Hazal Yıldız

Data Science Bootcamp Proje Raporu

Bu veriyi data science ile sağlık alanlarını birleştirmeyi sevdiğim için seçtim. Veri 193 ülkenin çeşitli sosyo-ekonomik değişkenlerini ile beklenen yaşam süresini(response) içeren bir veri. Veri DSÖ'nün veri tabanından alınmış ve 2000-2015 yılları arasında toplanmış. Her ne kadar verinin zamana bağlı bir yapısı olsa da bu analizde amacım time-series analizi yapmak değil, ülkelerin farklı sosyo-ekonomik değişkenlerinin beklenen yaşam süresini nasıl etkilediğini görmek. Ayrıca veride her ülke için her yıl kayıt tutulmamış. Veri de response dahil 22 değişken ve 2938 kayıt var. Veride toplam 512 tane gelişmiş, 2426 tane de gelişmekte olan ülkeye ait kayıt var.

Bazı değişkenler şu şekilde: yetişkin ve çocuk ölüm oranları, kızamık, difteri, çocuk felci gibi hastalıkların görülme sıklığı, BMI, GDP, toplam nüfus, okuma oranı.

Hem verideki NA değerleri sorgularken hem de verinin tanımında yazdığı üzere, NA verilerin geldiği ülkeler gelişmekte olan ve az bilinen ülkeler. Veride gelişmiş ülkeler de olduğu herhangi bir imputation metodu kullanmak bize çok doğru olmayan analiz sonuçları verebilir. Bu yüzden verideki NA değerlerini veriden çıkarmayı tercih ettim. Bu işlemden sonra toplam 1649 kayıt kaldı.

EDA kısmına geçtiğimde, tüm ülkeler baz alındığında yıllar içinde ortalama beklenen yaşam süresinde göze çarpan değişimler olmadığını gördüm. Genel trende baktığımızda ortalama beklenen yaşam süresinin 2000-2015 yılları arasında 70 yıl civarında olduğunu söyleyebiliriz.

En yüksek ortalama yaşam süresine sahip ülkeler sırasıyla İrlanda, Kanada, Fransa, İtalya ve İspanya. Bu ülkelerin ortalamaları 82 ile 83.4 arasında değişiyor ve aralarında çok az fark var. En düşük ortalama sahip ülkeler ise sırasıyla Sierra Leone, Lesotho, Zimbabwe, Malawi ve Angola ve bu ülkelerin ortalamaları da 48.2 ile 50.7 arasında değişiyor.

En yüksek GDP'ye sahip ülke Luxemburg. Onu Hollanda, Avusturalya, Avusturya ve İsveç takip ediyor. Bu ülkeler ile en yüksek ortalama yaşam süresine sahip ülkeler arasında kesişen bir ülke yok. Bu bize GDP ile beklenen yaşam süresi arasında bir ilişki olmadığı konusunda bir ihtimal olduğunu düşündürüyor. GDP ile beklenen yaşam süresi arasındaki korelasyonu hesapladığımızda 0.44 çıkıyor. Bu da arada doğrusal bir ilişki olmadığını gösteriyor bize.

Tüm sayısal değişkenler arasındaki korelasyonu daha iyi görebilmek için heatmap yaptım ve beklenen yaşam süresi ile arasında kayda değer korelasyon olan üç değişken olduğunu gördüm: adult mortality, ICR, ve schooling. Bu üç değişken ile beklenen yaşam süresi arasında doğrusal pozitif ilişki olduğunu görüyoruz(correlation coeff > 0).

Gelişmiş ve gelişmekte olan ülkelerin ortalama yaşam sürelerine baktığımızda arada bir fark görüyoruz. Gelişmiş ülkeler için ortalama 78.7 iken gelişmekte olan ülkeler için ortalama 67.7.

Modellemeye geçtiğimizde ben KNN modeli kurmayı tercih ettim. Bu model bir makine öğrenmesi modeli olduğu için one-hot encoding yaptım. Ayrıca modelleme aşamasında yıl ve ülke değişkenleri bize anlamlı sonuçlar vermeyeceği için bu değişkenleri modelleme kısmına dahil etmedim. Modelin doğruluğunu ölçmek için %70-%30 oranında train-test split yaptım. KNN modelinin neighbor sayısını 10 aldım. Modelin doğruluğunu ölçmek için MAE ve RMSE gibi değerleri hesapladım. MAE ve RMSE değerlerinin düşük olması, modelin accuracy'sinin yüksek olduğu anlamına geldiği için ve modelimizde de bu değerler düşük olduğu için bu modelin kullanılabilir bir model olduğunu söyleyebiliriz.