Bitirme Projemin adı “Yıllara ve Ülkeler Göre Kanser Tiplerine Bağlı Ölüm Oranları Veri Setinin İncelenmesi, Analizi”. Eğitimimi tamamladığım Bootcamp dersleri boyunca aldığım bilgiler, yetkinlikler ve farklı analiz tekniklerini kendi bölümüm ve alanım olan Moleküler Biyoloji ve Genetik çatısı altında kullanmak istedim.

Lisans eğitimim boyunca Kanser çalışmalarında yer aldım ve bu sebeple veri seti araştırması yaparken aklımda kanser ile ilişkili bir set bulmak vardı. Sonuç olarak üzerinde çalışmaya nihai karar kıldığım veri setinde, ülkeler, 1990-2016 yılları, kanser çeşitleri ve ölüm oranları bulunmaktadır.

Bu veri setiyle çalışma yapmanın sonucunda, alandaki teorik bilgilerimi sayısal şekilde de destekleyecek yetkinliğe sahip olacağımı düşündüm. Örneğin, en yaygın kanser tipinin Akciğer Kanseri olduğunu biliyordum ve yaptığım çalışma sonucunda da bunu destekleyen nitelikte verilere ulaşabildim.

İlk adım olarak, ihtiyacım olacağını düşündüğüm bütün kütüphaneleri import ettim. Daha sonra veri setine aşina olabilmek adına baştaki 100 değer, sondaki birkaç değer, veri setinin uzunluğu, genel bilgisi, sütunların ismi gibi kodlar yazarak incelemede bulundum. Veri setinin genel bilgisine baktıktan sonra Data Type’larını sonraki adımlarda yapacağımı düşündüğüm işlemlere uygun hale getirdim.

Veri setini daha yakından incelemek adına 10 temel soru sorarak cevaplarını aradım. Bunlar arasında ilgimi çeken noktaların üzerine gittim. Örneğin 1990-1994 yılları arasında, Afganistan’da en yüksek ölüm oranı artışına sahip olan kanser tipine ulaşabilmek benim için heyecan verici bir deneyimdi. İleri adımlarda kullanacağımı düşündüğüm için değişkenlerin ortalama, standart sapma, minimum ve maksimum değerlerini de bu noktada hesapladım. Bunların yanında, eksik bilgilere sahip veriler tespiti de yaptım, temizledim. Biri çubuk grafiği biri korelasyon matrisi olmak üzere iki görselleştirme ile de bu temel basamaktaki işlerimi tamamladım. Veri setine aşinalığım, hakimiyetim arttı, bir tık üst düzey sorularla incelemeye devam etmeden merak ettiğim değerlerin grafiklerini de oluşturdum. Böylece görselleştirme açısından zenginlik elde ettim.

Diğer bir 10 soruluk sete başlarken daha detaylı incelemede bulunmaya çalıştım. Farklı grafik tipleri, tablolar çalıştım. Akciğer, karaciğer, meme kanserleri en çok çalıştığım ve üzerine literatür taraması yapmış olduğum kanser tipleri olduğundan bu kanser türleri ile en fazla ilgilendiğimi söyleyebilirim. Pandas kütüphanesinin “Profiling”ini kullanarak 5 farklı bölgedeki en yüksek ölüm oranına sahip kanser türlerini kapsamlı inceledim, belki de benim için en verimli ve ilginç kısmı burası oldu.

Machine Learning kısmına geçmeden önce, özellikle incelemek istediğim birkaç noktanın üzerine gittim. Ülkemize ait oranları da özellikle görmek istediğim için projeme böyle bir kısım da ayırdım. Yapmış olduğum bir çıkarım, “Mide kanserindeki yıllar arasında dramatik düşüşü görülüyor, bu güzel bir gelişme. akciğer kanseri dünya üzerinde bilinen en yaygın kanser tipi olduğu gibi Türkiye'de de durumun aynı seyrettiği çıkarımı yapılıyor.” İleride bu alanda çalışma yaparsam aklımın köşesinde duracak bir çıktı oldu benim için.

Python çalışmalarında yeni olduğum, Machine Learning alanında da kendimi biraz daha ilerletmem gerektiğini düşündüğüm için projemin son bölümü olan bu alanda küçük bir çalışma gerçekleştirdim. Karaciğer kanserinin 2010-2016 yılları arasında Türkiye’de sahip olduğu oranlar kullanılarak bir model oluşturuldu ve R2 skoru hesaplandı.

Fatma Müge TEKİN