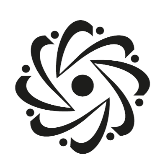
****

**Bursa Teknik Üniversitesi**

**Veri Madenciliğine Giriş**

*Proje Raporu*

**Ad:** Eren

**Soyad:** KÖSE

**Öğrenci Numarası:** 22360859075

**Proje Hakkında**

* **Açıklama:** Öğrencilerin 2 yıllık Matematik, Bilim, İngilizce dersleri için toplanan not verileri, cinsiyetleri, 2017/18 öğretim yılından hesaplanan yaşları, önerilen yıl/sınıfları, okula kabul yılları, önceki okullarında okudukları müfredatları, şu anki müfredatları, önceki okuldaki yıl/sınıfı, okula giriş sınavı matematik notu, okula giriş sınavı bilim notu, okula giriş sınavı ingilizce notu bilgileri bu veri setinde yer almaktadır. Bu veriler kullanılarak öğrencilerin “High”, “Medium”, “Low” şeklinde 3 farklı seviyeye ayrılması amaçlanmıştır. Bu amaç doğrultusunda öğrencilerin not ortalamalarının 85 ve üzeri olması durumu “High” seviye öğrenci, 75 ve 85 aralığı “Medium” seviye öğrenci, 75 altı ise “Low” seviye öğrenci olarak tanımlanmıştır.
* **Kullanılan Veri Seti:** <https://www.sciencedirect.com/science/article/pii/S235234092100192X>
* **Referans Alınan Makale:** <https://link.springer.com/article/10.1007/s43926-022-00023-0>
* **Kullanılan Yöntem:** Random Forest Classifier
* **Github Reposu:** <https://github.com/erennkose/data-mining-project>
* **Tanıtım Videosu:** eren

**Modelim Hakkında Bilgiler**

Projemde geliştirdiğim model öncelikle veri setini içeri aktarır. Bu veri setindeki boşluk, tırnak gibi gereksiz karakterlerin temizliği yapılır. Bu temizlik sayesinde model performansı yükselmektedir. Veri setinden çekilen verilerde kullanılan öznitelikler “Gender”, “Age as of Academic Year 17/18”, “Previous Curriculum (17/18)2”, “Math20-1”, “Science20-1”, “English20-1”, “Math20-2”, “Science20-2”, “English20-2”, “Math20-3”, “Science20-3”, “English20-3” öznitelikleridir. dropna() fonksiyonu ile eksik değer içeren satırlar kaldırılmıştır. Modelimin hedef değişkeni ise öğrencilerin kategorik seviyeleridir. Belirtilen öznitelikler kullanılarak öğrenciler belirli seviyelere atanmıştır. Tahmin için kullanılacak bağımsız değişkenler ve bu bağımsız değişkenlere bağlı olacak öğrenci seviyesi belirtilmiştir. Bunların ardından kategorik değişkenler (Gender, Previous Curriculum gibi) one-hot encoding yöntemiyle binary formata çevrilir. Ayrıca öğrencilerin not değerleri yaş gibi değerlere kıyasla değer olarak çok daha baskın olacağından bunu önleyip gerçekçi bir deneyim elde etme amaçlı öğrenci notlarına normalizasyon işlemi uygulandı. Sonrasında ise veri seti % 70 eğitim (train), % 10 doğrulama (validation), %20 test verisi olacak şekilde ayrılır. Stratify parametresi ile sınıfların orantılı dağılması sağlanır. Bu ayrılan veriler kullanılarak 4 farklı model (50, 100, 200 ve 500 ağaçla) eğitilir. Bunlar arasından genel olarak en dengeli sonuçları veren 100 ağaçlık modelin sonuçlarını paylaşacağım fakat diğer modellerin de değerleri aşağıdaki görselde bulunmaktadır.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil Modellerin Değerleri

**Eğittiğim Modelin Sonuçları**

Projemde yaptığım araştırmalar ve yapay zekalara danışmalarım sonucunda veri setimde Random Forest Classifier yönteminin amacıma daha uygun bir yöntem olduğunu keşfettim ve bu yöntemi seçtim. Bu yöntem sayesinde makalede bulunan değerlerin biraz üstüne çıkmayı başardım. Bunun bir diğer sebebi de gürültü yaratabilecek ve model için mantıksız olan özniteliklerin kullanılmamasıdır. Aşağıda verilen ROC eğrileri sonucunda görüldüğü üzere train verilerinin başarısı kadar test verileri de başarılı şekilde tahmin edilmiş durumda. Bu da overfittingin modelimde söz konusu olmadığını gösteriyor.

Şekil 4’te görüldüğü üzere karmaşıklık matrisinde öğrenciler gayet başarılı şekilde gerekli seviyelere ayrılıyor. Veri seti büyüklüğüne kıyasla ufak seviyede hatalarla karşılaşılıyor.

Şekil 5’e baktığımızda da modelimin performans metriklerini görebiliyoruz. Bu performans metriklerinden Accuracy değerleri modelin tahminlerinin genel olarak doğru olduğunu belirtiyor. Bu, tüm sınıflarda modelin başarılı olduğunu ortalama olarak gösteriyor. AUC değerleri ise modelin sınıfları ayırt etme kabiliyetinin çok yüksek olduğunu belirtiyor. ROC eğrisi altında kalan alan çok yüksek olduğu için modelin sınıfları ayrıştırma gücü de o kadar yüksek oluyor. Precision, Recall ve F-1 Skoru değerlerinin biraz diğerlerine kıyasla düşük kalmasının sebebi ise veri setimde bulunan verilerin dengesiz dağılımından kaynaklanıyor. Bundan dolayı bu değerler accuracy ve AUC değerlerine kıyasla biraz düşük kalıyor fakat ona rağmen kabul edilebilir durumda.

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

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil ROC Eğrisi (Train)

metin, çizgi, diyagram, ekran görüntüsü içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil ROC Eğrisi (Test)

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

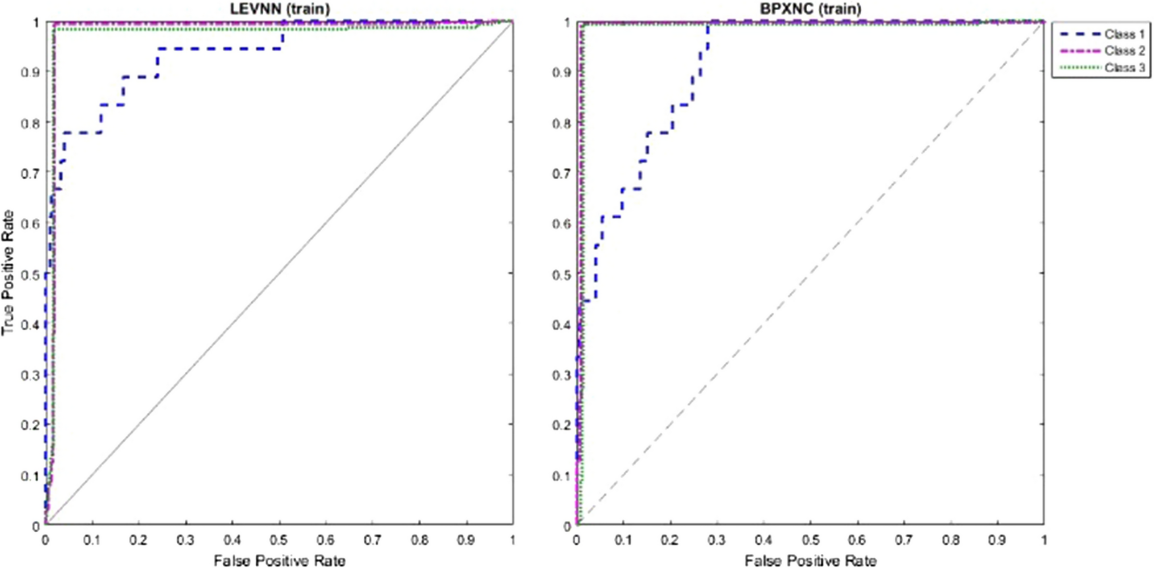
Şekil Karmaşıklık Matrisi

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

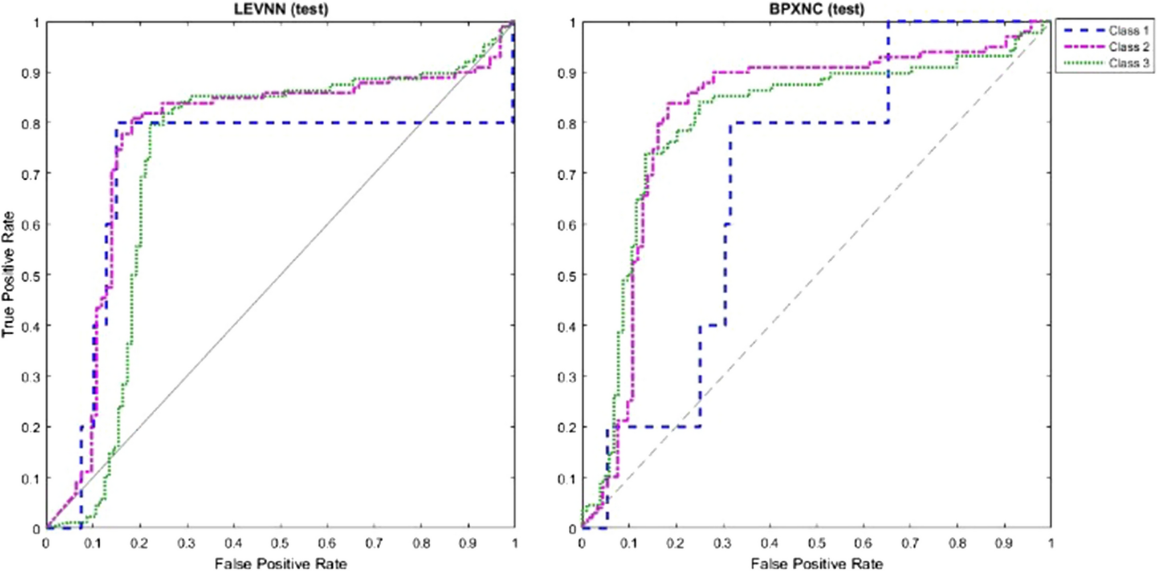
Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Şekil Performans Metrikleri

**Referans Aldığım Makaledeki Modelin Sonuçları**



Şekil Makale ROC Eğrisi (Train)



Şekil Makale ROC Eğrisi (Test)