**R ile Makine Öğrenmesine Giriş**

**Final Ödevi**

Değerli arkadaşlar, ödevinizde iki ayrı kısım bulunmaktadır.

**1. Kısım Denetimli Öğrenme**

Ödeviniz için “**credit.csv**” veri setini kullanacaksınız. Ödev kapsamında **k en yakın komşuluk (knn)**, **karar ağacı (decision tree) ve rassal orman (random forests)** yöntemlerinin bu veri seti üzerinde uygulanması ve **performanslarının karşılaştırılması** beklenmektedir.

Ödeviniz için R’de izleyeceğiniz adımlar şöyle (**Her bir adım için gerekli kodu yazmalısınız**):

* Veriyi içe aktarın ve inceleyin. Veri setindeki değişken isimlerini güncelleyelim. Sırasıyla; id, gelirduzeyi, yas, borc, borcgelirorani ve düzenliodeme olarak belirleyelim.
* Oluşan yeni veri setini inceleyin. Sırasıyla **id, gelirduzeyi, yas, borc, borcgelirorani** ve **duzenliodeme (0:ödeyen, 1:ödemeyen)** olmak üzere 6 değişkeniniz olmalı. Ancak **id** değişkenini kullanmayacağız. Veri setinden çıkarın.
* Nihai amacımız, **kalan** **diğer dört değişkeni kullanarak düzenli ödeyip ödememe** **durumlarını tahmin etmek**.
* Şimdi, yeni nesnemizin yapısını (str) ve betimsel istatistiklerini yazdırıp görelim.
* Eğitim ve test verilerini oluşturacağız. Bunun için farklı yollar kullanmıştık, siz istediğinizi seçin. İlgili komutu kullanarak **%80-%20** olacak biçimde eğitim ve test verilerini oluşturalım. **set.seed(41)** olacak.
* Şimdi, modelimizin tahmin gücünü ve dolayısıyla hatasızlığı (**accuracy**) artırmak için **“preprocessing”** yapacağız. Preprocessing neydi ve ne amaçla uygulanıyordu? Lütfen birkaç cümleyle belirtin.
* Bu amaçla **pp <- c("center", "scale")** kodunu çalıştırın. Buradaki “pp” yi, **veriyi eğitirken kurduğumuz modelde kullanacağız.**
* **k-kat** **çapraz geçerlik (k-fold cross validation**) yapacağız şimdi de..**. k=10** olacak. Daha önce bunu caret kullanarak yapmıştık. Yine train () içinde uygun şekilde bunu da kullanalım.
* Modelimizi, **Preprocessing ve k-fold** kullanarak, knn ile eğitelim. Ayrıca yine model içinde **“tuneLength = 10”** diyelim.
* Şimdi, modelimizi yazdıralım. Modele ilişkin hatasızlık (accuracy) değeri ne çıktı? Model için kullanılan nihai **“k”** değeri kaç olarak belirlendi? (buradaki k, k-en yakın komşuların “k”sı, karıştırmayalım).
* **Şimdi aynı modeli bir de “preprocessing” ve “k-kat çapraz geçerlik” olmadan (ikisini birden değil, sırayla çıkararak) deneyelim. Bu iki durum için “accuracy” ve “k” nasıl değişti?**
* Sonra, modelimizi ilk-orjinal haline döndürerek test verisi üzerinde tahminde (prediction) bulunalım (yani hem preprocessing hem de k-fold yapılmış haline).
* “**ConfusionMatrix**” kullanarak çapraz tabloyu oluşturalım ve raporlayalım. “Accuracy” ile birlikte “Sensitivity” ve “Specifity” değerleri nasıl çıktı? Yorumlayalım.
* Şimdi, aynı veri setiyle **karar ağacı** oluşturalım. Eğitim için yine **train ()**’i kullanacağız. **Tek fark, method için “knn” yerine “rpart” kullanmak**...
* Modeli eğittikten sonra test verisiyle tahmin (prediction) yapalım. Modeli “print” edelim. **“cp” değerini ne olarak belirledi** (yani en iyi “cp” değeri)**?** Raporlayalım.
* Sonra yine **ConfusionMatrix** oluşturalım ve “**accuracy, sensitivity ve specifity**” değerlerini yorumlayalım.
* Son aşamada yine ayrı veri setiyle **rassal ormal algoritmasını** çalıştıralım.
* Optimum mtry ve ntrees sayılarını belirlemeye çalışalım.
* Değişken önem sırasını yazdırıp ve değişkenleri önem sırasına göre sıralayalım.
* Yine **ConfusionMatrix** oluşturarak “**accuracy, sensitivity ve specifity**” değerlerini yorumlayalım.
* Son olarak, **bu üç yönteme göre elde edilen modellerden hangisini tercih edersiniz? Neden? Kısaca açıklayınız.**

**2. Kısım. Denetimsiz Öğrenme (K-ortalamalar Yöntemi)**

Bu sefer “**housing.csv**” dosyasında çalışacağız. Ödev kapsamında **k-ortalamalar** yönteminin bu veri seti üzerinde uygulanması beklenmektedir.

Ödeviniz için R’de izleyeceğiniz adımlar şöyle (**Her bir adım için gerekli kodu yazmalısınız**):

* Veriyi içe aktarın ve inceleyin. Veri setinden yalnızca ‘longitude’, ‘latitude’ ve ‘median\_income’ değişkenlerini seçin ve değişken isimlerini ‘enlem’, ‘boylam’ ve ‘gelir’ olacak şekilde güncelleyelim.
* Oluşan yeni veri setini inceleyin. Sırasıyla **enlem, boylam ve gelir** olmak üzere 3 değişkeniniz olmalı.
* Şimdi, yeni nesnemizin yapısını (str) ve betimsel istatistiklerini yazdırıp görelim.
* K sayısını nasıl belirlediniz? K yı belirleme yöntemleri nelerdir? Farklı denemeler yapınız ve en uygun sonucu raporlayınız.
* Sonucunuzu görseliyle (plot) birlikte raporlayıp yorumlayınız.

**Ödevler, 23 Ocak 2025 saat 22:00’da dersimizin “final ödevi” bölümüne yüklenmelidir.**

Ödevi **bireysel ya da grup** olarak (en fazla 3 kişi) yapabilirsiniz. Grup ödevlerinde tüm **grup üyeleri, ödevlerini** **ayrı ayrı yüklemelidir**. Ödevin kapağında grup üyelerinin isimleri yer almalıdır.

Ödeviniz bir **“R script” (komut dosyası) ve Word dosyası**ndan oluşmalıdır. Komut dosyasında bütün işlemler yukarıda istenilen sırayla yapılmalıdır. Her birinin başına **bir “#” koyarak kısaca açıklamasını yapabilirsiniz**. Ben kodları çalıştırdığımda tüm sonuçları, sizin raporladığınız biçimde alabilmeliyim. **Bu sonuçlar, raporladıklarınızla aynı olmalı.**

Kodları aşama aşama word’e yazmanızı beklemiyorum. Gerektiği yerde sonuçları sırası ile paylaşır ve hemen altına kısaca, istenmişse yorumunuzu yazarsınız.

**Başarılar,**

**Başak ERDEM KARA**