 2012).

3. VERİ SETİ VE İŞLEM AKIŞ DİYAGRAMI

3.1. Veri Seti

UCI veri tabanı sitesinde https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation internet adresi ile belirli, kayıp verisi bulunmayan 8 sınıflı, 33 kategorik öznitelikli (31 kategorik öznitelik kullanılmıştır), 145 bilgi içeren " Yükseköğretim Öğrencileri Performans Değerlendirme (Higher Education Students Performance Evaluation)" adlı veri seti ile makine öğrenmesi algoritmaları kullanılarak elde edilen matematiksel modellemelerin model başarı analizi yapılmıştır (Yılmaz & Şekeroğlu, 2019). Kurulan matematiksel modellemelerin model başarısını arttıracak hiper parametreler incelenmiştir. Ayrıca sınıflandırma başarısını maksimize etmek amacıyla öznitelik ve sınıf sayısı çeşitli makine öğrenmesi yöntemleri ile azaltılarak, öznitelik ve sınıf sayısını azaltmanın sınıflandırma başarısı üzerindeki etkileri gözlemlenmiştir. 145 öğrenciye ait 31 adet demografik değişken kategorik olduğu için pandas kütüphanesi kullanılarak tüm öznitelikler ikili olarak temsil edilmiş ve ön ad eklenmiş yeni 116 adet alt öznitelik için rastgele orman algoritması kullanılarak öğrenci başarısına en fazla etki eden faktörler belirlenmiştir. Alt öznitelik ve veri setine ait sınıf etiketi sayısı, kurulan matematiksel modelin başarısını arttırmak için azaltılmıştır. Bu amaçla Şekil 1 de gösterildiği gibi, 8 sınıf etiketi 3 ve 2 olacak şekilde dönüştürülmüş, elde edilen yeni veri setlerinin boyut indirgeme yöntemlerinden olan PCA ve NCA kullanarak azaltılmış öznitelik sayılarına göre model başarıları incelenmiştir. İlgili matematiksel modellerin başarıları, Anaconda Navigator ortamında Phyton programlama dilinde Numpy, Pandas, MatPlotLib, Seaborn, Scikit-learn kütüphaneleri kullanılarak K-en yakın komşu, Naive Bayes, Destek Vektör Makineleri, Karar Ağaçları, Rastgele Orman, Boosting makine öğrenmesi algoritmaları ile (bu algoritmalardan bazıları, verileri eğitim ve test verisi olarak ayırırken dağılımdan kaynaklı aşırı öğrenme, yanlılık veya öğrenememe gibi hata oluşma ihtimallerine karşın k-katlı çapraz doğrulama ve katmanlı k-katlı çapraz doğrulama model değerlendirme yöntemleriyle birlikte tekrar uygulanarak) elde edilmiştir (Albon, 2018), (Jake, 2017). Şekil 1 verinin mevcut sınıf etiketlerine göre dağılımı ile, bu sınıf etiketlerinin sırasıyla 0:0, 1:0, 2:0, 3:1, 4:1, 5:1, 6:2, 7:2 ve 0:0, 1:0, 2:0, 3:0, 4:1, 5:1, 6:1, 7:1 tasvirleri ile oluşturulmuş iki yeni veri setinin sınıf etiket sayılarına göre dağılımını göstermektedir.



Şekil 1. Verinin 8, 3 ve 2 Sınıf Etiketi Sayısına Göre Dağılım Grafiği

3.2. İşlem Akış Diyagramı

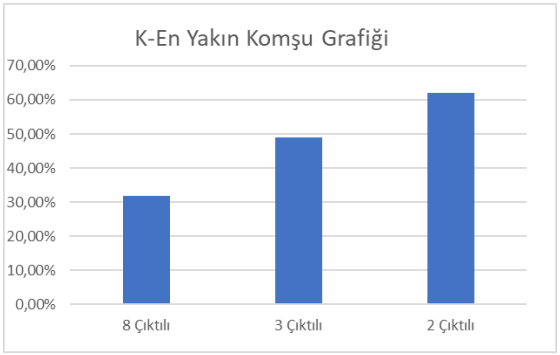
Şekil 2'de, 31 öznitelik 8 sınıf etiketine sahip veri setine dönüşüm uygulayarak verilerin oluşturulması ve kurulan matematiksel modellemelere dair genel akış diyagramı verilmiştir.

Şekil 2. Çalışmaya Ait İşlem Akış Diyagramı

4. MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE MODEL BAŞARI ANALİZİ

4.1. K-En Yakın Komşu Algoritması

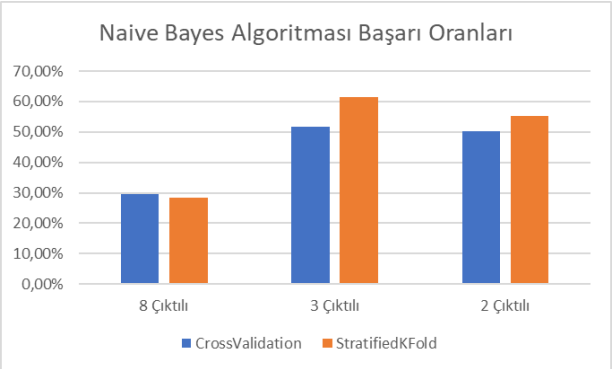
K-en yakın komşu algoritması yukarıda bahsedilen üç farklı veri seti için ayrı ayrı ele alınarak ızgara taraması (grid-search) ile en yakın komşu sayısı (n_neighbours) için 1, 3, 5, 7, 9, 11, 12 değerleri; Manhattan, Öklid, Minkowski metrikleri (p) için 1, 2, 3 değerleri kullanılarak cv=5 çapraz doğrulama ile veriler toplamda 105 kez fit edilmiş ve en uygun parametreler k=5, p=1 olarak belirlenmiştir. Şekil 3'te kurulan modelin başarı oranları verilmiş ve en yüksek model başarısının 0,62 ile 2 sınıf etiketli veriye ait olduğu görülmektedir.



Şekil 3. K-En Yakın Komşu Algoritmasının Çıktı Sayısına Göre Başarı Oranları

4.2. Naive Bayes Algoritması

Naive Bayes algoritması k katlı çapraz doğrulama (K-Fold Cross Validation) ve örnek yüzdeliklerini koruyan k katlı çapraz doğrulama (StratifiedKFold) yöntemi ile ayrı ayrı uygulanmıştır. Şekil 4'te elde edilen sonuçlar görselleştirilmiş ve en yüksek model başarısının örnek yüzdeliklerini koruyan k katlı çapraz doğrulama (StratifiedKFold) yöntemi kullanılarak 0.61 model başarısı ile 3 sınıf etiketine sahip veri seti olduğu görülmektedir.



Şekil 4. Sınıf Etiketi Sayısına Göre Hesaplanan Naive Bayes Algoritması Başarı Oranları

4.3. Destek Vektör Makineleri Algoritması

Destek vektör makinelerinde önemli iki parametre C Regularizasyon Parametresi ve γ Gamma-Uzaklık Parametresidir. Mevcut üç farklı veri seti %80-%20 train-test data olarak bölünmüş ve model başarım oranını maksimum yapacak en uygun parametreler için ızgara taraması (grid search) uygulanmıştır. 8 sınıf etiketli veri seti için C=10, γ=scale ve RBF Kernel en uygun parametreleri ile model başarısı 0,41 olarak hesaplanmıştır. 3 sınıf etiketli veri seti için C=100, γ=auto ve RBF Kernel en uygun parametreleri ile model başarısı 0,62 olarak belirlenmiştir. 2 sınıf etiketli veri seti için C=10, γ=scale ve RBF Kernel en uygun parametreleri ile model başarısı 0,72 olarak bulunmuştur.

4.4. Karar Ağaçları Algoritması

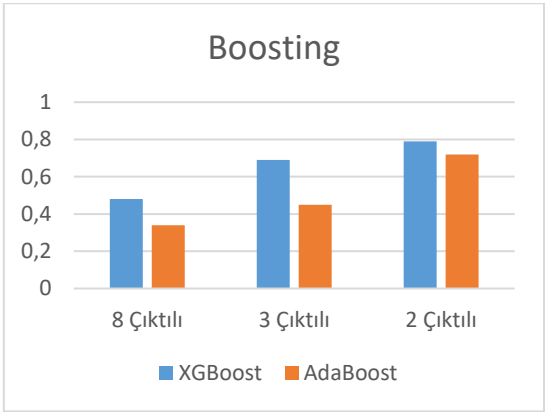
Karar ağaçları algoritması %80 eğitim, %20 test olarak ayrılmış verilere maksimum derinlik (max_depth) parametresi 3 alınarak "gini" ve "entropy" kriterleri için uygulandığında elde edilen sonuçlar Tablo 3'te özetlenmiştir.

Tablo 3. Bölünme Kriterinin Çıktı Sayısına Göre Hesaplanmış Karar Ağacı Algoritması Başarısı

	8 Çıktı	3 Çıktı	2 Çıktı
Gini Kriteri	%34,48	%68,97	%86,21
Entropy Kriteri	%37,93	%75,86	%86,21

4.5. Boosting Algoritması

Boosting algoritmalarından XGBoost ve AdaBoost, %80'e %20 eğitim ve test olmak üzere ayrılmış mevcut üç veri setine {n_estimators=400, learning_ rate=1, max_depht=6} parametreleri ile uygulandığında elde edilen model başarı ölçütleri Şekil 5'te özetlenmiştir. Buna göre en yüksek model başarısına sahip algoritma 2 sınıf etiketine sahip veriler için 0,79 model doğruluğu ile XGBoost algoritmasıdır.

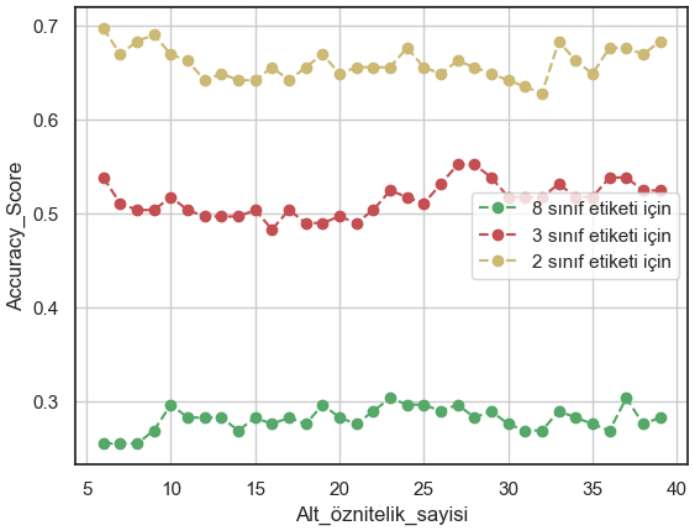


Şekil 5. Farklı Yöntemlerle Sınıf Sayısına Göre Hesaplanan Boosting Algoritması Başarı Oranları

4.6. Boyut Azaltma ve Öznitelik Önem Sıralaması

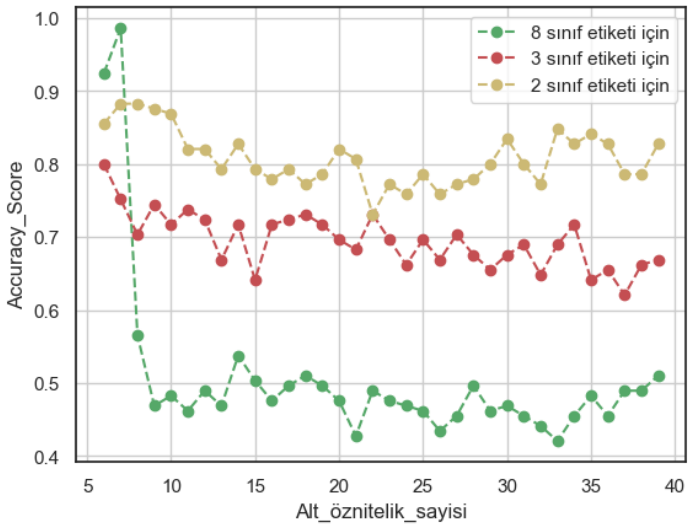
Rastgele orman algoritması ile verinin 31 özniteliğinin model başarısı üzerinde yüzdelik etkileri araştırıldığında; akademik ilk 5 özniteliğin sırasıyla: course-ıd, son yarıyıl genel not ortalaması, mezuniyette beklenen genel not ortalaması, bursluluk türü, dersi dinleme olduğu; demografik bilgiler içeren ilk 5 özniteliğin ise sırasıyla: öğrenci yaşı, baba mesleği, kardeş sayısı, anne eğitimi ve anne mesleği olduğu gözlenmiştir.

Ayrıca, yukarıda 31 kategorik özniteliğin 116 alt özniteliğin 0-1 şeklinde ifadesi sonrası, şekil 1 de gösterilen 8,3 ve 2 sınıf etiket sayısı ile düzenlenmiş yeni üç veri seti ile incelemelerinin yapıldığı algoritmalardan farklı olarak, veri boyutunun indirgenmesi verilerin yorumlanmasını kolaylaştırıp sınıflandırma başarısını arttıracağı için, boyut azaltma teknikleri olan PCA ve NCA sonrasında rastgele orman algoritması uygulanarak elde edilen sonuçlar karşılaştırılmıştır. Öncelikle PCA algoritması ile pca.explained_variance_ratio_ fonksiyonu kullanılarak 116 alt öz nitelikli veri setinin 40 adet bileşen tarafından %90 varyans yüzdesi ile ifade edildiği tespit edilmiştir. Bu amaçla alt öznitelikler PCA algoritması ile azaltılarak; 6 bileşenden 40 bileşene kadar her bir durum için rastgele orman algoritması kullanılarak ızgara taraması (grid search) ile belirlenmiş en iyi parametrelerle, 8, 3 ve 2 sınıf etiketli verilere uygulanmış ve elde edilen model başarıları Şekil 6'da gösterilmiştir.



Şekil 6. PCA Yöntemi ile Azaltılmış Alt Özniteliklerle Kurulan Rastgele Orman Algoritmasının Model Başarılarının Sınıf Etiket Sayısına Göre Değişimi

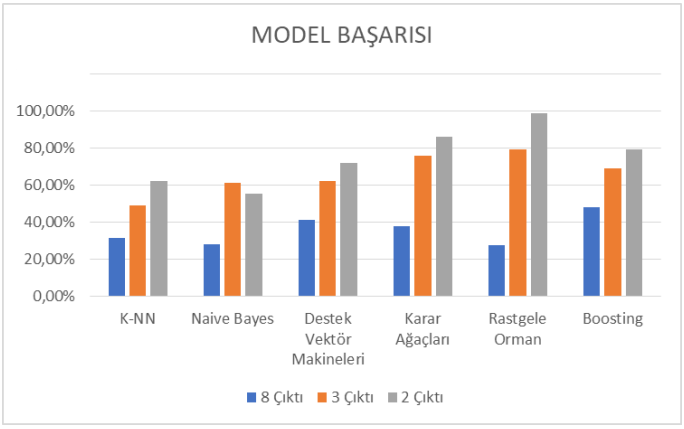
Aynı işlemeler NCA boyut indirgeme yöntemi kullanılarak tekrarlanmış ve sonuçlar Şekil 7'de verilmiştir:



Şekil 7. NCA Yöntemi ile Azaltılmış Alt Özniteliklerle Kurulan Rastgele Orman Algoritmasının Model Başarılarının Sınıf Etiket Sayısına Göre Değişimi

5. SONUÇ VE ÖNERİLER

Gelişen dünyada geleceğin teknolojilerini şekillendiren, hemen hemen her alanda kullanılan ve sıkça duyduğumuz yapay zeka uygulama alanlarından biri olan makine öğrenmesi, eğitim alanında da kullanılmaya başlamıştır. Yapılan literatür araştırmalarına göre öğrencilerden elde edilen çok sayıda verinin işlenerek bunlardan anlamlı sonuçlar çıkarılması ve bu sonuçların yorumlaması amacıyla kullanılan makine öğrenmesi yöntemleri ve altındaki matematiksel yapı günümüzde eğitimcilere faydalı içerikler sunmaktadır. Bu bağlamda yapılan bu çalışma eğitim öğretim kurumlarının belirleyecekleri stratejiler veya alacakları önlemlerde yol gösterici olabilir. Hatta daha büyük ana kütle için daha farklı okul türü ya da farklı kademelerde, farklı sektörlerde vb. uygulanarak sonuçların genelleştirilmesine fayda sağlayabilir. Geleneksel eğitimden ziyade öğrenci başarılarının önceden tahmin edilmesini sağlayan bu yöntem ile kurulan modellemeler karar verici konumundaki eğitim birimlerinde önleyici tedbirler alınmasına katkı sağlayabilir. Bu doğrultuda verilerin doğru bir şekilde incelenmesi önem arz etmektedir. Çalışmamızda kullanılan 8 sınıf etiketli veri seti, uygulanan sınıflandırma algoritmalarının tahmin başarılarını yükseltmek için, 3 ve 2 sınıf etiketli olacak şekilde düzenlenmiştir. Böylece model performans ölçütlerinden başarı kriterine göre daha yüksek sonuç verecek matematiksel modellemeler kurmak mümkün olmuş ve kurulan bu matematiksel modellemeler veri setini daha yorumlanabilir kılarak analizde kolaylıklar sağlamıştır. Akademik ve demografik bilgiler içeren 31 kategorik özniteliğin akademik başarı üzerindeki etkileri gözlenen ve Şekil 8 de model performans metriklerinden başarı ölçütüne göre sonuçları özetlenen bu çalışmanın literatüre katkı sağlayacağı düşünülmektedir.



Şekil 8. Kurulan Matematiksel Modellemelerin Sınıf Etiketi Sayılarına Göre Başarı Yüzdeleri

Yazarların Katkısı
Yazarların makaleye katkıları eşit orandadır.

Çıkar Çatışması Beyanı
Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı
Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

Bu çalışma Yıldız Teknik Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Birimince Desteklenmiştir. Proje Numarası: FYL-2024-6152

KAYNAKÇA

Abbasoğlu, B. (2020). Ortaokul Öğrencilerinin Akademik Başarılarının Eğitsel Veri Madenciliği Yöntemleri İle Tahmini. *Veri Bilimleri Dergisi*, 3(1), 1-10.

Akarsu, C. (2016). *Twitter verileri ile türk televizyonları izlenme oranı sıralamaları tahmini* [Yüksek Lisans Tezi]. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Albon, C. (2018) Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (1. Baskı). O'Reilly Media.

Alpaydın, E. (2010). Introduction to Machine Learning. *The MIT Press, Cambridge, Massachusetts*, London, England.

Amasyalı, M.F. (2008). *Yeni makine öğrenmesi metodları ve ilaç tasarımına uygulamalar* [Doktora Tezi]. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Bezek Güre, Ö. (2023). Investigating the Performance of Feature Selection Methods in Classifying Student Success, International Journal of Education Technology and Scientific Researches, 8(24), 2695-2728

Bilgin, M. (2018). Makine Öğrenmesi Teorisi ve Algoritmaları. *Papatya Yayıncılık Eğitim*, İstanbul.

Bozinovski, S. (1981). *Teaching space: A representation concept for adaptive pattern classification*. Department of Computer and Information Science, University of Massachusetts, Amherst, COINS Technical Report No. 81-28.

Breiman, L. (2001). Random Forests. *Machine Learning*,45,5-32,2001.

Çalışkan, M.& Talu M.F. (2020). Boyut indirgeme yöntemlerinin karşılaştırmalı analizi. *Türk Doğa ve Fen Dergisi*, 9(1), 107-113. DOI:10.46810/tdfd.707200.

Cortes, C.& Vapnik, V. (1995).” Support-Vector Networks”, *Machine Learning*, 20(3), 273-297.

Fernandes, E., Holand, M., Victorino, M., Borges, V., Carvalho, R. & Erven, G.V. (2019). Educational data mining: predictive analysis of academic performance of public-school students in the capital of Brazil. *Journal of Business Research*, 94(C), 335-343, DOI: 10.1016/j.jbusres.2018.02.012.

Goldberger, J., Roweis, S., Hinton, G., & Salakhutdinov, R. (2005) Neighbourhood Components Analysis, Advances in Neural Information Processing Systems, 17, 513-520.

Gök, M. (2017). Makine öğrenmesi yöntemleri ile akademik başarının tahmin edilmesi. *Gazi Üniversitesi Fen Bilimleri Dergisi Part(C): Tasarım ve Teknoloji*, 5(3), 139-148.

Han, J., Kamber, M.& Pei J. (2011). Data Mining: Concept and Techniques. *Morgan Kaufmann Publications*, USA.

Harrington, P. (2012). “Machine Learning In Action”, *By Manning Publications Co*, USA.

Hengpraprohm, K., Hengpraprohm, S., & Sudjitjoon, W. (2022). A Study of Factors Affecting Learning Efficiency on Higher Education Student Performance Evaluation Dataset Using Feature Selection Techniques. Information Technology Journal, 18(2), 34-43.

Jabardi, M. H. (2022). Machine learning techniques for assessing students'environments'impact factors on their academic performance. International Journal of Advanced Research in Computer Science, 13(2). http://dx.doi.org/10.26483/ijarcs.v13i2.6813

Jake, V. (2017), Python Data Science Handbook: Essential Tools for Working with Data, O’Reilly Media, Inc.

Kaur, P., Singh, M., & Singh Josan, G. (2015, Mart, 12-13). *Classification and prediction-based data mining algorithms to predict slow learners in education sector*. 3 rd International Conference on Recent Trends in Computing. Procedia Computer Science Journal, 57, India, 500-508.

Kayalı, S.& Buyrukoğlu, S. (2022, Haziran, 23-26). *Makine öğrenmesi yöntemleri ile öğrencilerin akademik başarılarının sınıflandırılması*. 2nd International Conference on Educational Technology and Online Learning-ICETOL2022 Full Paper Proceedings, Balıkesir, 330-336.

Lantz, B. (2013). Machine Learning with R. *Packt Publishing*, 35 Livery Street, Barmingham B3 2PB, UK.

Marquez-Vera, C., Cano, A., Romero, C., Noaman, AYM., Fardoun, HM. & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert System*, 33(1), 107-124, DOI:10.1111/exsy.12135.

Mitchell T.M. (1997). Machine Learning. *McGraw-Hill Science/Engineering/Math Publisher*, Kaliforniya.

Nayak S., Bhat M., Reddy N V S., & Rao B A. (2022) Study of distance metrics on k - nearest neighbor algorithm for star categorization, *Journal of Physics: Conference Series* 2161, 012004.

Nedeva V., & Pehlivanova T. (2021) Students' Performance Analyses Using Machine Learning Algorithms in WEKA. IOP Conf. Ser. Mater. Sci. Eng., vol. 1031, no. 1, p. 12061, 2021, doi: 10.1088/1757- 899X/1031/1/012061.

Pallathadka H., Wenda, A., Ramirez-Asís E., Asís-López M., FloresAlbornoz J. & Phasinam K. (2021). Classification and prediction of student performance data using various machine learning algorithms, Mater. Today Proc.doi: https://doi.org/10.1016/j.matpr.2021.07.382.

Phatai, G., & Luangrungruang, T. (2023, March, 18-20). A Comparative Study of Hybrid Neural Network with Metaheuristics for Student Performance Classification. In 2023 11th International Conference on Information and Education Technology (ICIET) (pp. 448-452). IEEE. Fujisawa, Japan

Romero, C., Espejo, PG., Zafra, A., Romero, JR., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use moddle courses. *Computer Applications in Engineering Education*, 21(1), 135-146, DOI:10.1002/cae.20456.

Sokolova, M. & Lapalme G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management* ,45, 427–437.

Sweeny, M., Lester, J., Rangwala, H. & Johri, A. 2016. Next-Term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining*, 8(1), 22-51, https://doi.org/10.5281/zenodo.3554603.

Şahin, S. (2021). *Makine öğrenmesi yöntemleri ile ortaokul öğrenci başarılarının tespiti ve bir uygulama* [Yüksek Lisans Tezi]. İstanbul Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Şekeroğlu, B., Dimililer, K., & Tuncal, K. (2019) Student Performance Prediction and Classification Using Machine Learning Algorithms ICEIT 2019: Proceedings of the 2019 8th International Conference on Educational and Information Technology 7–11.

Tipping, M.E. & Bishop C.M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2),443-482. DOI:10.1162.

Tosunoğlu, R., Yılmaz, E., Özeren, E. & Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi,* 3(2), 178-199, DOI:10.38151.

UCI Machine Learning Repository (2024): Higher Education Students Performance Evaluation Dataset https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation. adresinden 25 Şubat 2024 tarihinde erişildi.

Yang, W., Wang, K.& Zuo W. (2012). Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, 7(1), 161-168.

Yıldız, M. & Börekçi, C. (2020). Predicting Academic achievement with Machine learning algorithms. *Journal of Educational Technology & Online Learning*, 3(3), 372-392, DOI:10.31681/jetol.773206.

Yılmaz, N.& Şekeroğlu, B. (2019, Ağustos, 27-28). *Student Performance Classification Using Artificial Intelligence Techniques*. 10 th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions-ICSCCW-2019, Prag, 596-603.

Xu, G., Zong, Y. &Yang, Z. (2013). Applied Data Mining. *CRC Press*, NewYork.

Zhang H., (2004). The Optimality of Naive Bayes. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004).