Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(1) (2023) 43-51
DOI: 10.38016/jista.1183353

Makine Öğrenmesi Yöntemleri Kullanılarak Öğrencilerin Kazanım Bilgileri ile Sınavlardaki Başarı Durumunun Tahmini

Muhammed Fatih Adak[1*], Ömer Duralioğlu[2]

[1] Sakarya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Sakarya, Türkiye

[2] Sakarya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Sakarya, Türkiye

fatihadak@sakarya.edu.tr, oduralioglu@hotmail.com

Öz

Yapılan bu çalışmada öğrencilerin girdiği sınav verilerine göre sonraki sınavlardaki performansları tahmin edilmek istenmiştir. Veri kümesi olarak, 2021-2022 eğitim öğretim yılı 1. döneminde, İstanbul ili Ataşehir ilçesinde bulunan Dr. Nureddin Erk-Perihan Erk Mesleki ve Teknik Anadolu Lisesi, Bilişim Teknolojileri alanındaki 10 ve 11'inci sınıfta okuyan 87 öğrencinin Nesne Tabanlı ve Programlama dersinde uygulanan 3 sınavdaki puan dağılımları kullanılmıştır. Sınavlardaki sorular ders bilgi formundaki kazanım başlıklarıyla eşleştirilmiş, her öğrencinin kazanım başlıklarına göre performans oranları tablo haline getirilmiştir. Verilerin kısıtlı olmasından dolayı toplanan gerçek veriler kullanılarak sentetik veriler üretilmiştir. Sentetik verinin gerçeğe yakınlık derecesi detaylı sonuç raporu ile teyit edilmiştir. Birden çok sayıda performans değeri tahmin edileceğinden, çok çıkışlı regresyonu destekleyen doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmıştır. Algoritmaların başarı değerlendirmesi için k katmanlı çapraz doğrulama uygulanmıştır. Performans ölçümleri için MAE, MSE, R2 ve standart sapma hesaplanmıştır. Aşırı uyum çözümü için KNN ve karar ağacı algoritmalarında en iyi parametre değerleri bulunarak performans iyileştirilmiştir. Sonuçlara göre en iyi performans değerleri KNN ile elde edilmiştir. Bu çalışmanın devamı olarak tüm derslerin sınav verilerinin girileceği bir sistem tasarlanarak dersler arasındaki performans bağlantıları analiz edilebilir. Böylece performans tahminlerine göre öğrencilerin gelecekteki başarısızlıkları bugünden önlenebilir ve eğitim kalitesi arttırılabilir.

Anahtar kelimeler: Eğitsel veri madenciliği, öğrenci performans tahmini, çok çıktılı regresyon, sentetik veri, makine öğrenmesi

Estimation of Students' Achievement Information and Success in Exams Using Machine Learning Methods

Abstract

In this study, it was aimed to predict the performance of the students in the next exams according to the exam data they entered. As the data set, in the 1st semester of the 2021-2022 academic year, Dr. The score distributions of 87 students studying in the 10th and 11th grades of Nureddin Erk-Perihan Erk Vocational and Technical Anatolian High School in the field of Information Technologies in the 3 exams applied in the Object-Oriented and Programming course were used. The questions in the exams were matched with the achievement titles in the course information form, and the performance rates of each student were tabulated according to the achievement titles. Synthetic data were produced using real data collected due to limited data. The degree of closeness of the synthetic data to reality was confirmed by the detailed result report. Since more than one performance value will be estimated, linear regression, k-nearest neighbor and decision tree algorithms supporting multi-output regression are used. K-layer cross validation was applied to evaluate the success of the algorithms. MAE, MSE, R2 and standard deviation were calculated for performance measurements. For overfitting solution, the performance is improved by finding the best parameter values in KNN and decision tree algorithms. According to the results, the best performance values were obtained with KNN. As a continuation of this study, a system in which exam data of all courses will be entered can be designed and the performance connections between courses can be analyzed. Thus, future failures of students can be prevented and the quality of education can be increased according to performance predictions.

Keywords: Educational data mining, student performance prediction, multi-output regression, synthetic data, machine learning

* Sorumlu yazar.
E-posta adresi: fatihadak@sakarya.edu.tr

Alındı : 2 Ekim 2022
Revizyon : 24 Ekim 2022
Kabul : 31 Ekim 2022

1. Giriş (Introduction)

Gelişen teknoloji ile saklanan ve kullanılan veri boyutları her geçen gün artmaktadır. Verilerin hangi şekilde kullanılarak daha anlamlı hale getirilebileceği araştırıldıkça veri madenciliği ve makine öğrenmesi yöntemleri önem kazanmıştır. Veri madenciliği, geliştirilen yöntemler ile çok büyük miktardaki veri içerisinden işe yarayacak olanları kullanılabilir hale getirmektir. Makine öğrenmesi ise sisteme girilen veriler ve kullanılan algoritmalar ile bilgisayarın öğrenmesini ve gelecek hakkında tahminler üretmesini sağlayan tekniktir. Veri madenciliği ve makine öğrenmesi yöntemleri ile birçok alanda olduğu gibi eğitim alanında da yapılan çalışmalar artarak önem kazanmaya devam etmektedir.

Bu çalışmanın amacı öğrencilerin girdiği sınav verilerini kullanarak gelecekte gireceği sınav performanslarını tahmin etmektir. Benzer çalışmalar için yapılan literatür taramasında; İçeli çalışmasında Cumhuriyet Üniversitesi Divriği Nuri Demirağ MYO (Meslek Yüksek Okulu)'da Temel Bilgisayar Bilimleri dersini alan öğrencilere uygulanan anket ile elde edilen veriler kullanılarak Weka programı ile karar ağacı uygulayarak bir derse ait başarı analizi uygunlamış ve anlamlı bilgiler elde edilmiştir. (İçeli, 2012). Güvenç vd. öğrencilerin başarı tahmini ve eksikliklerini gidermek için bir dersi almadan önce öğrencilerin bilgi düzeylerini ölçmüş ve dönem sonu başarı puanlarınıda alarak öğrencileri başarılarına göre dört kategoriye ayırmıştır. Veriler yetersiz olduğu için SMOTE (The Synthetic Minority Over-sampling TEchnique) yöntemi ile sentetik veri üretilmiştir. Uygulanan altı makine öğrenmesi yönteminde en iyi sonucu SVM (destek vektör makinesi) algoritması vermiştir (Güvenç vd., 2022). Şengür, Fırat Üniversitesi BÖTE (Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü) bölümünden mezun öğrencilerin ders ve mezuniyet notu verilerini kullanarak yeni öğrenciler hakkında tahminde bulunmuştur. Yapay sinir ağı ve karar ağacı kullanmış, yapay sinir ağının daha iyi sonuçlar verdiği görülmüştür (Şengür, 2013). Aghalarova çalışmasında Kalboard 360 E-öğrenme sisteminden elde edilen verileri kullanarak otomatik makine öğrenmesi (AutoML) ile en iyi modelin seçimini araştırılmıştır. En iyi algoritma dağıtılmış rastgele orman algoritması olarak elde edilmiştir (Aghalarova, 2022). Gök çalışmasında öğrencilere uygulanan anket sonuçlarına göre dönem sonu genel başarı notları 6 makine öğrenmesi algoritması ile tahmin etmiştir. Puan tahmininde rastgele orman, not tahmininde ise öznitelik seçme yöntemi ile birlikte lojistik sınıflandırma algoritması en iyi sonucu vermiştir (Gök, 2017). Abbasoğlu çalışmasında dört farklı resmi ortaokuldaki öğrencilerin verilerini kullanarak uyguladığı 8 sınıflandırıcı algoritma analizi sonucunda lojistik regresyon en iyi sonucu vermiştir (Abbasoğlu, 2020). Aydemir vd. üniversitede Türk Dili dersini alan öğrenci verilerini kullanarak bu dersteki başarıları beş farklı ağaç algoritması ile tahmin edilmiştir. En iyi sonuçlar random forest algoritması ile elde edilmiştir (Aydemir vd., 2019). Can vd. üniversite öğrencilerine uygulanan anket ile iki adet ders başarı değişkeni tanımlamıştır. Bu değişkenleri etkileyen soru cevaplarıyla ders başarıları arasındaki ilişki lojistik regresyon ile tahmin edilmek istenmiş, analizler sonucunda %91,6 oranında tahmin başarısı sağlanmıştır (Can vd., 2018). Güner vd. Pamukkale Üniversitesi Mühendislik Fakültesi öğrencilerinin verilerini kullanarak destek vektör makinesi yöntemiyle öğrencilerin matematik-1 dersindeki alacakları puanları tahmin etmiştir. Yapılan analizeler sonucunda %86,36 doğruluk oranı ile tahmin sonuçları elde edilmiştir (Güner vd., 2011). Akgün vd. eğitsel veri madenciliğiyle ilgili yapılan 102 çalışmayı incelemiş ve çalışmalarda en çok kullanılan yazılımın WEKA olduğunu, en çok kullanılan tekniklerin ise karar ağaçları ve yapay sinir ağları olduğunu belirtmiştir (Akgün vd., 2020). Üniversite düzeyinde teknik ders öneri sistemi geliştirilmiş bulanık modelin başarısı yüksek olmuştur (Adak vd., 2016).

2. Materyal ve Metot (Material and Method)

Öğrencilerin girdiği sınavlardaki kazanımlarda gösterdiği performans verileri kullanılarak sonraki gireceği sınavlardaki performans değerleri sayısal olarak tahmin edileceğinden dolayı çok çıktılı regresyon kullanılmıştır.

2.1. Veri Kümesi (Data Set)

Veri kümesi olarak 2021-2022 eğitim öğretim yılı 1.döneminde, İstanbul ili Ataşehir ilçesinde bulunan Dr. Nureddin Erk-Perihan Erk Mesleki ve Teknik Anadolu Lisesi, Bilişim Teknolojileri alanındaki 10 ve 11'inci sınıfta okuyan 87 öğrencinin Nesne Tabanlı ve Programlama dersinde uygulanan 3 sınavın not dağılımları kullanılmıştır. Dersin kazanımları, MEB (Milli Eğitim Bakanlığı) tarafından hazırlanan ders bilgi formlarına bakılarak elde edilmiştir (MEB, 2022). Tablo 1'de Nesne Tabanlı Programlama dersinin kazanım başlıkları numaralandırılmış şekilde görülmektedir.

Sınav soruları kazanım başlıklarına göre gruplanmıştır. Gruplama sonucunda 1. sınavda 1, 3, 4, 5 numaralı kazanımlar, 2. sınavda 1, 3, 4, 5 numaralı kazanımlar, 3. sınavda 1, 3, 5, 7 numaralı kazanımlarla ilgili sorular olduğu görülmüştür. Excel programı kullanılarak her sorudan alınan puanlar tablo haline getirilmiştir. Daha sonra sorular kazanım başlıkları eşleştirilerek öğrencinin her kazanımdan toplamda kaç puan aldığını gösteren yeni bir tablo oluşturulmuştur. Sınavdaki sorulara bağlı olarak her kazanımın toplam puanı farklı olduğu için kazanımlardan alınan puanlar yüzlük sisteme göre yeniden düzenlenmiştir.

Tablo 1. Kazanım başlıkları (Educational attainment titles)

Kazanım Numarası	Kazanım Başlıkları
Kazanım 1	Yazım hatalarını dikkate alarak nesne tabanlı programlama çalışma ortamını kullanır.
Kazanım 2	Yazım hatalarını dikkate alarak isim uzaylarını kullanır.
Kazanım 3	Tanımlama kurallarını dikkate alarak değişkenleri ve temel veri türlerini kullanır.
Kazanım 4	İşlem önceliğine göre aritmetiksel operatörleri kullanır.
Kazanım 5	Yazım kurallarına dikkat ederek şart ifadelerini kullanır.
Kazanım 6	Mantıksal operatörleri öncelik sırasına uygun kullanır.
Kazanım 7	Yazım formatına dikkat ederek döngü yapılarını kullanır.
Kazanım 8	Programda hata ayıklaması yapar.

Yani öğrencinin sınavda her kazanımdan yüzde kaç başarı gösterdiği yeni bir tabloda hesaplanmıştır. Böylece çalışmadaki modellerde kullanılacak veri kümesi elde edilmiştir.

Toplanan verilerin kısıtlı olmasından dolayı Gretel sistemi ile farklı epochs ve batch size değerleri verilerek denemeler sonucunda epochs=10000 ve batch size=5000 değerleri kullanılarak 86 sentetik veri kalite puanı ile 5000 sentetik veri üretilmiştir. Gretel sisteminde 80 ve üzeri puanlama mükemmel olarak değerlendirilmektedir. Şekil 1'de görüldüğü gibi, alan korelasyon kararlılığı, derin yapı stabilitesi ve alan dağılım kararlılığı değerlerinin 80 ve üzerinde puanlandığından dolayı üretilen verilerin makine öğrenmesi çalışmalarında kullanılabilir olduğu görülmüştür.

Şekil 1. Veri özeti istatistikleri (Data Summary Statistics)

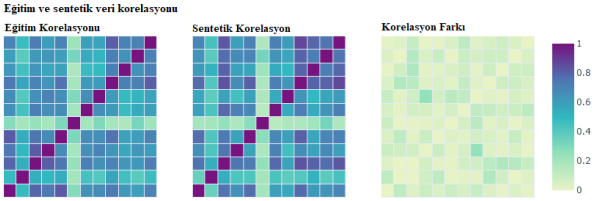
Şekil 2'de üretilen sentetik verilerin başlıklarına göre bilgiler yer almaktadır. Veri kümesi başlığındaki birinci sayı sınav numarasını, ikinci sayı ise o sınavdaki kazanım numarasını temsil etmektedir. Numaralarına göre kazanım başlıkları Tablo 1'de görülmektedir.

Eğitim Kalitesi

Alan	Tekil	Eksik Veri	Ort. Uzunluk	Tür	Dağılımın Kalitesi
K_2_3	2	0	1.94	İkili Veri	Mükemmel
K_3_4	12	0	5.06	Sayısal	Orta
K_1_3	14	0	10.31	Sayısal	İyi
K_2_5	15	0	7.09	Sayısal	İyi
K_3_1	29	0	9.48	Sayısal	İyi
K_3_5	22	0	5.51	Sayısal	İyi
K_1_5	20	0	6.15	Sayısal	Mükemmel
K_2_1	52	0	10.90	Sayısal	Mükemmel
K_1_4	29	0	8.99	Sayısal	Mükemmel
K_1_1	67	0	10.64	Sayısal	Mükemmel
K_2_4	18	0	7.22	Sayısal	Mükemmel
K_3_7	24	0	5.37	Sayısal	Mükemmel

Şekil 2. Eğitim alanına genel bakış (Training field overview)

Şekil 3'teki Gretel rapor ekranında gerçek ve sentetik verilerin ayrı ayrı olmak üzere kendi içindeki alanları arasındaki korelasyon verilmiştir. Sağda ise gerçek verilerin ve sentetik verilerin alanları arasındaki fark verilmiştir. Her iki veri grubu ayrı ayrı değerlendirildiğinde korelasyon oranlarının birbirine benzediği, ikisi arasındaki korelasyon farkının ise az olduğu görülmüştür.



Şekil 3. Eğitim ve sentetik veri korelasyonu (Training and Synthetic Data Correlation)

2.2. Kullanılan Regresyon Yöntemleri (Regression Methods Used)

Çok çıktılı regresyonu desteklediğinden dolayı doğrusal regresyon, k-en yakın komşu ve karar ağacı algoritmaları kullanılmıştır.

2.2.1.Doğrusal Regresyon (Linear Regression)

Doğrusal regresyon, bağımsız bir değişken ile bağımlı bir değişken arasındaki ilişkiyi tahmin etmek amacıyla kullanılan doğrusal bir yaklaşımdır. (Kumari vd., 2022)

Doğrusala yakın ilişki gözlendiğinde, basit doğrusal regresyon modelleri denklem (1)'deki gibi ifade edilir (Mardikyan, 2005):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1)$$

Çoklu doğrusal regresyon modelleri için ise denklem (2)'deki gibi ifade edilir:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} \varepsilon_i \quad (2)$$

2.2.2. K-En Yakın Komşu (K-Nearest Neighbors)

KNN algoritması, veriler arasında bir korelasyon olduğunu varsayarak, yeni gelen veriyi mevcut verilerle daha benzer olan kategoriye dahil eder. Burada k değeri, etraftaki kaç komşuya dikkat edileceğini temsil eder. Gözlemler arasındaki mesafe ölçümü için genellikle öklid mesafesi kullanılır (Altunkaynak vd., 2020).

$$Öklid\ mesafesi = \sqrt{\sum_{i=1}^{k}(xi - yi)^2} \quad (3)$$

2.2.3. Karar Ağacı (Decision Tree)

Karar ağacı algoritması, bir ağaç yapısı gibi kök düğüm, dallar ve yapraklardan oluşur. Veri kümesinin alt kümelere bölünmesiyle oluşur. En uçtaki yapraklar, regresyonda sayısal bir değeri temsil eder. Veri kümesi karmaşıksa ağaç dallanarak büyümektedir (Özlüer Başer vd., 2021).

2.3. K Katmanlı Çapraz Doğrulama (K-Fold Cross Validation)

Çapraz doğrulama, kullanılan verilere göre makine öğrenimi yöntemini değerlendirmek için kullanılır. K değeri verilerin kaç gruba bölüneceğini temsil eder. Gruplardan biri test için, k-1 tanesi ise eğitim için kullanılır. Tüm denemelerin sonucunda ortalama hata hesaplanır (Özlen, 2022).

2.4. Kullanılan Performans Ölçümleri (Performance Measurements Used)

Kullanılan regresyon modellerindeki performansı ölçmek amacıyla MAE, MSE, R^2 ölçümleri kullanılmıştır.

2.4.1. Ortalama Mutlak Hata (Mean Absolute Error-MAE)

Hatanın hedef çıktıya bölümüyle elde edilen değer toplamının veri kümesi sayısına bölünmesi sonucu elde edilir. Denklem 4'te, T değeri gerçek çıktı, O değeri ağda hesaplanan çıktı, n değeri ise veri kümesi sayısıdır (Adak, 2016).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|T-O|}{T} \quad (4)$$

MAE değerinin sıfıra yakın olması, modelin tahmin ettiği değerler ile gerçek değerlerin birbirine yakın olduğunu gösterir.

2.4.2. Ortalama Kare Hatası (Mean Squared Error-MSE)

Hata kareleri toplamının eğitimde kullanılan veri kümesi sayısına bölümü ile elde edilir. Denklem 5'te, T değeri gerçek çıktı, O değeri ağda hesaplanan çıktı, n değeri veri kümesi sayısıdır (Adak, 2016).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(T - O)^2 \quad (5)$$

MSE değerinin sıfıra yakın olması, modelin performansının iyi olduğunu gösterir.

2.4.3. Açıklayıcılık Katsayısı (Explanatory Coefficient-R^2)

Bağımsız değişken ile gerçek çıktıların ortalaması hesaplanarak kullanılır. Denklem 6'da, T değeri gerçek çıktı, T_{ort} değeri gerçek çıktıların ortalaması, O değeri ağda hesaplanan çıktı, n değeri veri kümesi sayısı, k değeri bağımsız değişken sayısıdır (Adak, 2016).

$$R^2 = 1 - \left[1 - \frac{\sum(T-O)^2}{\sum(T-T_{ort})^2}\right]\frac{n-1}{n-k-1} \quad (6)$$

R^2 değeri 0 ile 1 arasında bulunur. Hesaplanan sonucun bire yakın olması, modelinin performansının iyi olduğunu, sıfıra yakın olması ise modelin performansının kötü olduğunu gösterir.

2.5. Uygulama (Application)

Regresyon yöntemlerini uygulamak için Python, kodların çalıştırılması için Google Colab ve sentetik veri üretimi için Gretel kullanılmıştır. Veri işleme konusunda zengin kütüphaneleri sayesinde kullanıcılarına kolaylık sağlamaktadır. Analiz sonuçlarını kolaylıkla görme ve görselleştirme sayesinde genellikle tercih edilmektedir. Gretel, her türlü metin veya yapılandırılmış veriden yeni sentetik örnekler oluşturmak için Uzun Kısa Süreli Bellek (LSTM) yapay sinir ağını kullanmaktadır. Sentetik veri üretimi sonrasında üretilen verilerin kalitesiyle ilgili detaylı rapor sunmaktadır.

3. Uygulama ve Bulgular (Application and Findings)

Verilerin %80'i sistemin eğitimi için, %20'si test verisi olarak kullanılmıştır.

Öncelikle kullanılan algoritmalarda parametre olmadan performansları gözlenmiştir. İlk aşamada mevcut gerçek verilerle analiz yapılmıştır.

Tablo 2, Tablo 3 ve Tablo 4'te gerçek verilerle algoritmalardan elde edilen sonuçlar görülmektedir.

Tablo 2. Gerçek veriler ile doğrusal regresyon sonuçları (Linear regression results with real data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	26,477	951	0.471	7,609
1.sınavdan 3.sınav tahmini	21,256	535	0.622	6,324
2.sınavdan 3.sınav tahmini	21,981	653	0.526	8,922

Tablo 3. Gerçek veriler ile KNN sonuçları (KNN results with real data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	23,578	1139	0.389	4,899
1.sınavdan 3.sınav tahmini	19,523	721	0.497	7,626
2.sınavdan 3.sınav tahmini	25,331	749	0.45	9,742

Tablo 4. Gerçek veriler ile karar ağacı sonuçları (Decision tree results with real data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	29.659	1623	0.073	11.529
1.sınavdan 3.sınav tahmini	19.711	943	0.242	8.118
2.sınavdan 3.sınav tahmini	20.906	1354	-0.094	10.048

Gerçek verilerle elde edilen sonuçlara bakıldığında en iyi performansı doğrusal regresyon algoritması olduğu görülmektedir. MAE, MSE değerlerinin düşük olması hata oranının diğerlerine göre daha düşük olduğunu göstermektedir. R^2 değerinin 1'e yakın olması üretilen tahminlerin doğruluğunun diğerlerine göre daha iyi olduğunu göstermektedir.

Tablo 5, Tablo 6 ve Tablo 7'de sentetik verilerle yapılan analiz sonucunda algoritmalardan elde edilen sonuçlar görülmektedir.

Tablo 5. Sentetik veriler ile doğrusal regresyon sonuçları (Linear regression results with synthetic data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	20,423	768	0.448	0.578
1.sınavdan 3.sınav tahmini	14,980	451	0.567	1,034
2.sınavdan 3.sınav tahmini	16,320	500	0.525	0.991

Tablo 6. Sentetik veriler ile KNN sonuçları (KNN results with synthetic data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	19,493	859	0.39	0.766
1.sınavdan 3.sınav tahmini	14,927	483	0.539	1,155
2.sınavdan 3.sınav tahmini	15,889	554	0.478	1,179

Tablo 7. Sentetik veriler ile karar ağacı sonuçları (Decision tree results with synthetic data)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	22.802	1483	-0.034	1.536
1.sınavdan 3.sınav tahmini	18.003	759	0.27	1.417
2.sınavdan 3.sınav tahmini	19.354	904	0.145	1.621

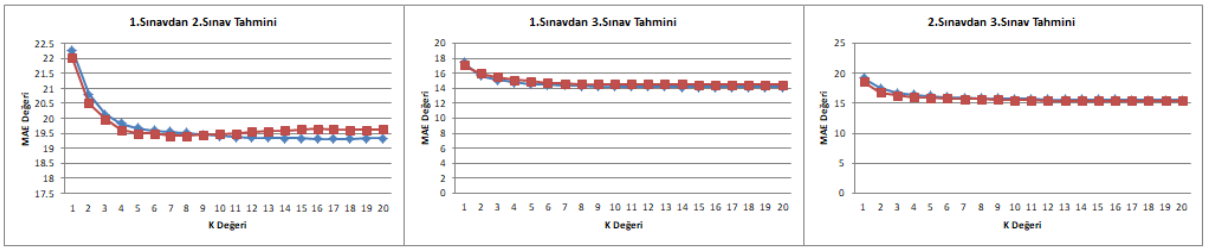
Gerçek verilerle elde edilen sonuçlada olduğu gibi sentetik verilerle elde edilen sonuçlara bakıldığında en iyi performansı doğrusal regresyon algoritması göstermiştir. Ayrıca sentetik verilerle elde edilen değerlerin, gerçek verilerle elde edilen değerlere göre daha iyi olduğu görülmüştür.

Ancak KNN ve karar ağacı algoritmalarında parametre kullanılmadığından dolayı eğitim aşamasında uygulanan k katmanlı çapraz doğrulama ile elde edilen analiz sonuçlarıyla, test verilerinin kullanılmasıyla elde edilen analiz sonuçları arasındaki farkın çok yüksek olduğu görülmüştür. Sistemin daha önce görmediği test verileri kullanıldığında tahmin performansında önemli düşmeler gözlenmiştir.

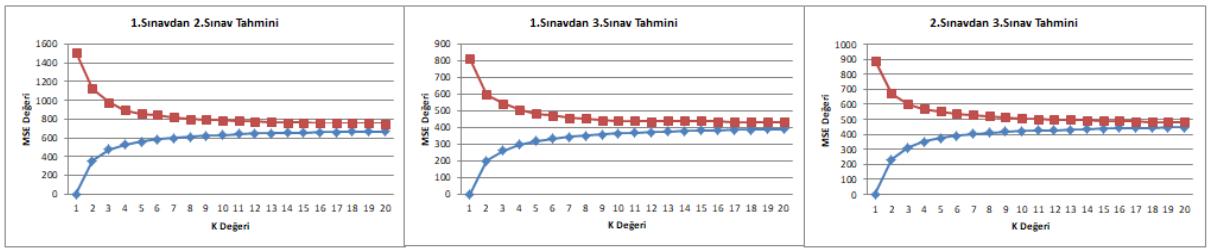
Şekil 4, Şekil 5, Şekil 6 ve Şekil 7'de KNN algoritmasındaki k değerine 1'den 20'ye kadar değerler verildiğinde performans ölçümlerindeki değişimler görülmektedir.

Şekil 8, Şekil 9, Şekil 10 ve Şekil 11'de ise karar ağacı algoritmasındaki derinlik değerine 1'den 20'ye kadar değerler verildiğinde performans ölçümlerindeki değişimler görülmektedir.

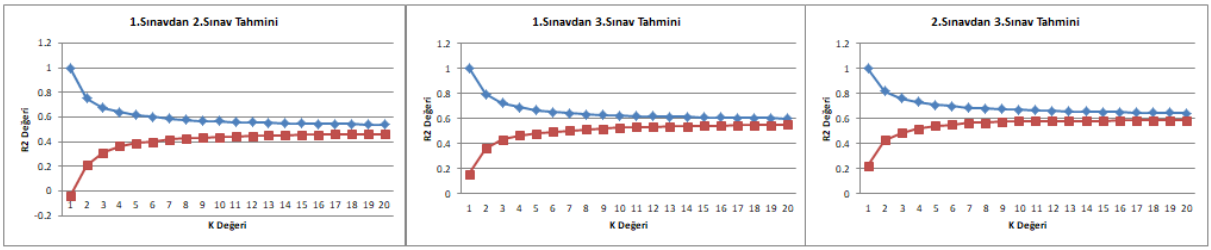
Şekillerde sistemin eğitim aşamasında k katmanlı çapraz doğrulama ile elde edilen analiz sonuçları mavi ile, test verilerinin uygulanması sonucu elde edilen sonuçlar ise kırmızı ile gösterilmiştir.



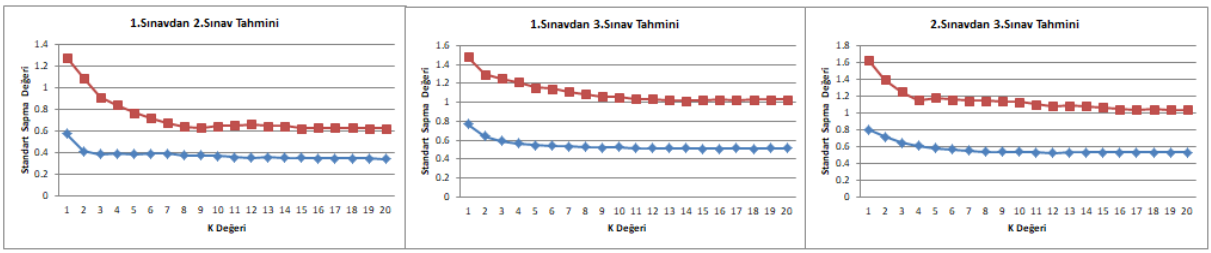
Şekil 4. KNN algoritmasındaki k değeri ile MAE ilişkisi (Relationship between k value and MAE in KNN algorithm)



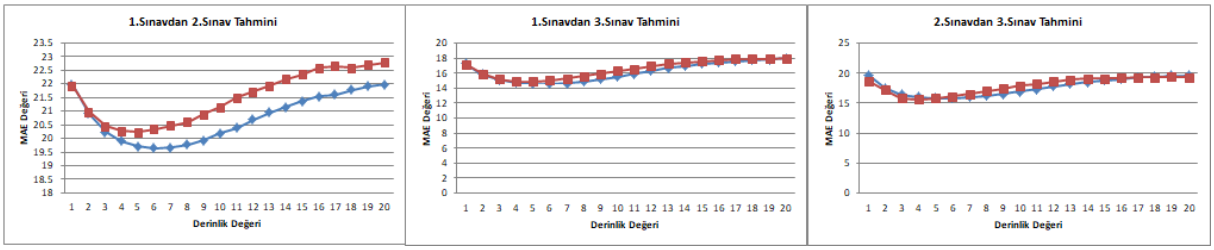
Şekil 5. KNN algoritmasında k değeri ile MSE ilişkisi (Relationship between k value in KNN algorithm and MSE)



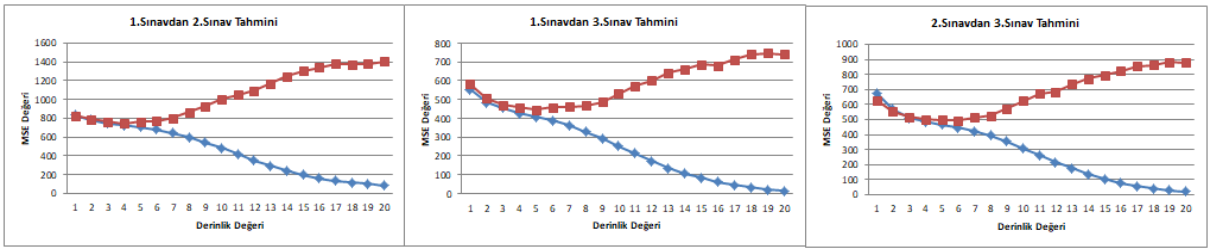
Şekil 6. KNN algoritmasında k değeri ile R^2 ilişkisi (Relationship between k value and R^2 in KNN algorithm)



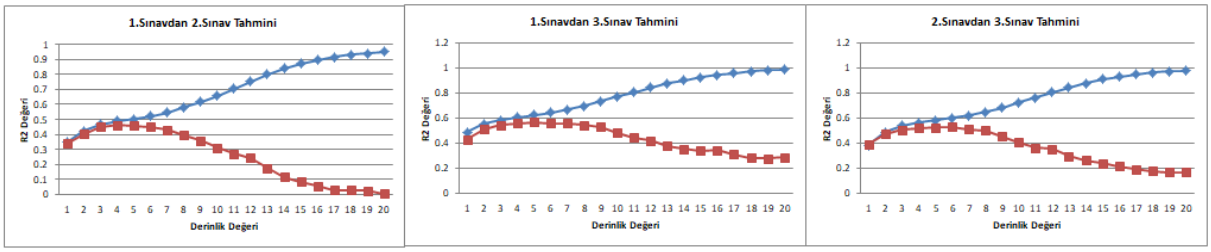
Şekil 7. KNN algoritmasında k değeri ile standart sapma ilişkisi (Relationship between k value and standard deviation in KNN algorithm)



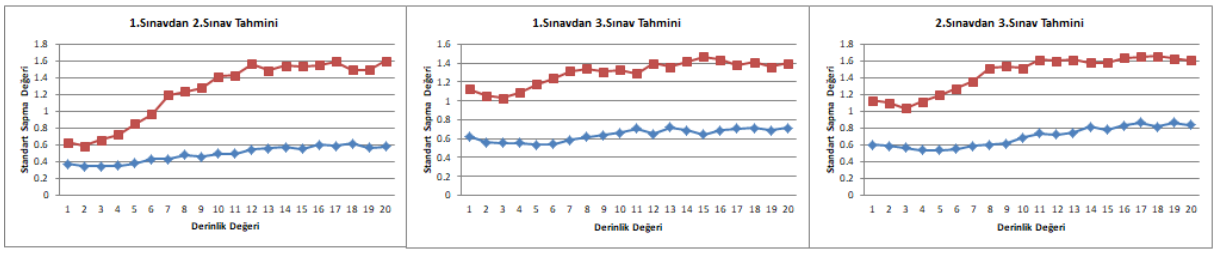
Şekil 8. Karar ağacı algoritmasında derinlik ile MAE ilişkisi (Relationship between depth and MAE in the decision tree algorithm)



Şekil 9. Karar ağacı algoritmasında derinlik ile MSE ilişkisi (Relationship between depth and MSE in the decision tree algorithm)



Şekil 10. Karar ağacı algoritmasında derinlik ile R^2 ilişkisi (Relationship between depth and R2 in decision tree algorithm)



Şekil 11. Karar ağacı algoritmasında derinlik ile standart sapma ilişkisi (Relationship between depth and standard deviation in decision tree algorithm)

KNN algoritmasındaki k parametresinin en iyi değerini bulabilmek için Python'da GridSearchCV yöntemi kullanılmıştır ve k=35 olarak bulunmuştur. Elde edilen değerlerin parametresiz elde edilen sonuçlara göre daha iyi olduğu görülmüştür. Tablo 8'de, k=35 uygulanarak elde edilen sonuçlar görülmektedir.

Tablo 8. KNN algoritmasında k=35 uygulanarak bulunan sonuçlar (The results found by applying k=35 in the KNN algorithm)

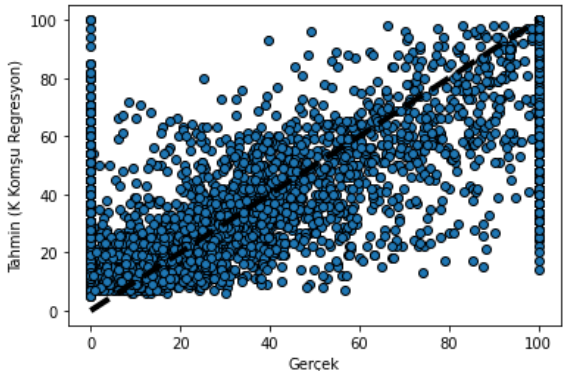
	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	19.828	746	0.47	0.603
1.sınavdan 3.sınav tahmini	14.452	427	0.589	1.027
2.sınavdan 3.sınav tahmini	15.587	470	0.557	1.016

Karar ağacı algoritmasından daha iyi performansı alabilmek ve aşırı uyum sorununu önlemek amacıyla derinlik (max_depth) ve düğümün bölünmeden önce gerekli örnek sayısı (min_samples_split) parametrelerinin en iyi değerlerini bulabilmek amacıyla Python'da GridSearchCV yöntemi kullanılmıştır. En iyi derinlik değeri 5 ve örnek sayısı değeri 18 olarak bulunmuştur. Elde edilen değerlerin parametresiz elde edilen sonuçlardan daha iyi olduğu görülmüştür. Tablo 9'da max_depth=35 ve min_samples_split=18 uygulanarak elde edilen sonuçlar görülmektedir.

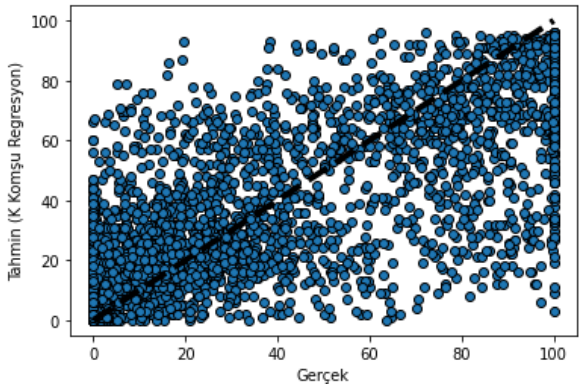
Tablo 9. Karar ağacı algoritmasında Max_depth=35 ve Min_samples_split=18 uygulanarak bulunan sonuçlar (The results found by applying Max_depth=35 and Min_samples_split=18 in the decision tree algorithm)

	MAE	MSE	R^2	Standart Sapma
1.sınavdan 2.sınav tahmini	20.227	762	0.456	0.81
1.sınavdan 3.sınav tahmini	14.713	446	0.57	1.138
2.sınavdan 3.sınav tahmini	15.813	492	0.532	1.171

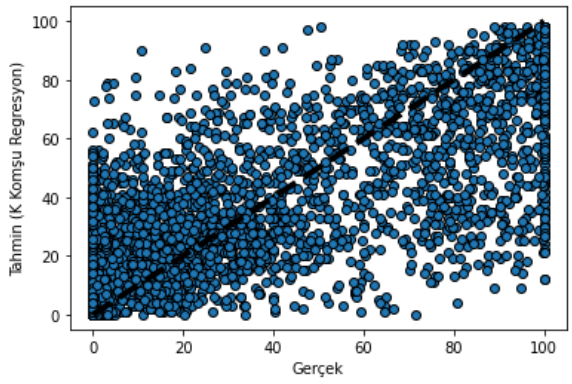
En iyi sonuçların elde edildiği KNN algoritmasında test verileriyle elde edilen grafikler Şekil 12, Şekil 13, Şekil 14'te görülmektedir.



Şekil 12. KNN ile 1. sınav verileri ile 2. sınav performansını tahmin grafiği (Prediction graph of 2nd exam performance with 1st exam data with KNN)



Şekil 13. KNN ile 1. sınav verileri ile 3. sınav performansını tahmin grafiği (Prediction graph of 3rd exam performance with 1st exam data with KNN)



Şekil 14. KNN ile 2. sınav verileri ile 3. sınav performansını tahmin grafiği (Prediction graph of 3rd exam performance with 2nd exam data with KNN)

4. Sonuçlar (Conclusions)

Bu çalışma ile öğrencilerin gelecekte başarılı yada başarısız olma durumlarını tahmin etmek yerine sayısal olarak gelecekte girecekleri sınav tahminleri yapılmak istenmiştir. Yapılan literatür taramasında genellikle öğrencilerin ileride başarılı olup olmama durumları tahmin edilmek istendiği görülmüştür.

Yapılan analiz sonuçlarına göre, çok az farklar olsada KNN algoritmasının diğerlerine göre daha iyi performans gösterdiği görülmüştür. Algoritmalarda en iyi parametre değerlerinin bulunarak eklenmesi performansı etkilemiştir. Özellikle karar ağacı algoritmasında sistemin eğitiminden sonra uygulanan çapraz doğrulamada R^2 değeri 1'e yakın çıkıyorken sistemin daha önce görmediği test veri kullanıldığında R^2 değeri çok düşmüştür. Yani tahmin başarı oranı düşmüştür. Aşırı uyum sorunu parametrelerin girilmesiyle giderilmiştir.

Sentetik verilerin kullanımı ile daha iyi sonuçlar elde edilmiştir. Gerçek verilerin yetersiz olduğu durumlarda sentetik verilerin performansı arttırdığı görülmüştür.

Bu çalışmada bir dersteki mevcut sınav verileriyle gelecekte girecekleri sınav kazanım performansları tahmin edilmiştir. Milli Eğitim Bakanlığı tarafından kullanılan e-okul (öğrenci veli bilgi sistemi), MEBBİS (MEB Bilişim Sistemleri), EBA (Eğitim Bilişim Ağı) gibi sistemlerde çok fazla veri işlenmeyi beklemektedir. Bu veriler ile öğrencinin tüm derslerinde aldığı puan dağılımlarına göre dersler arasındaki performans bağlantıları analiz edilebilir. Bir dersteki performans durumunun diğer derslere etkisi tahmin edilebilir. Gelecek hakkındaki bu tahminlere göre öğrenciye rehberlik edilebilir ve öğrenciye göre ders içerikleri planlanabilir. Böylece öğrencilerin gelecekteki başarısızlıkları bugünden önlenebilir ve eğitim kalitesi arttırılabilir.

Kaynaklar (References)

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

Adak, M. F., Yumusak, N., & Taskin, H. (2016). An elective course suggestion system developed in computer engineering department using fuzzy logic. In 2016 International Conference on Industrial Informatics and Computer Systems (CIICS) (pp. 1-5). IEEE.

Aghalarova, S. & Bozkurt Keser, S. (2022). Öğrencilerin Akademik Performanslarının Tahmin Edilmesi Için Automl Tekniğinin Uygulanması . El-Cezeri , 9 (2) , 394-412 . Doi: 10.31202/Ecjse.946505.

Akgün, K. & Bulut Özek, M. (2020). Eğitsel Veri Madenciliği Yöntemi İle İlgili Yapılmış Çalışmaların İncelenmesi: İçerik Analizi . Uluslararası Eğitim Bilim Ve Teknoloji Dergisi , 6 (3) , 197-213 . Doi: 10.47714/Uebt.753526.

Altunkaynak, A. , Başakın, E. E. & Kartal, E. (2020). Dalgacik K-En Yakin Komşuluk Yöntemi İle Hava Kirliliği Tahmini . Uludağ Üniversitesi Mühendislik Fakültesi Dergisi , 25 (3) , 1547-1556 . Doi: 10.17482/Uumfd.809938

Aydemir, E. , Kaysi, F. & Gülseçen, S. (2019). Üniversite Öğrencilerinin Türk Dili Dersi Sınav Sonuçlarının Sınava Hazırlık Düzeylerine Göre Tahminlenmesi . Alphanumeric Journal , 7 (2) , 351-356 . Doi: 10.17093/Alphanumeric.583502

Can, Ş. , Özdil, T. & Yılmaz, C. (2018). Üniversite Öğrencilerinin Ders Başarisini Etkileyen Faktörlerin Lojistik Regresyon Analizi İle Tahminlenmesi . International Review Of Economics And Management , 6 (1) , 28-49 . Doi: 10.18825/Iremjournal.349984.

Gök, M. (2017). Makine Öğrenmesi Yöntemleri İle Akademik Başarinin Tahmin Edilmesi . Gazi University Journal Of Science Part C: Design And Technology , 5 (3) , 139-148.

Güner, N. & Çomak, E. (2011). Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi . Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi , 17 (2) , 87-96.

Güvenç, E. , Sakal, M. , Çetin, G. & Özkaraca, O. (2022). Öğrencilerin Dersteki Niteliklerinin Makine Öğrenmesi Teknikleri Kullanılarak Sınıflandırılması . Düzce Üniversitesi Bilim Ve Teknoloji Dergisi , 10 (3) , 1359-1371 . Doi: 10.29130/Dubited.1017202.

İçeli, N. (2012). Veri Madenciliği Yöntemi İle Divriği Nuri Demirağ Meslek Yüksekokulu Öğrencilerinin Temel Bilgisayar Dersine Ait Başari Analizi Uygulamasi . Mesleki Bilimler Dergisi (Mbd) , "2012 Yılı Cilt:1 Sayı:1 (Syf 18-37)" , 18-37.

Kumari, S., Siwach, V., Singh, Y., Barak, D., & Jain, R. (2022). A Machine Learning Centered Approach For Uncovering Excavators' Last Known Location Using Bluetooth And Underground Wsn. Wireless Communications And Mobile Computing, 2022.

Mardikyan, S. (2005). İlişki Analizinde Varsayımlardan Sapmaların Belirlenmesi Ve Çözümlenmesine Yönelik Bir Bilgisayar Programı Geliştirilmesi, Doktora Tezi, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Meb, (2022). Http://Meslek.Eba.Gov.Tr/?P=Ders-Bilgi-Formu&Tur=Mtal, 20.02.2022

Özlen T. (2022). Servikal Kanserlerin Teşhisinde Kullanilan Makine Öğrenmesi Algoritmalarinin Karşilaştirmali Analizi, Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi, Lisansüstü Eğitim Enstitüsü, İstanbul

Özlüer Başer, B. , Yangın, M. & Sarıdaş, E. S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması . Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi , 25 (1) , 112-120 . Doi: 10.19113/Sdufenbed.842460

Şengür, D. & Tekin, A. (2014). Öğrencilerin Mezuniyet Notlarının Veri Madenciliği Metotları İle Tahmini . Bilişim Teknolojileri Dergisi , 6 (3) , 7-16.