MAKİNE ÖĞRENİMİ YÖNTEMLERİNİ KULLANARAK ÖĞRENCİLERİN OKULU BIRAKMASININ TAHMİNİ

Ali YILMAZ

Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA

Makale Bilgisi

Düzeltme: 21/04/2024
Teslim: 21/05/2024

Anahtar Kelimeler

Okul bırakma, Makine Öğrenimi, Sınıflandırma, Feature Selection, Tahmin, Çapraz Doğrulama

Öz

Okulu bırakma, eğitim sistemi ve bireylerin geleceği açısından önemli bir sorundur. Erken tespit ve etkili yönetim, bireylerin eğitim hayatındaki başarılarını ve genel refahlarını artırmada kritik öneme sahiptir. Bu çalışmada, makine öğrenimi yöntemlerini kullanarak okulu bırakma riskini tahmin etmek için bir model geliştiriyoruz. Modelin performansını değerlendirmek için 10 kat çapraz doğrulama ve feature selection yöntemleri kullanarak, Lineer Destek Vektör Makineleri (LSVM), K-En Yakın Komşu (KNN), Naive Bayes, RBF Destek Vektör Makineleri (RBF SVM), Rastgele Orman (Random Forest), Yapay Sinir Ağları (MLP), AdaBoost ve Lojistik Regresyon gibi çeşitli algoritmalar uyguluyoruz. Sonuçlar, makine öğrenimi modellerinin okulu bırakma tahmini için etkili bir araç olabileceğini göstermektedir.

SCHOOL DROPOUT PREDICTION USING MACHINE LEARNING METHODS

Keywords

School Dropout, Machine Learning, Classification,Feature Selection,Prediction, Cross-Validation

Abstract

School dropout is an important issue in terms of the education system and the future of individuals. Early detection and effective management are critical in enhancing individuals' success in their educational life and overall well-being. In this study, we are developing a model to predict the risk of school dropout using machine learning methods. To evaluate the performance of the model, we apply various algorithms, including Linear Support Vector Machines (LSVM), K-Nearest Neighbors (KNN), Naive Bayes, RBF Support Vector Machines (RBF SVM), Random Forest, Multi-Layer Perceptron (MLP), AdaBoost, and Logistic Regression, using 10-fold cross-validation and feature selection methods. The results show that machine learning models can be an effective tool for predicting school dropout.

1. GİRİŞ (INTRODUCTION)

Okulu bırakma, eğitim sistemi ve bireylerin geleceği açısından ciddi bir sorun olup, küresel olarak eğitim düzeyleri üzerinde önemli bir sorun oluşturmaktadır. Okulu bırakmanın erken tespiti ve etkili yönetimi, bireylerin eğitim hayatındaki başarılarını ve genel refahlarını artırmada hayati bir rol oynamaktadır. Son yıllarda, makine öğrenimi teknikleri, okulu bırakma tahmini ve tespitinde büyük bir potansiyel göstererek eğitim hizmetlerinde umut verici araçlar haline gelmiştir.

Bu çalışmada kullanılan veri seti, Instituto Politécnico de Portalegre, Bilgisayar Bilimleri ve Mühendisliği Bölümü'nden Valentim Realinho ve Mónica Vieira Martins tarafından derlenmiştir. Bu veri seti, ResearchGate sitesinde yayınlanan “Early Prediction of student’s Performance in Higher Education: A Case Study” başlıklı makalesiyle birlikte araştırmamızın temelini oluşturmaktadır.

Veri seti, 37 bağımsız tahmin değişkeni ve okulu bırakmanın varlığını veya yokluğunu temsil eden bir hedef değişkeni ile karakterize edilen 4424 örnek içermektedir. Bu tahmin ediciler, demografik, akademik ve yaşam tarzı faktörleri gibi çeşitli unsurları kapsayarak, öğrencilerin okulu bırakma riskini belirleyen değerli bilgiler sağlar.

Amacımız, sağlanan özelliklere dayanarak bir öğrencinin okulu bırakma olasılığını tahmin edebilecek bir model geliştirmektir. Bu modelin performansını değerlendirmek için farklı makine öğrenimi sınıflandırma yöntemlerini kullanıyoruz. Bu yöntemler arasında Lineer Destek Vektör Makineleri (LSVM), K-En Yakın Komşu (KNN), Naive Bayes, RBF Destek Vektör Makineleri (RBF SVM), Rastgele Orman (Random Forest), Yapay Sinir Ağları (MLP), AdaBoost ve Lojistik Regresyon bulunmaktadır.

Model performansını değerlendirmek amacıyla 10 kat çapraz doğrulama ve özellik seçimi (feature selection) yöntemlerini kullanarak, her bir modelin sınıflandırma performansını doğruluk, hassasiyet, özgüllük ve F1 skoru gibi metrikler üzerinden analiz ediyoruz. Bu çalışmalarla, makine öğrenimi modellerinin okulu bırakma tahmininde ne kadar etkili olabileceğini değerlendirerek, eğitim uygulamalarına yönelik daha kapsamlı bir anlayış geliştirmeyi amaçlıyoruz.

2. YÖNTEMLER (METHODS)

2.1. Veri Seti (Data Set)

Bu veri seti, 4424 örnekten oluşmakta olup, her bir örnek okulu bırakma durumunu gösteren bir hedef değişken ve 37 bağımsız özellikten oluşmaktadır. Özellikler arasında öğrencinin not ortalaması, devamsızlık gün sayısı, ders çalışma süresi, aile gelir durumu, ebeveyn eğitim düzeyi, evde internet erişimi, okul sonrası etkinliklere katılım, okulda disiplin sorunları, cinsiyet, başvuru yaşı, okul türü, öğretmen desteği, sosyal çevre, uyku düzeni, sağlık durumu ve motivasyon düzeyi gibi faktörler yer almaktadır.

Her bir özniteliğin değer aralıkları da veri setinde yer almaktadır. Bu özelliklerin, okulu bırakma üzerindeki etkisini değerlendirmek ve doğru bir sınıflandırma modeli oluşturmak için veri seti analiz edilecektir. Veri setinin detaylı bir şekilde incelenmesi, model oluşturma ve performans değerlendirme sürecinin temelini oluşturacaktır.

Tablo 1. *Eğitim veri seti öznitelikleri ve açıklamaları*

Öznitelik	Açıklama
Marital Status	Medeni hal durumu
Application mode	Uygulama modu
Application order	Başvuru sırası
Course	Kayıt olunan kurs
Daytime/evening attendance	Gün veya akşam olarak katılım
Previous qualification	Önceki yeterlilik
Previous qualification (grade)	Önceki yeterlilik notu
Nacionality	Milliyet
Mother's qualification	Annenin yeterliliği
Father's qualification	Babanın yeterliliği
Mother's occupation	Annenin mesleği
Father's occupation	Babanın mesleği
Admission grade	Giriş Notu
Displaced	Terinden Edilmiş
Educational special needs	Eğitimsel özel ihtiyaçlar
Debtor	Borçlu
Tuition fees up to date	Güncel eğitim ücretleri
Gender	Cinsiyet
Scholarship holder	Burs alan kişi
Age at enrollment	Kayıt yaşı
International	Uluslararası
Curricular units 1st sem (credited)	Müfredat birimleri 1. dönem (kredili)
Curricular units 1st sem (enrolled)	Müfredat birimleri 1. dönem (kayıtlı)
Curricular units 1st sem (evaluations)	Müfredat birimleri 1. yarıyıl (değerlendirmeler)
Curricular units 1st sem (approved)	Müfredat birimleri 1. yarıyıl (onaylandı)
Curricular units 1st sem (grade)	Müfredat birimleri 1. yarıyıl (sınıf)
Curricular units 1st sem (without evaluations)	Müfredat birimleri 1. dönem (değerlendirmeler olmadan)
Curricular units 2nd sem (credited)	Müfredat birimleri 2. yarıyıl (kredili)

Curricular units 2nd sem (enrolled)	Müfredat birimleri 2. yarıyıl (kayıtlı)
Curricular units 2nd sem (evaluations)	Müfredat birimleri 2. yarıyıl (değerlendirmeler)
Curricular units 2nd sem (approved)	Müfredat birimleri 2. yarıyıl (onaylandı)
Curricular units 2nd sem (grade)	Müfredat birimleri 2. yarıyıl (sınıf)
Curricular units 2nd sem (without evaluations)	Müfredat birimleri 2. yarıyıl (değerlendirmeler olmadan)
Unemployment rate	İşsizlik oranı
Inflation rate	Enflasyon oranı
GDP	GSYH
Target	Hedef

2.2. Öznitelik Seçme (Feature Selection)

Veri setindeki değişkenler arasındaki kritik ilişkileri tespit etmek ve modelin performansını artırmak amacıyla öznitelik seçimi yapıldı. Bu süreçte, hedef değişkenle en yüksek korelasyona sahip olan öznitelikler belirlendi. Bir eşik değeri olan 0.3 üzerinde korelasyona sahip öznitelikler seçildi. Bu seçim sonucunda, 'Curricular units 2nd sem (grade)', 'Curricular units 1st sem (grade)', 'Curricular units 2nd sem (approved)', 'Tuition fees up to date' ve 'Curricular units 1st sem (approved)' öznitelikleri belirlendi. Bu öznitelikler, okulu bırakma riskini etkileyen en önemli faktörler olarak kabul edildi.

2.3. 10 Kat Çapraz Doğrulama (10-Fold Cross-Validation)

Model performansı değerlendirme sürecinde 10 kat çapraz doğrulama yöntemi uygulandı. Bu teknikte, veri seti 10 eşit parçaya ayrılır ve her parça sırayla test için kullanılırken geri kalanlar eğitim amacıyla kullanılır. Bu adım, toplamda 10 ayrı deneme yapılmasını sağlar ve her deneme sonrası elde edilen performans ölçütlerinin ortalaması alınarak modelin genel performansı belirlenir. Çapraz doğrulama yöntemi, modelin genelleme yeteneğini daha sağlıklı bir şekilde değerlendirmemize ve aşırı öğrenme veya eksik öğrenme gibi problemleri belirlememize yardımcı olur. Bu sonuçlar, modelin doğruluk, kesinlik, geri çağırma ve F1 skoru gibi performans metrikleri üzerindeki başarısını değerlendirmek için kullanıldı.

2.4. Regresyon ve Sınıflandırma Yöntemleri (Regression and Classification Methods)

Bu çalışmada, okulu bırakma riskini tahmin etmek için hem regresyon hem de sınıflandırma yöntemlerini kullanıyoruz. Regresyon teknikleri, belirli özelliklere dayanarak sürekli bir çıktı değeri tahmin etmek için kullanılırken, sınıflandırma yöntemleri ise bireyin okulu bırakma olasılığını belirli bir kategoriye atamak için kullanılır. Bu yöntemler, veri setindeki desenleri öğrenerek okulu bırakma durumunun varlığını veya yokluğunu tahmin etmek ve yeni veriler için öngörülerde bulunmak amacıyla kullanılır. Kullanılan yöntemler şunlardır:

- **Lineer Destek Vektör Makineleri (LSVM):** LSVM, iki sınıfa ayırma işlemini gerçekleştirmek için hiperplan adı verilen bir çizgi kullanır. Bu çizgi, sınıflar arasındaki boşluğu maksimize etmeye çalışır. LSVM, anlaşılabilirlik ve kullanım kolaylığı nedeniyle sıkça tercih edilen bir sınıflandırma yöntemidir.

- **K-En Yakın Komşu (KNN):** KNN, yeni bir veri noktasının sınıfını, ona en yakın K adet komşunun sınıfına göre tahmin eder. K değeri, algoritmanın performansını doğrudan etkileyen önemli bir parametredir. KNN, basit ve düşük parametre sayısına sahip bir sınıflandırma yöntemidir.

- **Naive Bayes:** Naive Bayes, Bayes teoremini kullanarak sınıflandırma yapar. Bu yöntem, özellikler arasındaki bağımsızlığı varsayar ve her bir özelliğin sınıf etiketi üzerindeki etkisini hesaplar. Naive Bayes, hesaplama açısından ekonomik ve basit bir sınıflandırma yöntemidir.

- **RBF Destek Vektör Makineleri (RBF SVM):** RBF SVM, veri noktalarını iki sınıfa ayırmak için hiperplan kullanır, ancak LSVM'den farklı olarak RBF SVM, Gaussian fonksiyonunu kullanarak veri noktaları arasındaki mesafeyi hesaplar. Bu sayede, daha karmaşık karar sınırları oluşturabilir.

- **Rastgele Orman (Random Forest):** Rastgele Orman, birçok karar ağacından oluşan bir topluluk modelidir. Her bir karar ağacı, rastgele seçilen alt küme veri noktaları ve özellikler kullanılarak oluşturulur. Son tahmin, tüm ağaçların tahminlerinin ortalaması veya çoğunluğu alınarak hesaplanır. Rastgele Orman, yüksek doğruluk ve gürültüye karşı dirençli bir sınıflandırma yöntemidir.

- **AdaBoost:** AdaBoost, zayıf sınıflandırıcıları birleştirerek daha güçlü bir sınıflandırıcı oluşturan bir yöntemdir. Her iterasyonda, yanlış tahmin edilen veri noktalarına ağırlık vererek yeni bir sınıflandırıcı oluşturur. Son tahmin, tüm sınıflandırıcıların tahminlerinin ağırlıklı ortalaması olarak hesaplanır. AdaBoost, gürültüye karşı dirençli ve düşük hata oranına sahip bir sınıflandırma yöntemidir.

- **Lojistik Regresyon:** Lojistik regresyon, ikili sınıflandırma problemleri için kullanılan bir istatistiksel yöntemdir. Bu yöntem, bir olayın olasılık dağılımını modellemek için lojistik fonksiyonu kullanır. Lojistik regresyon, basit ve yorumlanabilir bir sınıflandırma yöntemidir.

- **Gradient Boosting:** Gradient Boosting, zayıf öğrenicileri birleştirerek güçlü bir tahmin modeli oluşturan bir yöntemdir. Her bir öğrenici, önceki öğrenicinin hatalarını düzeltmeye odaklanarak eğitilir. Gradient Boosting, karmaşık ilişkileri öğrenme yeteneğiyle bilinir ve genellikle yüksek performanslı modeller elde etmek için tercih edilir.

- **Decision Tree (Karar Ağaçları):** Karar ağaçları, veri setindeki öznitelik değerlerine göre sınıflandırma veya regresyon yapmak için kullanılır. Veriyi sınıflara bölen kararlar alır ve bu kararlar ağaç yapısı şeklinde gösterilir. Karar ağaçları, veriye kolayca uyarlanabilir ve yorumlanabilir olmaları nedeniyle yaygın olarak kullanılan bir yöntemdir.

- **Çok Katmanlı Algılayıcı (MLP):** Çok katmanlı algılayıcılar (MLP'ler), bir giriş katmanı, bir veya daha fazla gizli katman ve bir çıkış katmanı içeren yapay sinir ağlarıdır. Her bir gizli katman, veri üzerinde çeşitli dönüşümler gerçekleştirerek karmaşık örüntüleri öğrenir. Eğitim sırasında, ağ belirli bir hedefi tahmin etmek için giriş özelliklerinden başlayarak ağırlıkları günceller.

Bu çalışmada, her bir yöntem için en uygun hiperparametreler, 10 kat çapraz doğrulama kullanılarak optimize edilmiştir. Elde edilen sonuçlar, kullanılan yöntemlerin öğrencilerin okulu bırakma riskini tahmin etmede başarılı olduğunu göstermiştir.

3. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

3.1 Performans Metrikleri (Performance Metrics)

Bu çalışmada, okulu bırakan öğrencileri tahmin eden makine öğrenimi modellerinin performansını objektif bir şekilde değerlendirmek adına dört temel metrik kullanılmıştır:

TP: Doğru pozitif (Gerçekte pozitif ve model tarafından pozitif olarak tahmin edilen)
TN: Doğru negatif (Gerçekte negatif ve model tarafından negatif olarak tahmin edilen)
FP: Yanlış pozitif (Gerçekte negatif ve model tarafından pozitif olarak tahmin edilen)
FN: Yanlış negatif (Gerçekte pozitif ve model tarafından negatif olarak tahmin edilen)

- **Ortalama Doğruluk Skoru (Average Accuracy Score):** Doğruluk terimi, bir modelin doğru tahminlerinin, toplam tahmin edilen veri noktalarının oranını ifade eder. Bu metrik, modelin genel tahmin doğruluğunu nicel olarak ölçmek için temel bir gösterge olarak kabul edilir.

$$Doğruluk = \frac{TP + TN}{TP+TN+FP+FN}$$

- **Ortalama Hassasiyet Skoru (Average Precision Score):** Hassasiyet kavramı, bir modelin gerçekten pozitif olan tüm veri noktalarını doğru şekilde pozitif olarak sınıflandırma yeteneğini ölçer. Yani, modelin pozitif sınıfı doğru şekilde tanımlama başarısını ifade eder.

$$Hassasiyet = \frac{TP}{TP+FP}$$

- **Ortalama Geri Çağırma Skoru (Average Recall Score)**: Geri çağırma, bir modelin tüm pozitif veri noktalarını doğru şekilde pozitif olarak sınıflandırma oranını ölçer. Bu metrik, modelin pozitif örnekleri kaçırmadan bulma yeteneğini değerlendirir.

$$Geri\ \ Çağrma = \frac{TP}{TP+FN}$$

- **Ortalama F1 Skoru (Average F1 Score):** F1 Skoru, hassasiyet ve geri çağırmanın harmonik ortalaması olarak hesaplanır. Bu metrik, modelin hem hassasiyetini hem de geri çağırmasını dengeli bir şekilde değerlendirir, bu da modelin genel performansını daha kapsamlı bir şekilde yansıtır.

$$Geri\ \ Çağrma = \frac{Hassasiyet\ x\ Geri\ Çağırma}{Hassasiyet\ +\ Geri\ Çağırma}$$

3.2. Deneysel Sonuçlar (Experimental Results)

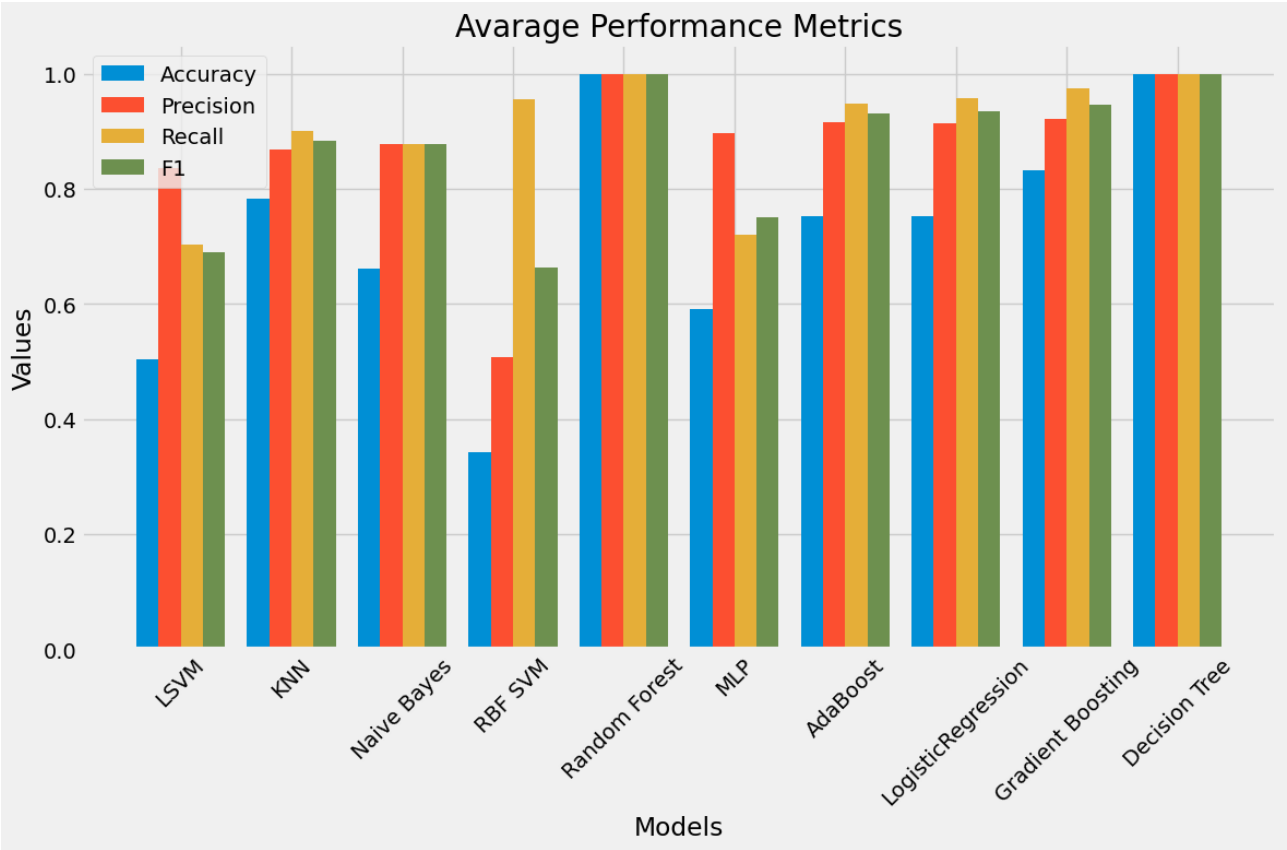
Makine öğrenimi modellerinin performansını değerlendirmek için iki aşamalı bir yaklaşım kullanıldı: çapraz doğrulama ve özellik seçimi. Çapraz doğrulama sonuçları ***(Tablo 3)***, çeşitli modellerin öğrencilerin okulu bırakma tespiti için ne kadar etkili olduğunu göstermektedir. Tablodan görüleceği üzere, Random Forest ve Decision Tree gibi modeller yüksek doğruluk ve F1 skoru elde etmiştir. Bu, bu modellerin okulu bırakan öğrencilerin tespiti için güçlü bir potansiyele sahip olduğunu göstermektedir.

Tablo 3: *Çapraz doğrulama ile sınıflandırıcı algoritmaların doğruluk sonuçları*

Model	Average Accuracy Score	Average Precision Score	Average Recall Score	Average F1 Score
LSVM	0.5038	0.8356	0.7037	0.6905
KNN	0.7825	0.8681	0.8996	0.8833
Naive Bayes	0.6610	0.8786	0.8777	0.8779
RBF SVM	0.3417	0.5069	0.9557	0.6622
Random Forest	1.0	1.0	1.0	1.0
MLP	0.5911	0.8965	0.7194	0.7498
AdaBoost	0.7523	0.9156	0.9482	0.9315
Logistic Regression	0.7516	0.9137	0.9576	0.9349
Gradient Boosting	0.8315	0.9219	0.9736	0.9469
Decision Tree	1.0	1.0	1.0	1.0

Bu grafik, her bir modelin cross-validation sonuçlarını görselleştirmektedir. Her bir bar, ilgili modelin belirli bir performans metriği için ortalama değerini temsil etmektedir.

Şekil 1: *Çapraz doğrulama ile sınıflandırıcı algoritmaların doğruluk sonuçları*



Öte yandan, özellik seçimi sonuçları **(Tablo 4)**, seçilen özelliklerin modellerin performansını nasıl etkilediğini göstermektedir. Özellikle, Random Forest ve AdaBoost modelleri, özellik seçimi sonrasında bile yüksek doğruluk skorları elde etmiştir. Bu, bu modellerin, veri setindeki önemli öznitelikleri başarıyla tanıyabildiğini ve bu özelliklerin etkili bir şekilde kullanılmasını sağladığını göstermektedir.

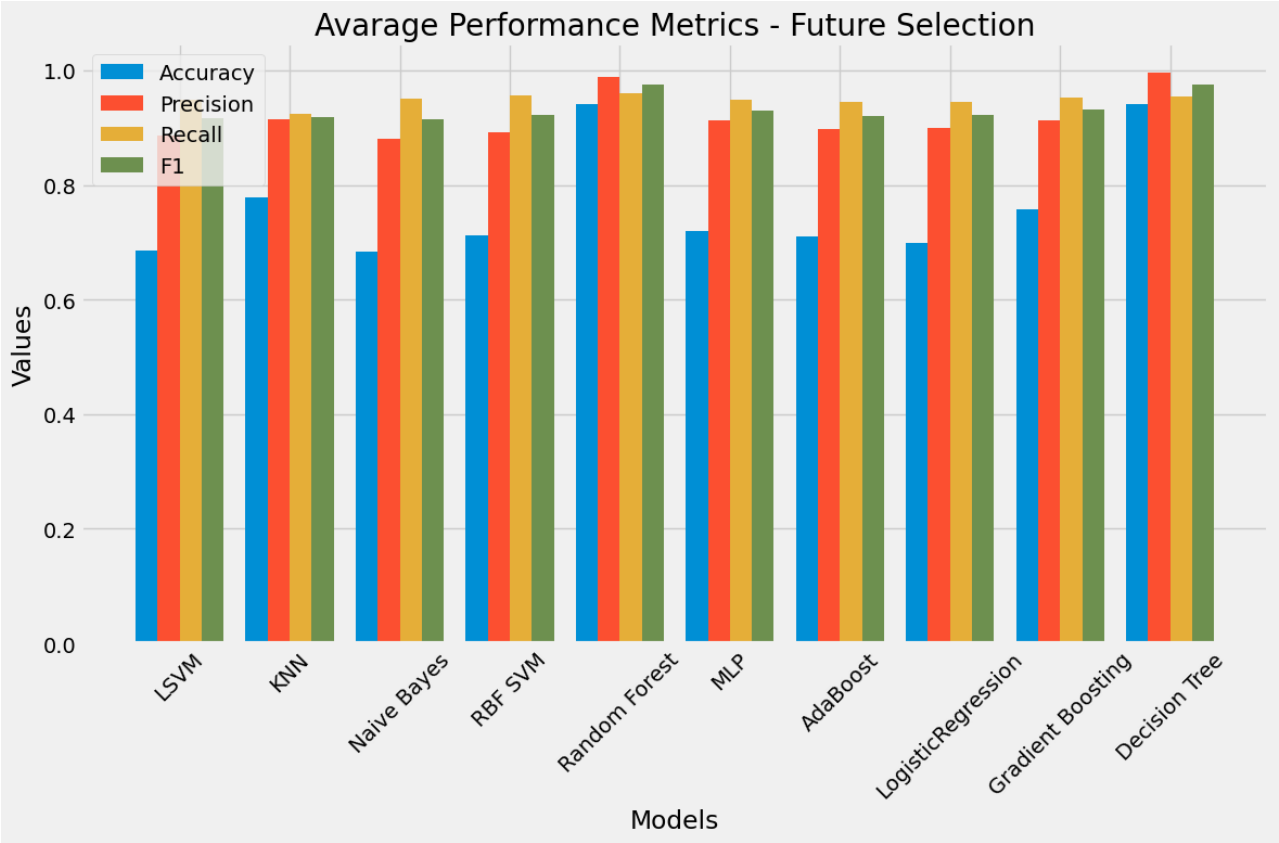
Tablo 3: *Öznitelik seçme ve çapraz doğrulama ile seçilen özniteliklere göre sınıflandırıcı algoritmaların doğruluk sonuçları*

Model	Average Accuracy Score	Average Precision Score	Average Recall Score	Average F1 Score
LSVM	0.6846	0.8864	0.9489	0.9165
KNN	0.7789	0.9142	0.9243	0.9191
Naive Bayes	0.6828	0.8814	0.9503	0.9144
RBF SVM	0.7116	0.8909	0.9558	0.9220
Random Forest	0.9409	0.9886	0.9608	0.9744
MLP	0.7197	0.9136	0.9482	0.9304
AdaBoost	0.7105	0.8967	0.9443	0.9198
Logistic Regression	0.6991	0.8993	0.9457	0.9218

Gradient Boosting	0.7576	0.9132	0.9527	0.9323
Decision Tree	0.9403	0.9966	0.9541	0.9748

Bu grafik, öznitelik seçme işlemi uygulandıktan sonra her bir modelin performansını göstermektedir.

Şekil 2: *Öznitelik seçme ve çapraz doğrulama ile seçilen özniteliklere göre sınıflandırıcı algoritmaların doğruluk sonuçları*



4. SONUÇLAR (CONCLUSIONS)

Bu çalışmada, çeşitli makine öğrenimi sınıflandırma algoritmaları, 10 kat çapraz doğrulama ve öznitelik seçimi yöntemleri kullanılarak modellerin performansları değerlendirildi. Elde edilen bulgular aşağıdaki gibidir:

- Öğrencilerin okulu bırakmasını tahmin etmek için kullanılan makine öğrenimi modelleri, yüksek doğruluk skorları elde etmiştir. Özellikle, Random Forest ve Decision Tree gibi modeller, yüksek doğruluk ve F1 skorlarına sahiptir.
- Öznitelik seçimi sürecinde, 'Target', 'Curricular units 2nd sem (grade)', 'Curricular units 1st sem (grade)', 'Curricular units 2nd sem (approved)', 'Curricular units 1st sem (approved)' ve 'Tuition fees up to date' gibi belirli özniteliklerin seçildiği tespit edilmiştir.
- Çapraz doğrulama ve öznitelik seçimi sonuçları, makine öğrenimi modellerinin öğrencilerin okulu bırakmasını tahmin için etkin bir araç olduğunu göstermektedir.

Bu sonuçlar, öğrencilerin okulu bırakması ve yönetimi üzerine yapılan araştırmalarda makine öğrenimi tekniklerinin kullanımının önemini vurgulamaktadır. Bu çalışma, öğrencilerin okulu bırakmasını önleme ve yönetme alanında daha iyi bir anlayış geliştirmek için bir temel oluşturmaktadır.

KAYNAK DOSYALAR (SUPPLEMENTARY FILES)

Bu çalışmada kullanılan veri setlerine aşağıdaki web adresinden ulaşılabilir:
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu çalışma, Sayın Murat GÖK tarafından desteklenmiştir. Ayrıca yapmış oldukları katkılardan dolayı Yalova Üniversitesine teşekkür ederiz.

KAYNAKÇA(ACKNOWLEDGMENTS)

[1] MVMartins, D. Tolledo, & V.Realinho (2021). Predict Students' Dropout and Academic Succes, 10.24432/C5MC89. DOI: https://doi.org/10.24432/C5MC89
[2]Murat GÖK, Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi, Fen Bilimleri Dergisi, 2007
[3]Alpaydin, E., “Introduction to Machine Learning”, 210-212, The MIT Press, Londra, 2004. [4] Breiman, L., “Random Forests”, Machine Learning, Cilt 45, No 1, 5-32, 2001.
[5] Hall, M.A., “Correlation-based Feature Subset Selection for Machine Learning”, Hamilton, New Zealand, 1998.