**BİL 210 – Veri Bilimi ve Makine Öğrenmesine Giriş**

**Dönem Projesi**

1. **Proje Detayları**

Bir alışveriş merkezinde teknolojik ürünler satan mağazaya ait tablet ürünlerinin özellikleri ve fiyatları **“tablet.csv”** olarak sizlerle paylaşılmıştır. Mağaza müdürü vitrinlerinde tablet ürünlerini en doğru şekilde sergileyebilmek için 2.000 adet tableti fiyat/performans durumlarına göre “Ucuz”, “Normal”, “Pahalı”, “Çok Ucuz” olarak etiketlemiştir.

Mağazasına yeni gelecek tablet ürünlerini etiketleyerek vakit kaybetmek istemeyen IT ekibi mevcut verilerle başarılı bir makine öğrenmesi algoritması geliştirmek istemektedir. Böylelikle yeni gelecek tablet ürünlerini hangi vitrinde sergilemeleri gerektiğini hızlı ve güvenilir şekilde elde edebileceklerdir.

IT ekip lideri bu projeyi keşifçi veri analizi, ön işleme ve tahmine dayalı analitik makine öğrenmesi olmak üzere 3 safha olarak planlamaktadır. Ekip liderinin bu projeyi detaylandırdığı açıklama maili aşağıda bulunmaktadır.

--------

“İlk safhada yapılacak çalışmalar veriyi tanımaya ve sindirmeye yönelik (EDA) görselleştirme süreçlerinin tamamını kapsamaktadır. Veri setinin hedef değişkeninin dengeli dağılıp dağılmadığı incelenmeli ve yorumlanmalıdır. Öte yandan korelasyon matrisleri çizdirilmeli ve anlamlı görünen değişkenler arasındaki korelasyon katsayıları yorumlanmalıdır. Veri setindeki sayısal değişkenlerin ortalama, medyan, standart sapma gibi değerleri görüntülenmeli ve yorumlanmalıdır. Sayısal ve kategorik değişkenler ideal tekniklerle ve araçlarla görselleştirilmeli ve önemli olduğu düşünülen kısımlar yorumlanmalıdır.

------

İkinci safhada veri setindeki ön işleme kısmı gerçekleştirilmelidir. Ön işleme kapsamında eğer veri setinde eksik veriler varsa bu eksik verilerin mantıklı gerekçelerle doldurulması beklenmektedir. Bu işlem esnasında hedef değişkenin anlamlılığını kaybetmemesi için eksik verilerin hedef değişken özelinde incelenip doldurulması beklenmektedir. Ayrıca bu safhada kategorik değişkenlerin sıralanabilir(ordered) veya sıralanamaz(nominal) olma durumlarına göre uygun sayısallaştırma işlemleri yapılmalıdır.

------

Ucuncu safhada makine öğrenmesi modellemeleri yapabilmek için hedef değişkenimiz olan “FiyatAraligi” değişkeni ve bu değişkeni tahmin etmemize yardımcı olacak diğer değişkenler bağımlı ve bağımsız değişkenler olmak üzere ayrılmalıdır. Sonrasında bağımlı ve bağımsız değişkenler %85 oranında eğitim ve %15 oranında test olmak üzere “train\_test\_split” metodu kullanılarak ayrılmalıdır.

------

Eğitim kümesi ile GaussianNB, DecisionTree ve KNN modelleri eğitilmeli ve her bir model için karmaşıklık matrisi çizdirilmeli, doğruluk skoru elde edilerek modeller arasındaki başarı durumu karşılaştırma yapılmalıdır. Modellerin başarısını değerlendirme konusunda F1\_Score, Precision, Recall değerleri “classification\_report” aracılığıyla yazdırılmalıdır.

---------

Ekip lideri DecisionTree algoritmasının varsayılan olarak (criterion = “gini”) parametre değerini kullandığını ancak bu problem için criterion parametresini (criterion = “entropy”) olacak şekilde değiştirdiklerinde başarının artacağı konusunda ekibi temin etmiştir. Buna ek olarak KNN algoritması konusunda komşu sayısının çok kritik olduğunu ve 2’den 50’ye kadar olan her sayıyı komşu sayısı olarak deneyerek her denemede modelin skorunu kaydetmeyi önermektedir. Komşu sayısının değişimi ile modelin skorunun da değişeceğini ve her komşu sayısına tekabül eden model skorunu bir plot çizdirerek görselleştirme yapılması gerektiğini rica etmiştir.

-------

**Teslim Edilecek ve Sisteme Yüklenecek Çıktılar**

1. Proje ve video çekimi 2 kişilik gruplar halinde yapılacaktır.
2. Video linkini içeren bir txt ya da word dosyası. Bu linke tıklandığında videonun youtube, vimeo vb. linki açılmalı. Video indirme işlemi olmamalı.
3. Projenin Kaggle URL adresini içeren txt ya da word dosyası. Bu linke tıklandığında projenizin Kaggle adresi açılmalı.