|  |  |  |
| --- | --- | --- |
|  | Tekirdağ Namık Kemal Üniversitesi  Fen Bilimleri Enstitüsü    <http://fbe.nku.edu.tr> |  |

**Dersin Adı: Yapay Zekâ Teknikleri ve Makine Öğrenmesi**

**Danışman Unvanı Adı Soyadı: Dr. Öğr. Üyesi Erkan Özhan**

**Öğrenci Adı Soyadı: Fatma Elvan Sarıaydın**

**Öğrenci Numarası: 1248132154**

1. **GİRİŞ**

Günümüzde, eğitim alanında öğrenci performansının tahmini, başarıyı arttırmak için önemli bir konudur. Makine öğreniminde regresyon algoritmaları, öğrencilerin akademik başarısını sayısal olarak modellemek için sıkça kullanılmaktadır. Bu çalışmada, öğrenci performansını analiz etmek amacıyla on farklı regresyon algoritması uygulanmış ve sonuçları karşılaştırılmıştır. Böylece, performans üzerinde etkili faktörler belirlenerek eğitim süreçlerinin iyileştirilmesinde katkı sağlanması hedeflenmiştir.

* 1. **Çalışmanın Amacı ve Kapsamı**

Bu çalışmanın temel amacı, bir web uygulamasında kullanılmak üzere bir yapay zekâ modeli geliştirmektir. Çalışmada on adet makine öğrenmesi algoritması kullanılmıştır. Sonrasında bu modellerden en iyi performansa sahip model tespit edilmiştir. Bu model üzerinde öznitelik önemi değerlendirmesi yapılmış ve sonrasında model yeniden eğitilmiştir. Son olarak, model geliştirmesi tamamlanıp web uygulamasına hizmet etmek üzere hazır hale getirilmiştir.

1. **VERİ SETİNİN İNCELENMESİ**

Veri seti, kaggle.com isimli siteden alınmıştır [1]. Veri seti, öğrencilerin çeşitli demografik özellikleri ve okul yaşamlarına dair bilgiler ile matematik, okuma ve yazma notlarını içermektedir. Python kodları ile veri seti üzerinde keşifsel veri analizi (EDA) yapılmıştır.

* 1. **Keşifsel Veri Analizi**

Veri seti, Python kodunda df isimli değişkene atanmıştır ve 1000 adet satır ve 8 adet sütundan oluşmaktadır ancak 7’ye düşürülmüştür. Model eğitimi ve uygulama için “race/ethnicity” isimli sütun net bilgi sağlamadığı için kullanılmamıştır. İlgili sütun düşürüldükten sonra sütun isimlerinde Türkçeleştirmeye gidilmiştir. Bu işlemden sonra df.info() fonksiyonu çağrıldığında, veri setinin durumunun Şekil 2.1’deki gibi olduğu görülür. Şekilden de anlaşılacağı üzere, 4 adet sütun nesne veri tipine, geri kalan 3 adet sütun tamsayı veri tipine sahiptir. Verileri analiz ederken görselleştirme araçları olan seaborn ve matplotlib kütüphaneleri kullanılmıştır.



metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 1 Veri seti bilgisi

df.describe() fonksiyonu ile matematik, okuma ve yazma notlarının minimum, maksimum, standart sapma, ortalama vb. değerleri elde edilmiştir. Şekil 2.2’de de görüleceği üzere, öğrencilerin ortalama matematik notu 66 iken ortalama okuma notu 69.1 ve ortalama yazma notu 68’dir.



metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 2 Öğrencilerin matematik, okuma ve yazma notlarının matematiksel değerleri

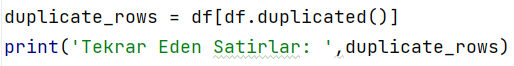
df.isnull().sum() fonksiyonu ile veri setinde eksik değer olup olmadığının kontrolü yapılmıştır. Veri setinde eksik bir değere rastlanmamıştır. Şekil 2.3’de ilgili fonksiyonun çıktısı gösterilmiştir.





Şekil 2. 3 Eksik veri kontrolü

Şekil 2.4’te görüleceği gibi veri setinde mükerrer veriye rastlanmamıştır.





Şekil 2. 4 Mükerrer veri kontrolü

* + 1. **Kategorik Değişkenlerin Dağılımı**

Kategorik değişkenler cinsiyet, ebeveynEgitim, ogleYemegi ve testHazirlik değişkenleridir. Bu değişkenlerin dağılımları sırasıyla Şekil 2.5, Şekil 2.6, Şekil 2.7 ve Şekil 2.8’de gösterilmiştir.

Şekil 2.5’den anlaşılacağı üzere, veri setinde 518 adet kadın, 482 adet erkek öğrenci bulunmaktadır. Veri setinin %51.8’ini kız öğrenciler oluşturuken %48.2’sini erkek öğrenciler oluşturmuştur.

metin, ekran görüntüsü, dikdörtgen, kare içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 5 Öğrencilerin cinsiyet dağılımı

metin, ekran görüntüsü, dikdörtgen, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 6 Öğrencilerin ebeveynlerinin eğitim durumu

Şekil 2.6’dan anlaşılacağı üzere, öğrencilerin %22.6’sının ebeveynlerinden en az biri bir üniversiteyi kazanmış ancak bitirememiş, %22.2’sinin ebeveynlerinden en az biri önlisans mezunu, %19.6’sının ebeveynlerinden en az biri lise mezunu, %17.9’unun ebeveynlerinden en az biri liseyi bitirememiş, %11.8’inin ebeveynlerinden en az biri lisans mezunu ve kalan %5.9’unun ebeveynlerinden en az biri yüksek lisans mezunudur.

metin, ekran görüntüsü, dikdörtgen, ekran, görüntüleme içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 7 Öğrencilerin öğle yemeği ücreti ödeme durumu

Şekil 2.7’de görüldüğü gibi, öğrencilerin %64.5’i öğle yemeklerini tam ücretli bir şekilde yerken, %35.5’i ücretsiz veya indirimli öğle yemeği yemiştir. Buradan öğrencilerin ekonomik durumu anlaşılabilir.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 8 Sınava hazırlık testi

Şekil 2.8’de, sınava hazırlık testini tamamlayan öğrenci sayısının 358, tamamlamayan öğrenci sayısının 642 olduğu görülür. Öğrencilerin %64.2’si hazırlık testini çözmemişken, %35.8’i bu testi tamamlamıştır.

* + 1. **Sayısal Değişkenlerin Dağılımı**

Veri setindeki sayısal değişkenler matematik, okuma ve yazma skorlarıdır. Şekil 2.9, bir histogramdır ve öğrencilerin matematik notlarının dağılımını göstermektedir. Anlaşılacağı üzere, öğrencilerin matematik notları 60 ile 80 arasında yoğunlaşmıştır. Bu da demektir ki, öğrencilerin aldığı matematik notları daha çok 60-80 arasındadır.

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 9 Öğrencilerin matematik notlarının histogramı

Öğrencilerin okuma notlarının histogramı Şekil 2.10’da gösterilmiştir. En sık alınan okuma notu 72 ile 76 arasındadır.

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 10 Öğrencilerin okuma notlarının histogramı

diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

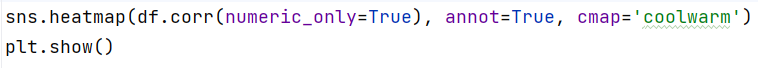
Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 11 Öğrencilerin yazma notlarının histogramı

Son olarak, Şekil 2.11’de öğrencilerin yazma notlarının histogramı verilmiştir. En sık alınan yazma notunun okuma notunda olduğu gibi 72 ile 76 arasında bir not olduğu görülür.

* + 1. **Korelasyonlar**

Seaborn kütüphanesi yardımıyla bir ısı haritası oluşturulmuştur. Şekil 2.12 kodu ve haritayı göstermektedir.



metin, ekran görüntüsü, dikdörtgen, kare içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 12 Notlara göre ısı haritası

Haritaya bakıldığında, matematik notu ile okuma notu arasında 0.82 değerinde, yazma notu ile 0.8 değerinde güçlü pozitif korelasyon olduğu görülür. Aynı şekilde, okuma notu ile yazma notu arasında 0.95 değerinde çok güçlü pozitif korelasyon olduğu görülür. Çıkarılacak sonuç, matematik notu, okuma notu veya yazma notunun değeri arttıkça diğer notların da arttığıdır.

* + 1. **Gruplama ile Veri Analizi**

Sınava hazırlık testini tamamlamış öğrenciler ile tamamlamamış öğrencilerin ortalama matematik notunu Şekil 2.13 ortaya koymaktadır. Şekilden yola çıkarak, sınava hazırlık testini tamamlayan öğrencilerin matematik notlarının daha yüksek olduğu sonucuna ulaşılabilir.



metin, yazı tipi, ekran görüntüsü, beyaz içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 13 Sınava hazırlık testi ile matematik notu arasındaki ilişki

Benzer şekilde, ebeveynlerinin eğitim seviyesine öğrencilerin ortalama matematik notları Şekil 2.14’deki gibi bir dağılım göstermiştir. Öğrencilerin ebeveynlerinin eğitim seviyesi arttıkça matematik notlarının arttığı (lise mezunu ve liseyi bitiremeyen ebeveynler hariç) gözlenmiştir.



metin, yazı tipi, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 14 Ebeveyn eğitim seviyesi ile matematik notu arasındaki ilişki

Son olarak, Şekil 2.15’te öğle yemeğini standart ücretten ödeyen öğrenciler ile ücretsiz bir şekilde veya indirimli alan öğrencilerin ortalama matematik notlarının karşılaştırması yapılmıştır. Sonuçlar standart ücretten ödeyen öğrencilerin daha yüksek matematik notunu aldığını göstermiştir. Buradan yola çıkarak ekonomik durumu daha düşük olan öğrencilerin daha düşük matematik notu aldığı söylenebilir.



metin, yazı tipi, ekran görüntüsü, beyaz içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 2. 15 Öğle yemeği ücreti ile matematik notu arasındaki ilişki

1. **KULLANILAN REGRESYON ALGORİTMALARI**

Veri seti ön işleme aşamasından geçtikten sonra %80’i eğitim verisi, %20’si test verisi olmak üzere ayrılmıştır. Modellerin performansları Ortalama Mutlak Hata (MAE), Hataların Karesinin Ortalamasının Karekökü (RMSE) ve R-Kare metrikleri göz önünde bulundurularak değerlendirilmiş olup en iyi modelin doğrusal regresyon modeli olduğu ortaya konmuştur.

* 1. **Lineer (Doğrusal) Regresyon Algoritması**

Doğrusal regresyon modelinin MAE değeri 4.6019, RMSE değeri 5.6332 ve R2 değeri 0.8732 olarak hesaplanmıştır.

* 1. **Lasso Regresyon Algoritması**

Lasso regresyon modelinin MAE değeri 4.6019, RMSE değeri 5.6332 ve R2 değeri 0.8732 olarak hesaplanmıştır.

* 1. **Ridge Regresyonu Algoritması**

Ridge regresyon modelinin MAE değeri 4.6019, RMSE değeri 5.6332 ve R2 değeri 0.8732 olarak hesaplanmıştır.

* 1. **Elastic Net Algoritması**

Elastic Net regresyon modelinin MAE değeri 4.6021, RMSE değeri 5.6333 ve R2 değeri 0.8732 olarak hesaplanmıştır.

* 1. **Karar Ağacı Regresyonu**

Karar ağacı regresyonu modelinin MAE değeri 6.8750, RMSE değeri 8.4808 ve R2 değeri 0.7128 olarak hesaplanmıştır. Bu değerlerle karar ağacı regresyon modeli en düşük performanslı model olmuştur.

* 1. **Random Forest Regresyonu**

Random forest regresyon modelinin MAE değeri 4.9050, RMSE değeri 6.0171 ve R2 değeri 0.8554 olarak hesaplanmıştır.

* 1. **Gradient Boosting Regresyonu**

Gradient boosting regresyon modelinin MAE değeri 4.8125, RMSE değeri 5.9592 ve R2 değeri 0.8581 olarak hesaplanmıştır.

* 1. **XGBoost Regresyonu**

XGBoost regresyonu modelinin MAE değeri 5.2713, RMSE değeri 6.7773 ve R2 değeri 0.8165 olarak hesaplanmıştır.

* 1. **Destek Vektörü Regresyonu (SVR)**

Destek vektörü regresyonu modelinin MAE değeri 4.6870, RMSE değeri 5.7253 ve R2 değeri 0.8691 olarak hesaplanmıştır.

* 1. **KNeighbors Regresyonu**

KNeighbors regresyonu modelinin MAE değeri 6.8160, RMSE değeri 8.4193 ve R2 değeri 0.7169 olarak hesaplanmıştır.

* 1. **Modellerin Performansları**

Modellerin performansları değerlendirildiğinde en iyi performansı doğrusal regresyon modelinin verdiği görülmüştür. Şekil 3.1’de görüldüğü üzere, bu veri setinde matematik notunu tahmin etmek için kullanılabilecek en iyi model lineer regresyon modelidir.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 3. 1 Modellerin performansları

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 3. 2 Modellerin performans metrikleri

* 1. **Öznitelik Önem Dereceleri (Feature Importance)**

Lineer regresyon modeli, bu proje için en iyi performans gösteren model olduğundan, bu model üzerinden öznitelik önem derecesi analizi yapılıp, model üzerinde en az etkiye sahip, modelin tahmin yaparken en az ağırlık verdiği öznitelikler belirlenmiştir. Bu öznitelikler eğitim ve test verileri içine dahil edilmeden model tekrar eğitilmiş ve başarısı hesaplanmıştır. Şekil 3.3’te özniteliklerin önem derecelerini gösteren grafik verilmiştir.

metin, ekran görüntüsü, diyagram, paralel içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil 3. 3 Özniteliklerin önem dereceleri

Yalnızca okumaSkoru elendiği ve model yeniden eğitildiğinde model skorlarında çok az iyileşme görülmüştür. okumaSkoru elenmeden önceki MAE skoru 4.6768, RMSE skoru 5.7860 ve R2 skoru 0.8663’tür. okumaSkoru elendikten sonra MAE skoru 4.6764, RMSE skoru 5.78515, R2 skoru 0.8663’tür. Sadece MAE ve RMSE skorlarında çok az miktarda iyileşme görülmüştür. Bu sebeple, okumaSkoru elenmeyip modelde kullanılmıştır.

1. **SONUÇ VE ÖNERİLER**

Modeller arasından doğrusal regresyon modeli en iyi performansa sahip olduğundan, web uygulaması için bu model kullanılmıştır.

**KAYNAKLAR**

[1] https://www.kaggle.com/datasets/allexanderspb/studentsperformance