İZMİR KÂTİP ÇELEBİ ÜNİVERSİTESİ
2010
FEN BİLİMLERİ ENSTİTÜSÜ

Hasta Teşhis ve Tahmin Uygulaması: Yapay Zeka Destekli Tıbbi Teşhis

Yazılım Mühendisliği Ana Bilim Dalı

Dönem Projesi

Ender KÖKSALDI

ORCID 0009-0008-2522-9318

Proje Danışmanı: Doç. Dr. Vahide BULUT

Ocak 2024

Hasta Teşhis ve Tahmin Uygulaması

Öz

Bu projede, çeşitli semptomların hastalıklar ile ilişkileri ele alınmıştır. Semptomların hastalıklar üzerindeki etkilerini görselleştirmek için çeşitli grafikler oluşturulmuştur. Bu grafikler üzerinden, hangi semptomların hangi hastalıkları tanımladığı incelenmiştir.

Daha sonra, semptom ağırlıklarıyla zenginleştirilmiş bir diğer veri seti kullanılarak çeşitli sınıflandırma algoritmaları değerlendirildi. Decision Tree, Random Forest gibi algoritmaların performansları karşılaştırıldı ve bu modellerin hastalık tahmininde ne kadar etkili olduğu belirlendi. Algoritmaların eğitim ve test doğruluk oranlarını bar grafikleri ile görselleştirildi, ayrıca bu oranların dağılımları standart sapmalarıyla birlikte incelendi.

Sonuç olarak verilen semptomlara göre hastalık tahminlemesi yapan bir yapay zekâ modeli geliştirildi ve sonuçlar raporlandı.

Anahtar Sözcükler: Veri görselleştirme, sınıflandırma algoritmaları, yapay zekâ, veri manipülasyonu

Patient Diagnosis and Forecasting App

Abstract

In this project, the relationship between various symptoms and diseases is discussed. Various graphs were created to visualize the effects of symptoms on diseases. Through these graphs, it was analyzed which symptoms define which diseases.

Then, various classification algorithms were evaluated using another dataset enriched with symptom weights. The performances of algorithms such as Decision Tree and Random Forest were compared and the effectiveness of these models in disease prediction was determined. The training and test accuracy rates of the algorithms were visualized with bar graphs, and the distributions of these rates were examined along with their standard deviations.

As a result, an artificial intelligence model that predicts disease according to the given symptoms was developed and the results were reported.

Keywords: Data visualization, classification algorithms, artificial intelligence, data manipulation

İçindekiler

Şekiller Listesi

Bölüm 1

Giriş

Sağlık sektöründeki hızlı teknolojik gelişmeler ve büyük veri analitiği, hastalıkların tanı ve tedavisinde yeni ve etkili yöntemlerin keşfini sağlamıştır. Bu bağlamda, hastalıkların belirtilerinin doğru bir şekilde analiz edilmesi, doğru tanının konulması ve uygun tedavi yöntemlerinin belirlenmesi büyük önem taşımaktadır. Bu doğrultuda, bu çalışma, semptom verilerinin analizi üzerine odaklanarak, hastalıkların belirtilerinin ağırlıklarını kullanarak hastalık tahminini amaçlamaktadır.

Yapay zekâ, insan beyninin işlevlerinden ilham alarak, insan benzeri yetilerin bilgisayarlar, robotlar, programlar gibi sistemlere aktarılmasıdır. Sağlık yönetimi alanında yapay zekâ uygulamaları oldukça geniş bir kapsama sahiptir. Dünya genelinde birçok ülkede, ulusal ve bölgesel sağlık kuruluşlarının yönetiminde yapay zekâ kullanılmaktadır.

Akalın ve Veranyurt (2021)'in de belirttiği gibi Türkiye'de, Sağlık Bakanlığı'nın yapay zekâ uygulamaları, Microsoft, Oracle gibi teknoloji firmalarının ürünlerini içermektedir. Bu uygulamalar arasında MHRS (Merkezi Hekim Randevu Sistemi) kullanım oranlarının izlenmesi, aile hekimliği performans raporlarının oluşturulması, hastaneye yatış, ameliyat, tanı gibi raporların değerlendirilmesi ve eNabız verilerinin analizi gibi işlemler bulunmaktadır.

> "*Makine öğrenmesinde kullanılan algoritmalar eldeki verileri analiz ederek kendisini eğitir. Bilgisayarın kendisine kazandırdığı bu yetenek sayesinde makineye sorulan işlemlerin sonucu için bir tahminde bulunur. Literatürde 26 adet makine öğrenme algoritması mevcuttur. Bu algoritmalar, eldeki verilere göre başarı sonucu değişmektedir. Bu*

sebeple belirli bir yöntem ile tüm veriler ile işlem yapmak doğru değildir (Kılıç, 2012)."

Bu projede, sağlık verilerinden elde edilen semptom ağırlıkları ile makine öğrenimi algoritmalarının kullanımı incelenmiştir. Semptom ağırlıkları, hastalıkların belirtilerinin önem derecelerini gösteren bir ölçüdür. Makine öğrenimi algoritmaları ise bu ağırlıkları kullanarak hastalıkları sınıflandırmak ve tahmin etmek için uygulanmıştır.

Çalışmanın amacı, hastalıkların belirtilerini semptom ağırlıklarıyla ilişkilendirerek, bu veri seti üzerinde sınıflandırma algoritmalarını değerlendirmek ve hastalık tahminindeki başarılarını değerlendirmektir. Bu bağlamda, projenin ilk bölümünde kullanılan veri seti ve semptom ağırlıkları açıklanacak, ardından sınıflandırma algoritmalarının kullanımı ve sonuçları detaylı bir şekilde ele alınacaktır.

Projenin sonucunda hedeflenen; sınıflandırma algoritmalarının doğruluk değerlerine göre karşılaştırılarak, semptomlara göre en doğru hastalık tahminlemesi yapan modelin geliştirilmesidir.

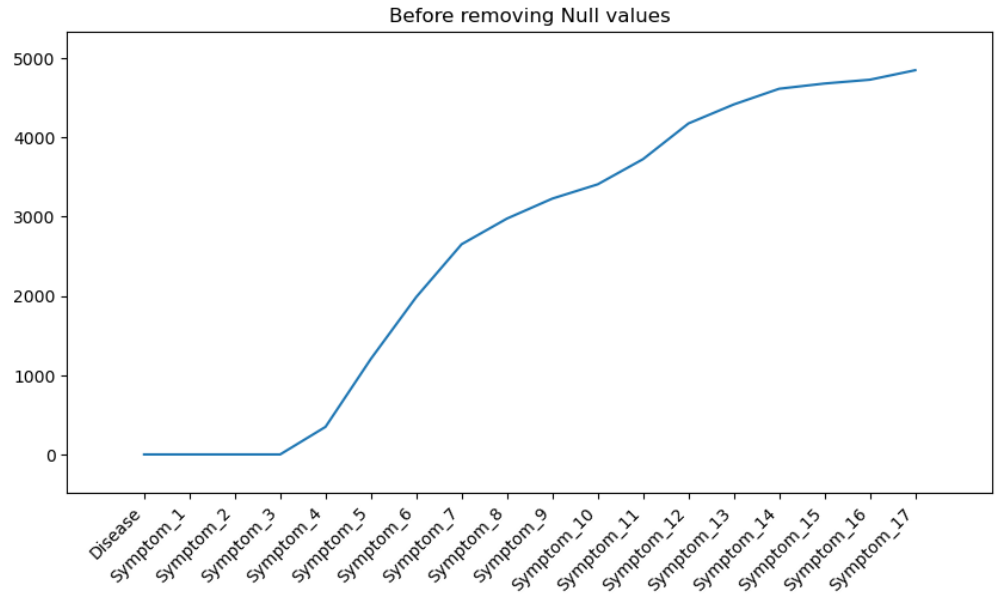
Bölüm 2

Veri Manipülasyonu

Bu bölümde, projenin temel veri manipülasyon süreçleri gerçekleştirildi. Veri setindeki eksik değerler ele alındı ve semptom ağırlıklarını içeren yardımcı bir veri setinden elde edilen değerler ana veri setine entegre edildi. Bu adımlar, projenin analiz aşamalarını daha güvenilir ve tutarlı hale getirmek amacıyla uygulandı.

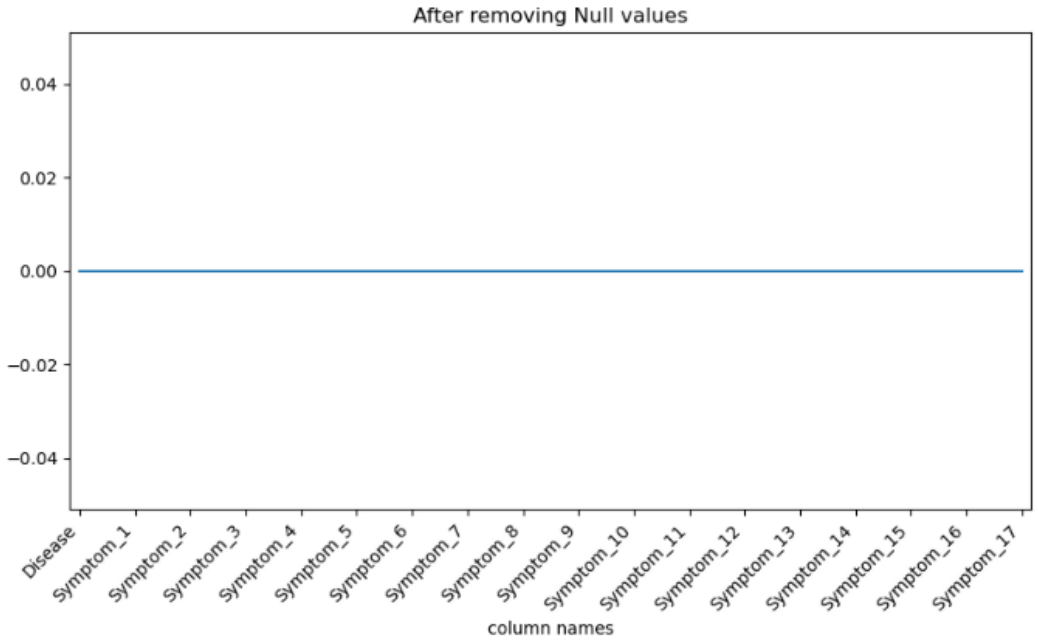
2.1 Eksik Değerlerin Doldurulması

Yapılan incelemede projemizde kullandığımız veri setinde, hastalık semptomları alanlarında ilk dört semptom sonrasında çokça eksik değer gözlenmiştir (bakınız Şekil 2.1).



Şekil 2.1: Eksik değer atamaları yapılmadan önceki veri seti

Bu doğrultuda, eksik değerler ilgili semptom ve hastalıkların özellikleri dikkate alınarak uygun değerlerle doldurulduldu (bakınız Şekil 2.2).



Şekil 2.2: Eksik değer atamaları yapıldıktan sonraki veri seti

2.2 Semptom Ağırlıklarının Güncellenmesi

2.2.1 Yardımcı Veri Setinin Aktarılması

Projede kullanılan ana veri setine, semptom ağırlıklarını içeren yardımcı bir veri seti entegre edildi. Bu entegrasyon sayesinde, semptomların hastalıklara olan etkilerini daha hassas bir şekilde değerlendirmek mümkün hale geldi. Bu aşama, projenin analitik derinliğini artırmaya yönelik bir adım olarak uygulandı (bakınız Şekil 2.3).

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Symptom_10
0	Fungal infection	1	3	4	dischromic patches	0	0	0	0	0	0
1	Fungal infection	3	4	dischromic patches	0	0	0	0	0	0	0
2	Fungal infection	1	4	dischromic patches	0	0	0	0	0	0	0
3	Fungal infection	1	3	dischromic patches	0	0	0	0	0	0	0
4	Fungal infection	1	3	4	0	0	0	0	0	0	0

Şekil 2.3: Aktarım sonrası veri seti

2.2.2 Ağırlık Değerleri Olmayan Verilere Atama

Aktarım sonrasında bazı semptomların ağırlık verilerinin yardımcı veri setinde yer almadığı görüldü. Bu veriler için sıfır değerleri atanarak semptomların herhangi bir ağırlığı olmadığı varsayıldı (bakınız Şekil 2.2).

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Symptom_10
0	Fungal infection	1	3	4	0	0	0	0	0	0	0
1	Fungal infection	3	4	0	0	0	0	0	0	0	0
2	Fungal infection	1	4	0	0	0	0	0	0	0	0
3	Fungal infection	1	3	0	0	0	0	0	0	0	0
4	Fungal infection	1	3	4	0	0	0	0	0	0	0

Şekil 2.4: Yardımcı veri setinde olmayan verilere atama yapıldı

Bölüm 3

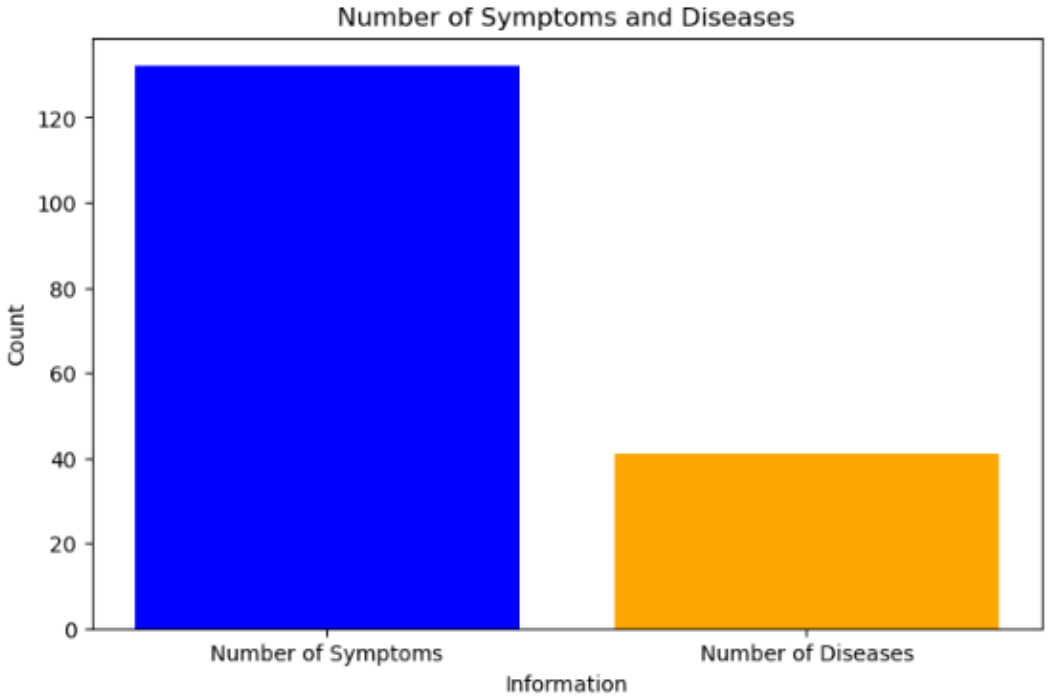
Veri Görselleştirme

Veri görselleştirme aşamasında, projenin ana veri seti üzerinde hastalıklar ve semptomlar arasındaki ilişkiler grafikler üzerinde incelendi. Bu aşama, projenin analiz ve yorumlama süreçlerini destekleyerek bilgi çıkarımına yardımcı oldu.

3.1 Etkili Semptomların Belirlenmesi

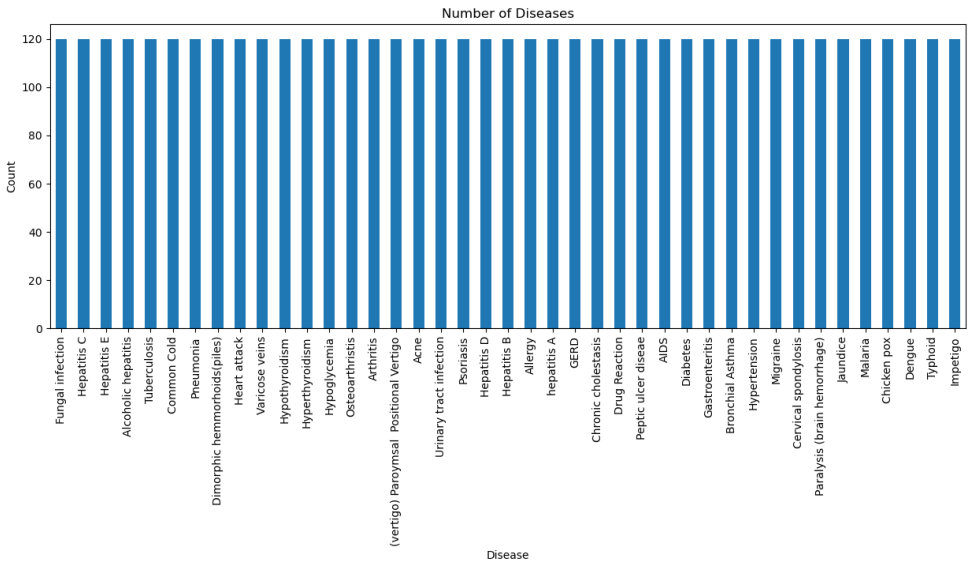
3.1.1 Analiz

Yapılan incelemede; veri setinde hastalıkları belirleyen semptomlarda genellikle ilk dört semptomun etkili olduğu görüldü (bakınız Şekil 2.1). Ayrıca veri setinde toplam 132 adet benzersiz semptom, 41 adet benzersiz hastalık bulunduğu görüldü (bakınız Şekil 3.1).



Şekil 3.1: Hastalık ve semptom sayıları

Bunun yanında verilerin dağılımını analiz etmek için veriler hastalık bazında gruplanarak incelenmiştir. Buna göre verilerin hastalık bazında eşit dağıldığı görülmektedir. Her bir hastalık için toplam 120 adet veri bulunmaktadır (bakınız Şekil 3.2).

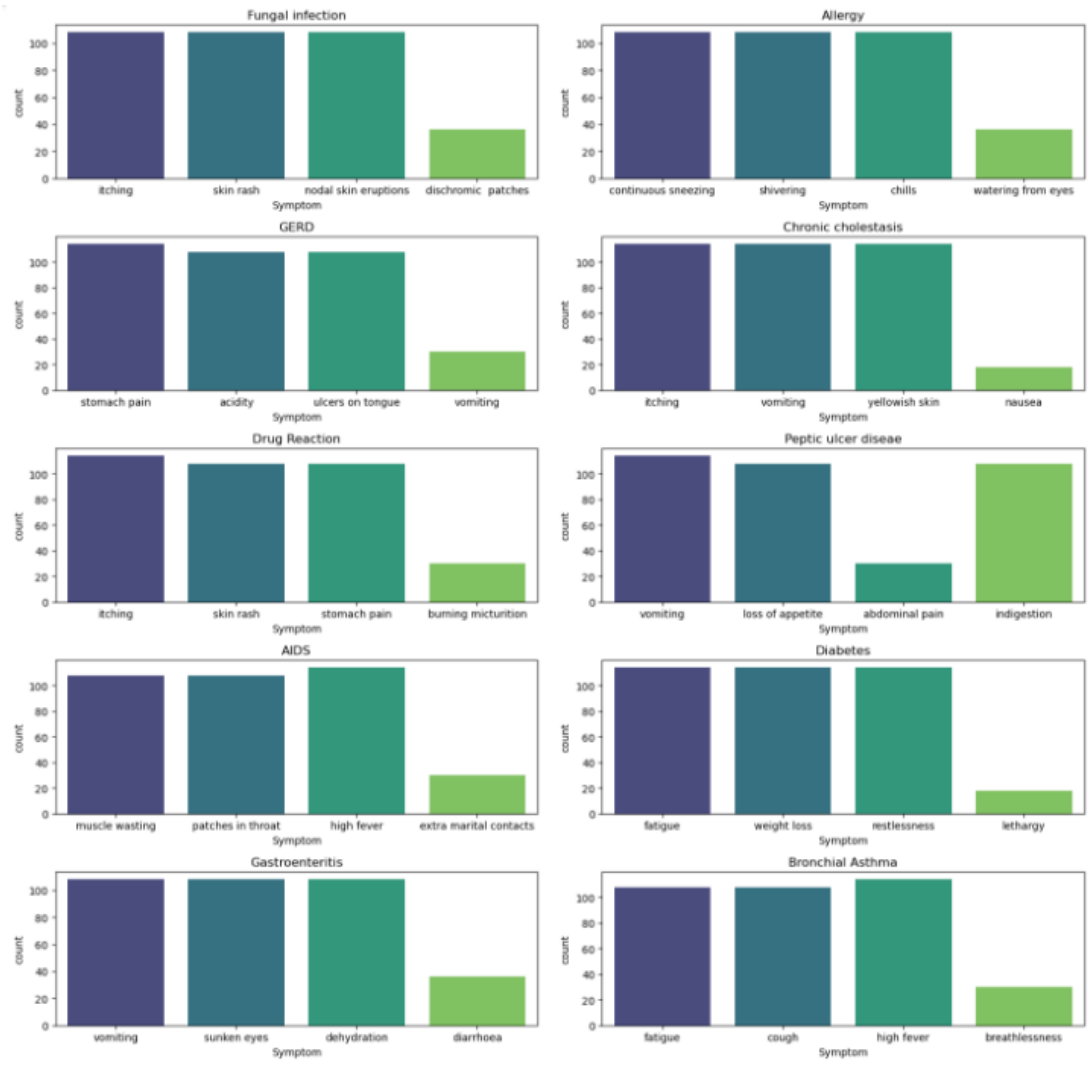


Şekil 3.2: Hastalık bazında veri sayıları

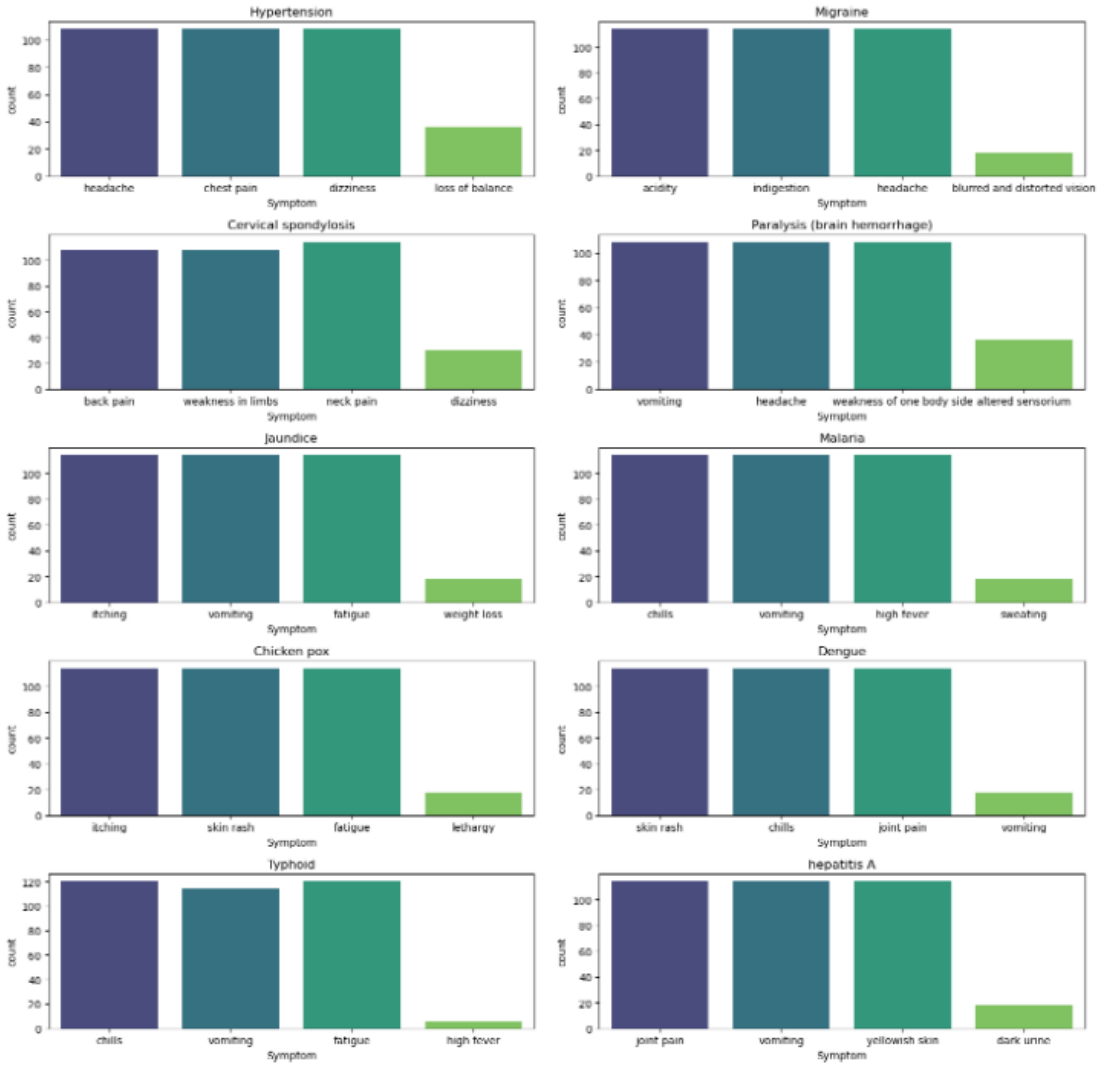
3.1.2 Görselleştirme

3.1.2.1 Hastalıklara Göre En Belirleyici Semptomlar

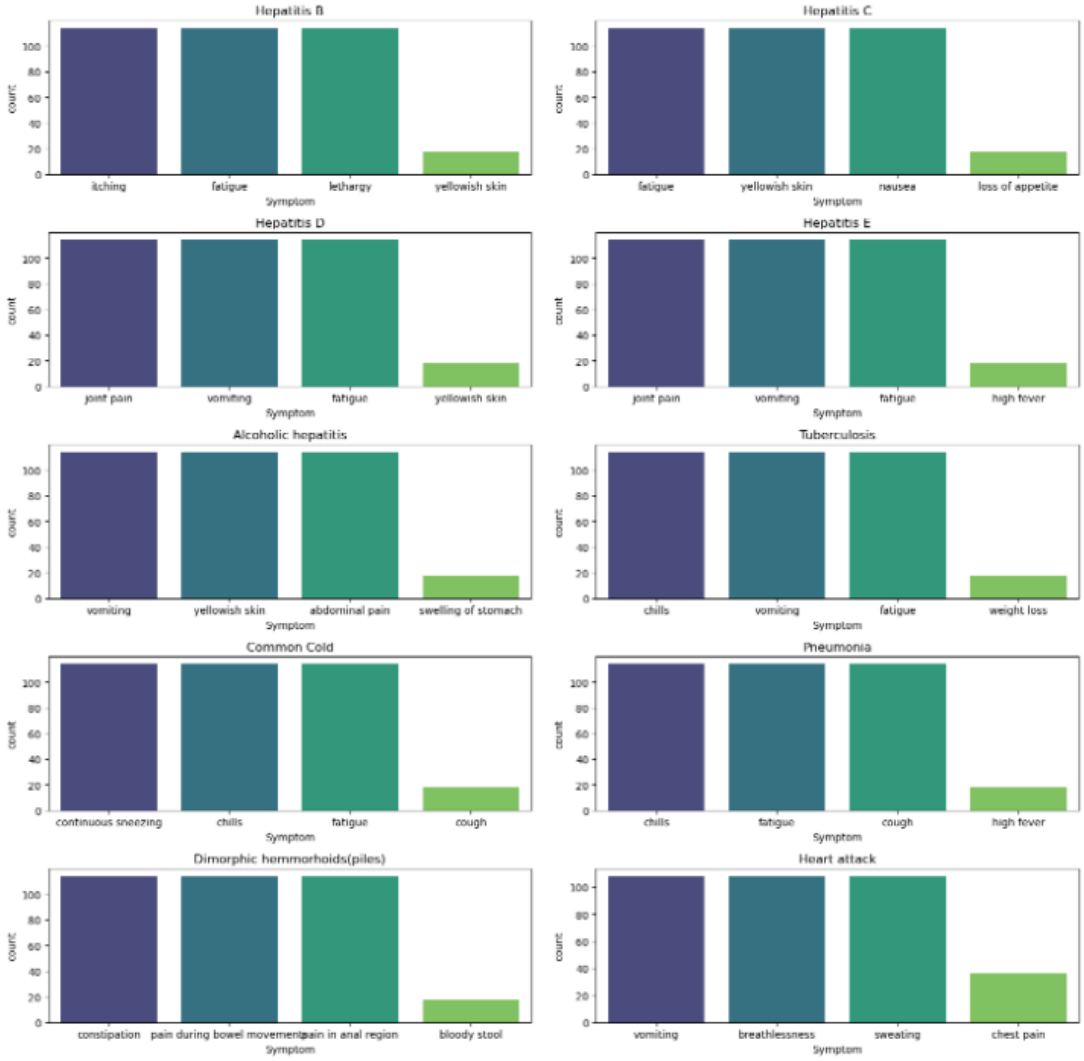
Yapılan analizler sonucunda verilerde genellikle dört semptomun belirleyici olduğu görülmüştü. Buna göre her bir hastalık için en belirleyici dört semptom görselleştirilmiştir (bakınız Şekil 3.3).



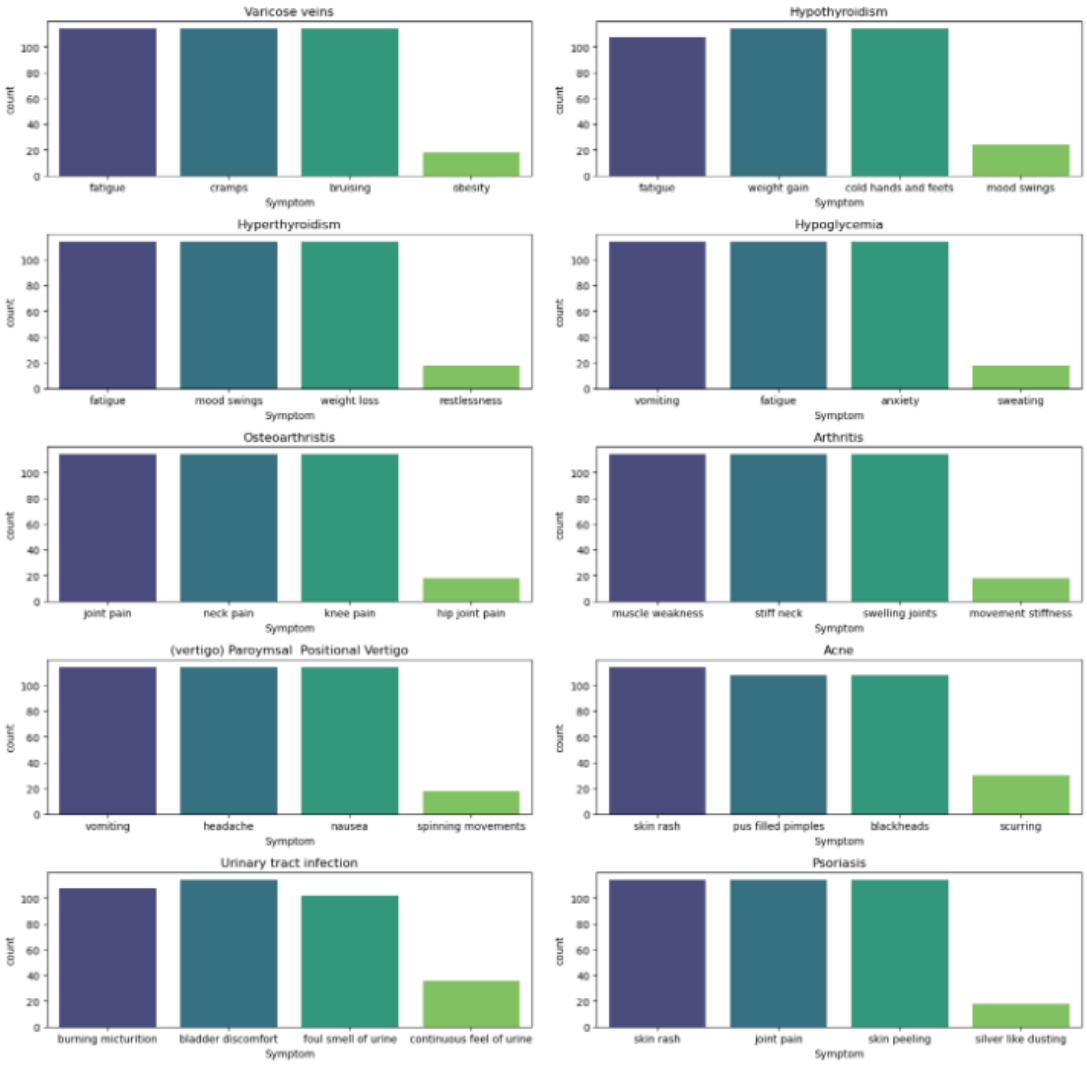
Şekil 3.3: Hastalıkların belirleyici semptomları



Şekil 3.4: Hastalıkların belirleyici semptomları



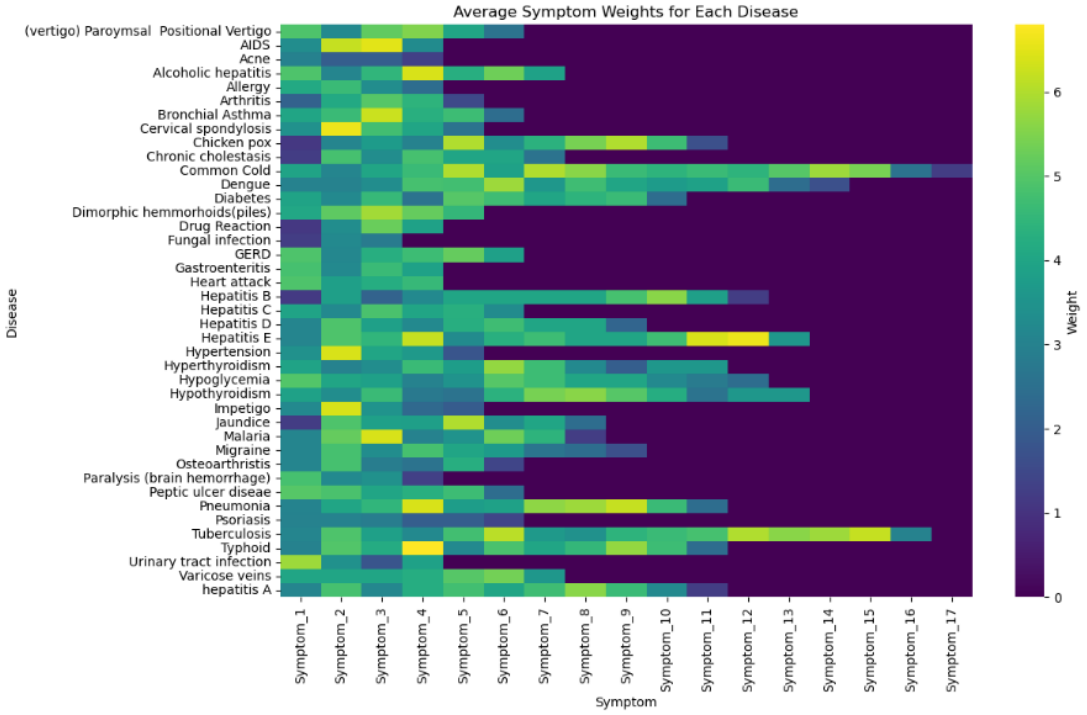
Şekil 3.5: Hastalıkların belirleyici semptomları



Şekil 3.6: Hastalıkların belirleyici semptomları

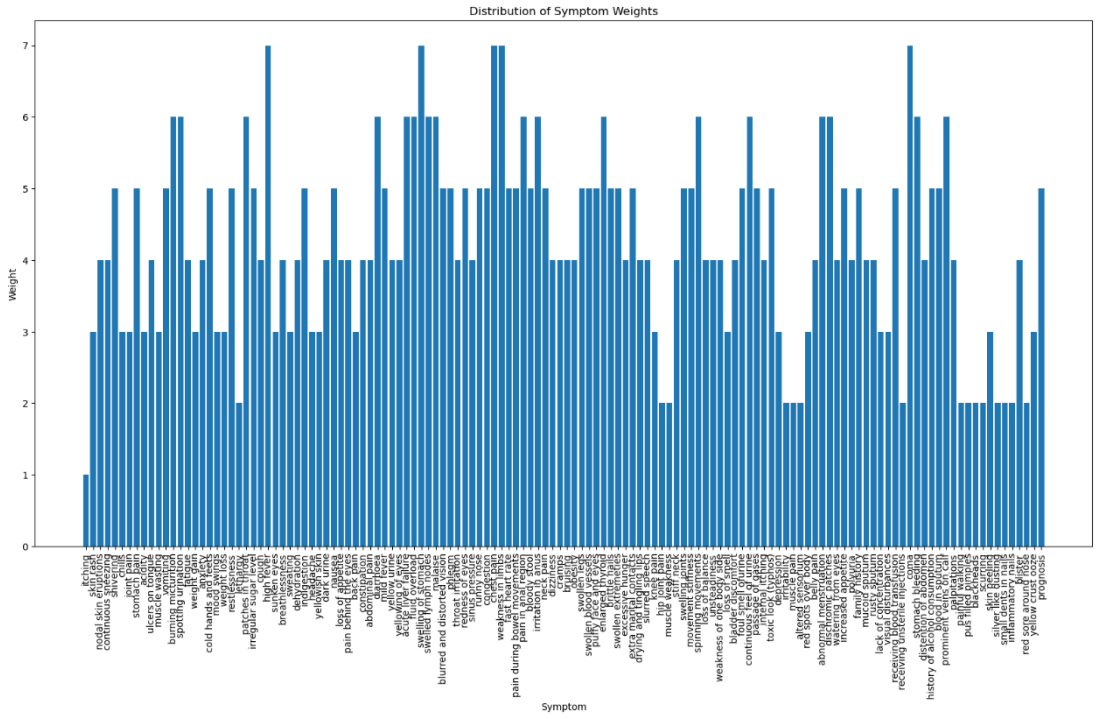
3.1.2.2 Semptom Ağırlık Grafikleri

Hastalıklara göre semptom ağırlıklarının ortalamaları, bir ısı haritası üzerinde görselleştirildi. Bu görselleştirme, hastalıkların semptom ağırlıkları konusunda anlamlı bir bilgi sunmaktadır (bakınız Şekil 3.7).



Şekil 3.7: Hastalıkların semptom ağırlık dağılımları

Ayrıca tüm semptomların ağırlık değerleri bir bar grafiği üzerinde incelendi (bakınız Şekil 3.8).



Şekil 3.8: Hastalıkların semptom ağırlık değerleri

Bölüm 4

Makine Öğrenmesi

Bu bölümde, hastalık teşhisi için kullanılan makine öğrenmesi sınıflandırma algoritmalarının eğitimi ve değerlendirmesi gerçekleştirilmiştir. Farklı algoritmaların performansları karşılaştırılarak en etkili modelin belirlenmesi amaçlanmıştır.

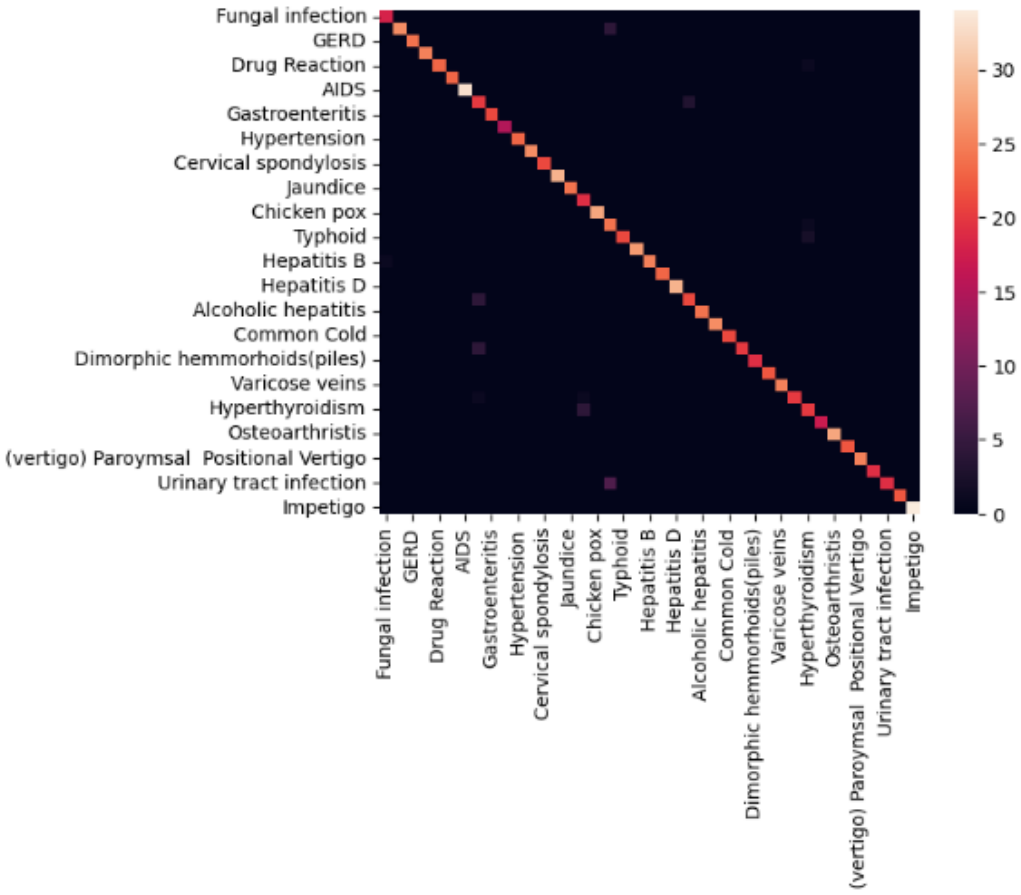
4.1 Veri Setinin Hazırlanması

Makine öğrenmesi modellerinin eğitimine geçmeden önce, veri seti üzerinde bazı ön işlemler gerçekleştirildi. Hastalık teşhisi için kullanılacak olan giriş verileri ve hedef etiketler ayrıldı. Veri seti, eğitim ve test veri setlerine ayrılarak modele uygun hale getirildi. Ayrıca, sınıflandırma algoritmalarının kullanabilmesi için kategorik veriler sayısal forma dönüştürüldü.

4.2 Decision Tree Algoritması ile Model Eğitimi ve Değerlendirme

4.2.1 Eğitim

Veri seti üzerinde ilk olarak Decision Tree sınıflandırma algoritması kullanılarak bir model eğitildi. Eğitilen model, test verileri üzerinde değerlendirildi ve elde edilen sonuçlar bir karışıklık matrisi ve sınıflandırma metrikleri ile görselleştirildi (bakınız Şekil 4.1).



Şekil 4.1: Hastalıkların semptom ağırlık dağılımları

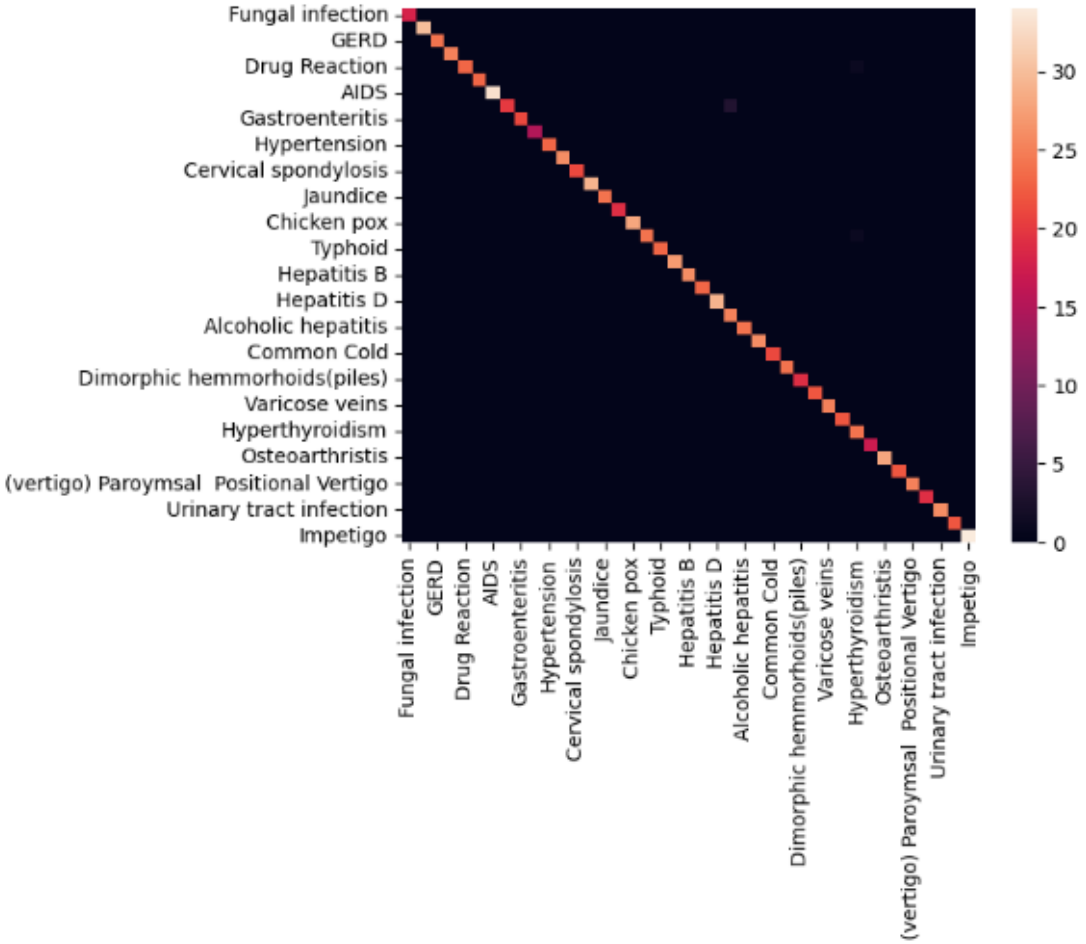
4.2.2 Değerlendirme

Decision Tree modelinin performansı, K-Fold Cross Validation kullanılarak daha geniş bir perspektiften değerlendirildi. Eğitim ve test doğrulukları ile standart sapma analizi sonuçları grafiklerle görselleştirildi.

4.3 Random Forest Algoritması ile Model Eğitimi ve Değerlendirme

4.3.1 Eğitim

Hazırlanan eğitim ve test verileri ile Random Forest Algoritması kullanılarak benzer şekilde bir başka model eğitildi ve test edildi. Ardından sınıflandırma performansını değerlendirmek için bir karışıklık matrisi oluşturuldu. Bu matris, modelin hangi sınıfları ne kadar doğru tahmin ettiğini görselleştirmesi amacıyla bir ısı haritasına dönüştürüldü (bakınız Şekil 4.2).



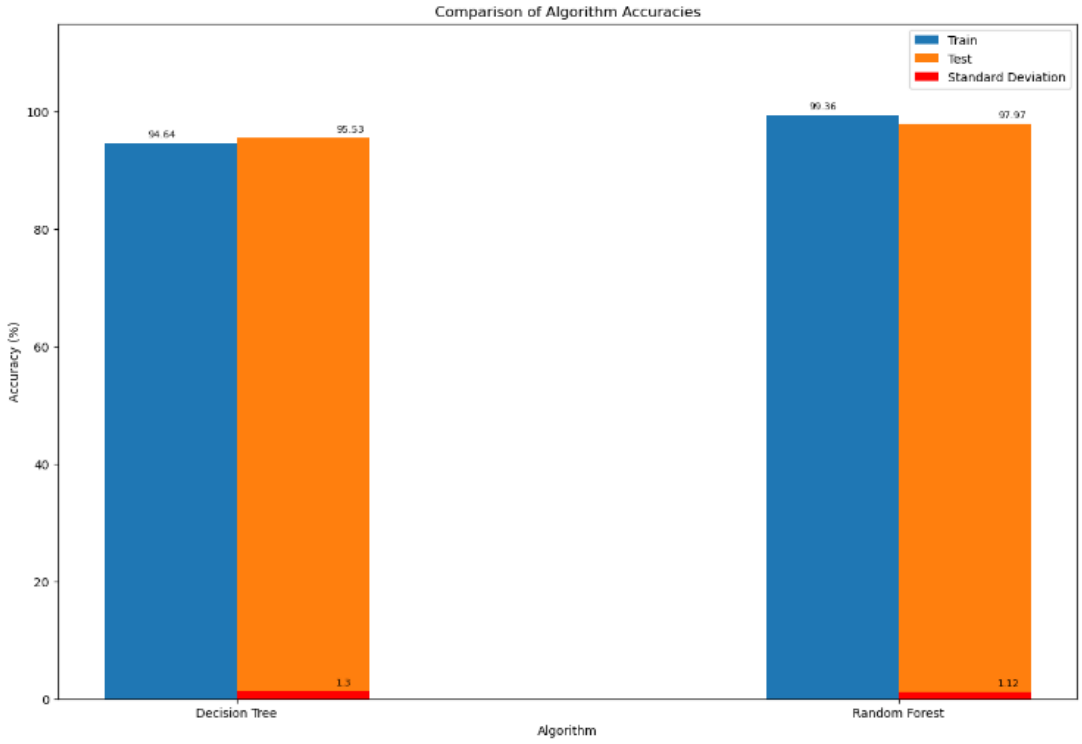
Şekil 4.2: Hastalıkların semptom ağırlık dağılımları

4.3.2 Değerlendirme

Random Forest algoritması kullanılarak yapılan eğitim ve test süreçleri K-Fold Cross Validation (çapraz doğrulama) ile hem test hem de eğitim veri seti üzerinde ayrı ayrı değerlendirildi ve sonuçlar incelendi.

4.4 Algoritma Perfomanslarının Karşılaştırılması

Bu bölümde, proje kapsamında kullanılan iki önemli sınıflandırma algoritması olan Decision Tree ve Random Forest'ın performansları karşılaştırmalı olarak değerlendirilmiştir. Her iki algoritmanın hastalık teşhisi yapma yeteneklerini karşılaştırarak, hangi algoritmanın daha etkili olduğunu belirlenmeye çalışılmıştır (bakınız Şekil 4.3).



Şekil 4.3: Hastalıkların semptom ağırlık dağılımları

Random Forest algoritması, eğitim veri seti üzerindeki çapraz doğrulama sonuçlarına göre oldukça etkileyici bir performans sergilemiştir. Ortalama doğruluk oranı %99.365 ve standart sapma sadece %0.47'dir. Test veri setinde de oldukça yüksek bir doğruluk

oranı (%97.966) ve düşük standart sapma (%1.12) elde edilmiştir. Bu sonuçlar, Random Forest algoritmasının genelde yüksek bir doğruluk sağladığını ve modelin stabilize olduğunu göstermektedir.

Decision Tree algoritması da çapraz doğrulama sonuçlarına göre başarılı bir performans göstermiştir, ancak Random Forest'a kıyasla biraz daha düşüktür. Eğitim veri setindeki ortalama doğruluk oranı %94.638 ve standart sapma %3.08 iken, test veri setinde bu oranlar sırasıyla %95.527 ve %1.30 olarak elde edilmiştir. Decision Tree, genelde yüksek doğruluk sağlamakla birlikte, Random Forest'a kıyasla biraz daha değişken bir performans sergilemiştir. Dolayısıyla hastalık tahminlemesi için Random Forest modelinin kullanılmasına karar verilmiştir.

Bölüm 5

Sonuç

Bu bitirme projesi, hastalık teşhisi alanında makine öğrenimi modellerinin kullanımını ve semptom ağırlıklarının analizini içeren detaylı bir çalışmayı kapsamaktadır. Projede gerçekleştirilen temel adımlar ve elde edilen sonuçlar şu şekildedir:

- Projenin ilk aşamasında, hastalık teşhisi veri setindeki eksik değerler düzenlenmiş ve semptom ağırlıklarını içeren yardımcı bir veri seti kullanılarak ana veri seti güncellenmiştir. Veri manipülasyonu süreci, veri bütünlüğünü sağlamak ve analiz için uygun bir veri seti elde etmek adına önemli bir adımdır.

- Hastalıklara göre en etkili semptomları belirlemek amacıyla bar grafikleri kullanılmıştır. Bu grafikler, her bir hastalığın hangi semptomları daha belirgin bir şekilde gösterdiğini anlamak için kullanılmıştır. Ayrıca, hastalıklara göre semptom ağırlıklarının ortalama değerlerini içeren bir ısı haritası oluşturulmuş ve görsel bir şekilde sunulmuştur.

- Decision Tree ve Random Forest algoritmaları kullanılarak hastalık teşhisi modelleri eğitilmiştir. Her iki modelin performansı, eğitim ve test doğrulukları incelenerek karşılaştırılmıştır. Random Forest modelinin, k-fold çapraz doğrulama sonuçlarına göre yüksek bir ortalama doğruluk sağladığı gözlemlenmiştir.

- Projede elde edilen sonuçlar, hastalık teşhisi için makine öğrenimi modellerinin başarıyla kullanılabileceğini göstermektedir. Random Forest modelinin yüksek doğruluk oranları, genel performansının güçlü olduğunu göstermektedir. Ancak, bu sonuçlara dayanarak karar ağacı modelinin de etkili bir performans sergilediği söylenebilir.

Bu çalışmanın sonuçlarına dayanarak, hastalık teşhisi alanında daha ileri çalışmalar yapılabilir. Özellikle, modelin daha geniş ve çeşitli bir veri seti üzerinde test edilmesi, modelin genelgeçerliğini artırabilir. Ayrıca, farklı özellik setleri ve daha gelişmiş makine öğrenimi tekniklerinin kullanılmasıyla daha kesin ve güvenilir modellerin geliştirilmesi düşünülebilir.

Kaynaklar

Veri Seti:

Kaggle, “Disease Symptom Prediction”, erişim: May 2017, https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset

Dergi:

Akalın, B. & Veranyurt, Ü. (2021). Sağlık Hizmetleri ve Yönetiminde Yapay Zekâ. *Acta Infologica, 5(1), 231-240*. https://dergipark.org.tr/en/download/article-file/1479836

Lisansüstü tez örneği:

Kılıç, N. (2023). Makine öğrenimi algoritmaları ile kredi kartı işlemlerinde dolandırıcılık tespiti (Yayın No. 792733) [Yüksek lisans tezi, Hitit Üniversitesi]. YÖK Ulusal Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Albaramani, H. H. A. (2023). Intrusion detection system using machine learning (Yayın No. 832344) [Yüksek lisans tezi, Bahçeşehir Üniversitesi]. YÖK Ulusal Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Arslan, E. (2023). Yapay zekâ destekli karar destek sistemlerinin geliştirilmesi: Çok katmanlı ERP yazılım mimari uygulaması (Yayın No. 825296) [Doktora tezi, İstanbul Aydın Üniversitesi]. YÖK Ulusal Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Mohamed, M. Y. (2015). Comparison of classification algorithms for predicting diabetes (Yayın No. 387459) [Yüksek lisans tezi, Erciyes Üniversitesi]. YÖK Ulusal Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/