Tarih: 06.07.2025

İsim: Zeynep Nazlı Yiğit

YAPAY ZEKA AKADEMİSİ BİTİRME PROJESİ

Problem Tanımı:

Bu projede, Student Performance veri seti kullanıldı. Veri seti, iki Portekiz okulundan öğrencilerin demografik, sosyal ve akademik bilgilerini içeriyor. Her öğrenciye ait dönem içi sınav notları ve final notlarını da içeriyor.

Bu projede, hedef değişkenimiz öğrencilerin yıl sonu notu olan G3'tür. Öğrencilerin hedef değişkeni olan yıl sonu notlarını (G3) tahmin edebilecek regresyon modelleri kuruldu. Böylece öğrencilerin demografik, sosyal ve akademik bilgileri kullanılarak yıl sonu notları tahmin edilmeye çalışıldı. Hedef değişkeni öğrencilerin yıl sonu notu olan G3'tür. Oluşturulacak tahmin projesi sayesinde öğrenci sınav notlarını artırmaya yönelik kararlar alınabilir.

Model Sonuçları:

Öncelikle veri analizi ve eksik veri kontrolü yapıldı. Ardından hedef değişkeninin dağılımına bakıldı. Sayısal değerler boxplot kullanılarak görselleştirildi ve aykırı değerlere bakıldı. Sayısal değerlerin birbiri ile ilişkisini incelemek için korelasyon matrisi de oluşturuldu. Kategorik verilerin dağılımını gözlemlemek için histogram grafiği kullanıldı. Aykırı değer analizi yapılarak aykırı değerler winsorizing ile kırpıldı. Kategorik verileri sayısal verilere dönüştürmek için LabelEncoder kullanıldı. Ardından oluşturulan bu sayısal değerler StandardScaler yöntemi ile ölçeklendirildi.

Ardından linear regression, random forest, gradient boosting ve SVR modelleri oluşturuldu ve grafikler ile başarı metrikleri bulundu. Bu modellerde en başarılı tahmin linear regression tarafından verilmiştir. Bütün metriklerde de en başarılı sonuçları vermiştir. Ancak bu modelde hatalar sıfır olduğu için aşırı öğrenme (overfitting) vardır. Hemen ardından gradient boosting başarılı sonuçlar vermiştir. Bu model yüksek tahmin ve gerçekçi sonuçlara sahiptir. En güvenilir ve kullanışlı modeldir. Random forest'ın hata oranı düşük ama gradient boosting'den fazladır. SVR'de ise yetersiz öğrenme gözlemlenmiştir çünkü kötü bir tahmin oranına sahiptir ve R^2 değeri düşüktür.

Çıkarımlar:

Oluşturulan modellerin büyük bir kısmı yıl sonu notlarını yüksek doğruluk oranıyla tahmin edebilmiştir. Veri seti çok sayıda kategorik ve sayısal parametre içermektedir. Bu parametrelerden G1 ve G2 yani öğrencilerin dönem içi sınav notları yıl sonu notlarını tahmin edebilmede etkili değişkenlerdendir. Ders çalışma süresi, anne ve baba eğitim düzeyi gibi parametreler sınav notları üzerinde yüksek bir etkide bulunmamıştır.

Farklı regresyon modelleri (Lasso, Ridge gibi) de kullanılarak daha iyi sonuçlar elde edilebilir ve karşılaştırmalar yapılabilir. Veri seti geliştirilerek de daha iyi sonuçlar elde edilebilir. Son olarak, hiperparametre optimizasyonu ile parametreler daha verimli olarak ayarlanabilir.