Ceyda DİZDAR

10310110912

1. Öğretim

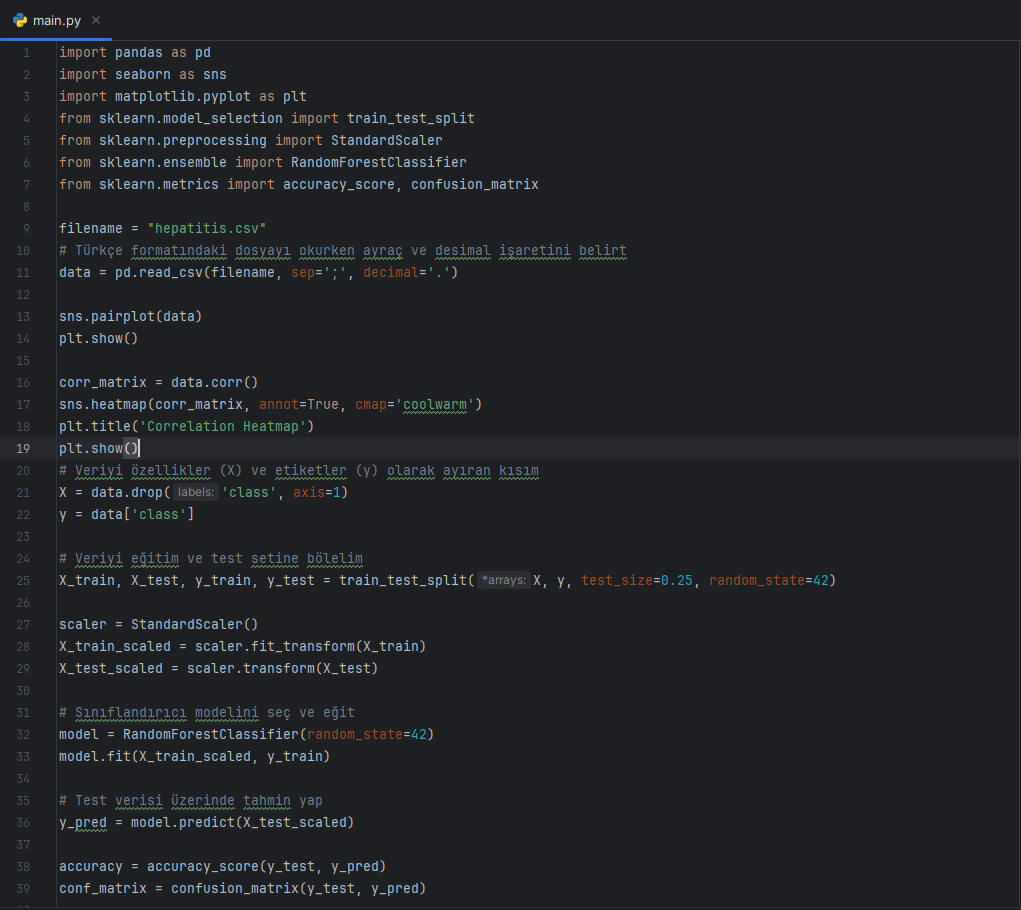
Python Programlama



Bize verilen “hepatitis.csv” adlı veri seti, hastaların klinik ve laboratuvar ölçümlerini içerir ve farklı özelliklere sahip hastalardan elde edilen verileri barındırır. Her bir örnek, hastanın sınıfını belirten "Class" sütunu ile başlar; bu sınıflar genellikle hastanın hayatta kalma veya ölüm durumunu temsil eder. Diğer önemli özellikler arasında "VARICES" (varislerin varlığı), "BILIRUBIN" (bilirubin düzeyi), "SGOT" (Serum Glutamic Oxaloacetic Transaminase) düzeyi, "ALBUMIN" (albümin düzeyi) ve "PROTIME" (prothrombin zamanı) bulunmaktadır.

Örneğin, "VARICES" sütunu hastaların varis varlığını belirtirken, "BILIRUBIN" sütunu bilirubin düzeyini temsil eder. "SGOT" sütunu, hastaların serum glutamik oksaloasetik transaminaz düzeyini, "ALBUMIN" sütunu albümin düzeyini ve "PROTIME" sütunu ise prothrombin zamanını gösterir. Bazı hastaların verileri eksik olabilir; bu durumda, ilgili hücreler boştur.

Bu veri setinin analizi için yapılan kod, veri setini daha iyi anlamak, görselleştirmek ve sınıflandırma modeli oluşturmak amacıyla kullanılmıştır. Önce, veri seti üzerinde çeşitli grafikler oluşturularak özellikler arasındaki ilişkiler ve korelasyonlar incelenmiştir. Daha sonra, RandomForestClassifier kullanılarak bir sınıflandırma modeli oluşturulmuş, bu model eğitilmiş ve test edilmiştir. Sonuçlar, modelin doğruluğunu ve sınıflandırma performansını değerlendirmek için kullanılmıştır. Kod aşağıdaki gibidir:



**Hepatitis Veri Seti Analiz Raporu**

Bu rapor, hepatit hastalığı ile ilgili klinik ve demografik bilgiler içeren bir veri setini ele almaktadır. Veri seti, Gail Gong aracılığıyla Carnegie-Mellon Üniversitesi'nden Bojan Cestnik'e ulaşılarak derlenmiş ve 1988 yılında oluşturulmuştur.

**1. Veri Seti Özellikleri:**

* Toplam örnek sayısı: 155
* Toplam özellik sayısı: 20 (Sınıf etiketi dahil)
* Eksik veri içeren özellikler: 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
* Sınıf etiketleri: "LIVE" (Hayatta) ve "DIE" (Ölü)

**2. Veri Seti İçeriği:**

* Yaş, cinsiyet, tedavi bilgileri, karaciğer durumu ve diğer klinik özellikler.
* Continous (sürekli) veri içeren özellikler: BILIRUBIN (Bilirubin Seviyesi).
* Histoloji sonucu: "yes" veya "no" olarak belirtilmiştir.

**3. Önceki Kullanımlar ve Başarılar:**

* Gail Gong'un çalışmasında %80 sınıflandırma doğruluğu elde edilmiştir.
* Cestnik'in Assistant-86'sı %83 sınıflandırma doğruluğu sağlamıştır.

**4. Veri Seti Analizi:**

* Veri seti üzerinde çeşitli görselleştirmeler yapılmıştır. Çiftler arası (pairplot) analiz ile özellikler arasındaki ilişkiler incelenmiş ve korelasyon ısı haritası (heatmap) ile de bu ilişkiler daha detaylı bir şekilde gösterilmiştir.

**5. Model Eğitimi ve Değerlendirme:**

* Veri seti, Random Forest sınıflandırıcı modeli kullanılarak eğitilmiş ve test edilmiştir.
* Eğitim ve test veri setleri ölçeklendirilmiş ve model eğitilerek test verisi üzerinde tahminler yapılmıştır.
* Modelin başarı oranı ve karışıklık matrisi hesaplanarak değerlendirilmiştir.

Son olarak kodu çalıştırdığımızda elde ettiğimiz grafikler:

