End to End Data Science Proje Dökümanı

Bu projede sigara içen ve içmeyen insanlardan toplanan verileri analiz ederek bir makine öğrenmesi projesi yaptık. Projede başta veriler arası ilişkileri görselleştirip gözle analiz ettikten sonra verimin yüzde 80’lik kısmı ile veriyi uygun kalıba soktuktan sonra eğittim ve ardından regresyon modelleri ile doğruluğumu ölçtükten sonra en yüksek doğruluk bulunan regresyon seçimimden sonra GridSearch ile doğruluğumu daha da yükselttim. Son olarak da modelin değerlendirmesini yaptım.

Projede istenen adımlar:

**1.** **Google Colaboratory Dosyası Açılması** ● Projenizin .ipynb uzantılı olmasına özen gösteriniz. ● Projenizdeki detayları açıklayan yorum satırları olmasına dikkat ediniz. ● Proje teslim edilirken bu .ipynb uzantılı dosyasının hücreleri çalıştırılmış ve sonuçları görünecek şekilde teslim ediniz.

**2.** **Gerekli Kütüphanelerin Eklenmesi** ● Projede kullanılacak olan kütüphanelerin Colab ortamına ekleyiniz. ● Pandas, numpy, Seaborn, Matplotlib ve Sklearn kütüphanelerini ekleyiniz.

**3.** **Keşifsel Veri Analizi Yapılması** ● Veriyi inceleyerek, analiz ederek veriden anlamlı sonuçlar çıkarınız. ○ Bmi(Vücut Kitle İndeksi)’nin dağılımını inceleyiniz ○ “smoker” ile “charges” arasındaki ilişkiyi inceleyiniz ○ “smoker” (Sigara tüketen) ile “region”(Bölge) arasındaki ilişkiyi inceleyiniz. ○ “bmi” ile “sex”(Cinsiyet) arasındaki ilişkiyi inceleyiniz. ○ En çok “children”’a sahip “region”’ı bulunuz. ○ “Age” ile “bmi” arasındaki ilişkiyi inceleyiniz. ○ “bmi” ile “children” arasındaki ilişkiyi inceleyiniz. ○ “bmi” değişkeninde outlier var mıdır? İnceleyiniz. ○ “bmi” ile “charges” arasındaki ilişkiyi inceleyiniz. ○ “region”, “smoker” ve “bmi” arasındaki ilişkiyi bar plot kullanarak inceleyiniz. ● Veriyi incelerken veri görselleştirme tekniklerini olabildiğince kulanmaya çalışınız. ● Analizlerden çıkardığınız anlamları yorum satırı olarak ekleyiniz.

**4.** **Veri Ön İşleme Yapılması** ● Bu kısımda elinizde olan veriyi model eğitmek için hazır hale getiriniz. ● Kategorik değişkenleri düzenlemek için Label ve One-Hot Encoding tekniklerini kullanınız. ● Veri setinizi X\_train,X\_test, y\_train, y\_test olacak şekilde bölüştürünüz. ● Veri setini normalize ederek ölçekleyiniz.

**5.** **Model Seçme** ● Birkaç regresyon modeli seçiniz bunları ön işleme yapılan veri ile eğitiniz ● Seçilen modellerin performanslarını çapraz doğrulama kullanarak inceleyiniz. ● En iyi performans gösteren modeli seçiniz

**6. Hiper-parametre Optimizasyonu** ● Bir önceki adımda seçilen modelin hiper-parametrelerinin optimize ediniz. ● Grid Search ile parametreleri optimize ediniz.

**7.** **Modeli Değerlendirme** ● Regresyon modeli değerlendirme metriklerini kullanarak optimize edilmiş olan modelin değerlendirmesini yapınız. (Ör. Mean Squared Error, Mean Absolute Error vb.)

Alp Erem