



**T.C.
GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ**

**YÜKSEK
LİSANS
TEZİ**

**KALP HASTALIKLARINI TAHMİN ETMEDE VERİ
MADENCİLİĞİ TEKNİKLERİYLE ETKİLİ
ALGORİTMANIN TESPİT EDİLMESİ**

AYŞEGÜL NART

SAĞLIK BİLİŞİMİ ANA BİLİM DALI

ŞUBAT 2023



**KALP HASTALIKLARINI TAHMİN ETMEDE VERİ MADENCİLİĞİ
TEKNİKLERİYLE ETKİLİ ALGORİTMANIN TESPİT EDİLMESİ**

Ayşegül NART

**YÜKSEK LİSANS TEZİ
SAĞLIK BİLİŞİMİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
BİLİŞİM ENSTİTÜSÜ**

ŞUBAT 2023

Ayşegül NART tarafından hazırlanan “KALP HASTALIKLARINI TAHMİN ETMEDE VERİ MADENCİLİĞİ TEKNİKLERİYLE ETKİLİ ALGORİTMANIN TESPİT EDİLMESİ” adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi Sağlık Bilişimi Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Danışman: Prof. Dr. Mehmet Ali ERGÜN

Sağlık Bilişimi Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Başkan: Doç. Dr. Ceren SUCULARLI

Sağlık Bilimleri Enstitüsü, Hacettepe Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Üye: Doç. Dr. Oktay YILDIZ

Bilgisayar Mühendisliği Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum.

.....

Tez Savunma Tarihi: 24/02/2023

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

.....

Prof. Dr. Aslıhan TÜFEKÇİ

Bilişim Enstitüsü Müdürü

ETİK BEYAN

Gazi Üniversitesi Bilişim Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Ayşegül NART

...../...../.....

KALP HASTALIKLARINI TAHMİN ETMEDE VERİ MADENCİLİĞİ TEKNİKLERİYLE ETKİLİ ALGORİTMANIN TESPİT EDİLMESİ

(Yüksek Lisans Tezi)

Ayşegül NART

GAZİ ÜNİVERSİTESİ

BİLİŞİM ENSTİTÜSÜ

Şubat 2023

ÖZET

Teknolojik gelişmelerin artmasıyla birlikte hayat standartları yükselmiş, bunun sonucunda insan ömrü uzamıştır. Bu durum beraberinde yaşlı nüfus artışı, beslenme alışkanlıklarının değişmesi ile kalp hastalıkları ve çeşitli ölümcül hastalıkların ortaya çıkmasına neden olmuştur. Son yıllarda yapılan araştırmalara göre; kalp hastalıkları dünyada insan ölümlerinin en büyük nedenlerinden biri olarak bilinmektedir. Bu nedenle kalp hastalıklarının teşhis edilerek tedavi edilmesi, ölüm oranlarının azalmasında önemli bir yere sahiptir. Bu kapsamda; hem ölüm oranlarının yüksek olması ve erken teşhis ile bu hastalıkların önlenabiliyor olması hem de çok sayıda kalp hastalığı teşhisi konulan hastadan elde edilen büyük bir verinin olması bu konun ile ilgili çalışmaların yapılmasına olanak sağlamaktadır. Ayrıca bu büyük verinin varlığı veri madenciliği teknikleri aracılığıyla işlenmeye oldukça uygundur. Bu çalışmada; kalp hastalıklarını tahmin etmede kullanılabilecek en etkili algoritma ya da algoritmaları tespit edebilen java programlama diliyle android tabanlı mobil bir uygulama geliştirilmiştir. Yapılan bu uygulama ile cep telefonu üzerinden uygun veri seti yüklenerek algoritmaların doğru sonucu tespit etme başarısı hesaplanmaktadır. Bu uygulama kullanılarak, sağlık çalışanları ve bireyler kalp sağlıklarını daha iyi anlayabilir ve yönetebilirler, ve potansiyel olarak kalp hastalıklarının gelişme riskini azaltabilirler. Ayrıca, android tabanlı bir platform kullanımı, uygulamanın çeşitli cihazlardan kolay ve uygun bir şekilde erişilebilmesini sağlar, daha büyük bir izleyici kitlesine erişilebilirliği artırır. Genel olarak, bu uygulama, ileri veri analizi ve veri madenciliği tekniklerinin kullanımıyla, kalp hastalığı tanı ve tedavisinin iyileştirilmesine katkıda bulunmayı ve potansiyel olarak yaşamları kurtarma amacı güder.

Bilim Kodu : 92429

Anahtar Kelimeler : Kalp hastalığı, Naive Bayes, Logistik Regresyon, Android, Java.

Sayfa Adedi : 47

Danışman : Prof. Dr. Mehmet Ali ERGÜN

DETERMINATION OF EFFECTIVE ALGORITHM WITH DATA MINING
TECHNIQUES IN PREDICTING HEART DISEASES

(M. Sc. Thesis)

Ayşegül NART

GAZİ UNIVERSITY
INSTITUTE OF INFORMATICS

February 2023

ABSTRACT

With the increase in technological developments, living standards have risen and as a result, life expectancy has increased. This has led to an increase in the elderly population and changes in nutrition habits, resulting in heart diseases and various fatal diseases. According to recent research, heart disease is known to be one of the leading causes of human deaths worldwide. Therefore, the diagnosis and treatment of heart disease is important in reducing mortality rates. In this context, both the high mortality rates and the fact that these diseases can be prevented by early diagnosis, as well as the large amount of data obtained from a large number of patients diagnosed with heart disease, make it possible to carry out studies on this subject. In addition, the existence of this large data is very suitable for processing through data mining techniques. In this study, an android-based mobile application that can identify the most effective algorithm or algorithms that can be used to predict heart disease was developed. With this application, the accuracy of algorithms in correctly detecting the appropriate data set used in the diagnosis of heart disease is calculated using a mobile phone. By using this application, healthcare professionals and individuals can better understand and manage their heart health, and potentially reduce their risk of developing heart diseases. Additionally, the use of an android-based platform allows for convenient and easy access to the application from a wide range of devices, making it more accessible to a larger audience. Overall, this application aims to contribute to the improvement of heart disease diagnosis and treatment, and potentially save lives through the use of advanced data analysis and data mining techniques.

Keywords: Heart disease, Naive Bayes, Logistic Regression, Android.

Science Code : 92429

Key Words : Heart Disease, Naive Bayes, Logistic Regression , Android, Java.

Page Number : 47

Supervisor : Prof. Dr. Mehmet Ali ERGÜN

TEŞEKKÜR

Yüksek Lisans'a kabulüm aşamasından tez danışmanlığıma kadar geçen süreçte desteklerinden dolayı değerli hocam Prof. Dr. Mehmet Ali ERGÜN'e çok teşekkür ederim. Hayatımın her aşamasında desteğine esirgemeyen daima arkamda olan sevgili aileme ve bugüne kadar almış olduğum eğitime katkı sağlayan bütün hocalarıma desteklerinden ötürü çok teşekkür ederim.

Ayşegül NART

Ankara,2023

İÇİNDEKİLER

	Sayfa
ÖZET.....	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
SİMGELER VE KISALTMALAR.....	xvi
1. GİRİŞ VE AMAÇ.....	1
2. LİTERATÜR TARAMASI.....	3
3. GENEL KAVRAMLAR.....	9
3.1. Kalp Hastalığı.....	9
3.1.1. Kalp Hastalığı Türleri ve Ölüm Oranları.....	11
3.2. Veri Madenciliği.....	12
3.3. Veri Madenciliği Algoritmaları.....	15
3.3.1. Naive Bayes.....	15
3.3.2 Karar Ağacı.....	17
3.3.3. K En-Yakın Komşu Algoritması	19
3.3.4. Lojistik Regresyon Algoritması.....	19
3.3.5. Destek Vektör Makinesi	20
3.3.5. Yapay Sinir Ağları	20
3.4. Veri Madenciliğinde Kullanılan Programlar	21
3.4.1. RapidMiner Studio.....	21
3.4.2. R.....	22
3.4.3. Spss	23
3.4.4. Weka	23
3.4.5. Python	24
3.4.6. Matlab	25
3.4.7. Tanagra	26
3.4.8. Orange.....	26
3.4.9. Knime.....	27
3.4.10. SAS	27
3.5. Veri Setleri	28

4. YÖNTEM.....	31
4.1. Veriler ve Toplanması.....	31
4.2. Model Performans Ölçütleri	32
5. UYGULAMA.....	35
6. SONUÇ	37
KAYNAKLAR.....	41
ÖZGEÇMİŞ.....	45



ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 3.1. Veri seti örneği.....	29
Çizelge 4.1. Öznitelik tablosu.....	31
Çizelge 5.1. Algoritma Performansları.....	37
Çizelge 5.2. Karışıklık Matrisi.....	37
Çizelge 5.3. KNN Algoritması Karışıklık Matrisi.....	37
Çizelge 5.4. Naive Bayes Algoritması Karışıklık Matrisi.....	37
Çizelge 5.5. Lojistik Regresyon Algoritması Karışıklık Matrisi.....	37

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 3.1. Kalbin Yapısı.....	9
Şekil 3. 2. Dolaşım sistemi hastalıklarından ölenlerin oranı.....	12
Şekil 3. 3. Bilgi Piramidi.....	14
Şekil 3. 4. Örnek karar ağacı.....	17
Şekil 3. 5. RapidMiner Ekran Görünümü.....	22
Şekil 3. 6. R programına ilişkin ekran görüntüsü	23
Şekil 3. 7. Weka programına ilişkin ekran görüntüsü.....	24
Şekil 5.1. Veri analizi dosya seçme ara yüz görüntüler.....	35
Şekil 5.2.Veriseti bilgisinin ara yüz görünümü.....	36
Şekil 5.3.Hesaplanan değerin ara yüzde gösterilmesi.....	36

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

g

Gram

ml

Mililitre

Kısaltmalar

Açıklamalar

ESK

Elektronik Sağlık Kayıtları

KNN

K-En Yakın Komşu

TÜİK

Türkiye İstatistik Kurumu

1. GİRİŞ VE AMAÇ

İnsanlık var olduğundan beri insanlar uzun yaşamın sırrını aramışlardır. Bu amaçla çeşitli yöntemlerle ve gelişen teknolojilerle beraber insan ömrünü uzatmaya başarmışlardır. Ancak insan ömrünün uzamasıyla birlikte yaşlı nüfusu artış göstermiş ve buna bağlı olarak kalp ve damar hastalarının sayısında da artış gözlemlenmiştir[1]. Kalp hastalıklarının artması ülkelerin sağlık yükünün oluşmasına neden olmaktadır. Bu yük, hem ekonomik hem de sosyal açıdan büyük bir etkiye sahiptir. Ekonomik açıdan, kalp hastalıklarının tedavisi ve bakımı oldukça maliyetli olabilmektedir. Sosyal açıdan ise, bu hastalıklar insanların iş gücünden ayrılmasına ve hayat kalitesinin düşmesine neden olabilmektedir. Bu nedenle ülkelerin sağlık sistemleri, kalp hastalıklarının önlenmesi ve tedavisi için gerekli tedbirleri almalıdır.

TÜİK verilerine göre 2019 yılında, dolaşım sistemi hastalıklarından kaynaklı ölümler alt ölüm nedenlerine göre incelendiğinde, ölenlerin %39,1'inin iskemik kalp hastalığından, %22,2'sinin serebro-vasküler hastalıklardan, %25,7'sinin diğer kalp hastalıklarından öldüğü görülmüştür[2]. Tüm bu durumlar göz önüne alındığında kalp hastalıklarını önlemek için çeşitli araştırmalar yapılması ihtiyacı doğmuştur.

Son yıllarda gelişen teknolojiler sayesinde kağıt kayıtlardan Elektronik Sağlık Kayıtları(ESK) sistemlerine geçiş ile birlikte hastalara ait tanı, tedavi, kullanılan ilaçlar, alerjik durumlar ve laboratuvar vb. bilgilerin elektronik ortamda tutulabilmektedir. Örneğin, elektronik sağlık kayıtları (EHR) ve internet tabanlı araçlar sayesinde hastaların tıbbi geçmişleri ve sağlık durumları daha kolay ve daha hızlı toplanabilmektedir. Ayrıca, çeşitli sensörler ve cihazlar sayesinde hastaların canlı olarak takip edilmesi ve verilerin gerçek zamanlı olarak toplanması mümkün hale gelmiştir. Bu sistemlerin dünya genelinde yaygın olması hastalıklara yönelik tedavi yöntemlerinin geliştirilmesine de olanak sağlamıştır.

Böylelikle dünyada en önemli ölümcül hastalıklardan biri olan kalp hastalığı verisinin çok sayıda hastadan toplanabilmesine olanak sağlanmıştır. Elde edilen veriler, veri madenciliği yöntemleri ile işlenerek, kalp hastalığının risk faktörleri, erken tanı ve tedavi seçenekleri gibi önemli bilgiler elde edilebilmektedir. Özellikle, yapay zeka ve makine öğrenimi teknolojileri ile birlikte veri madenciliği, kalp hastalığı ile ilgili çalışmalar için önemli bir rol

oynamaktadır. Örneğin, veri madenciliği yöntemleri kullanarak, risk faktörleri belirlenerek, hastalığın önlenmesi veya erken tanısı için uygun stratejiler geliştirilebilmektedir. Ayrıca, veri madenciliği yöntemleri kullanarak, tedavi seçenekleri veya ilaçların etkililiği hakkında da bilgi elde edilebilmektedir. Bu sayede hastalığın önlenmesi veya erken tanısı yapılması ve tedavisi konusunda daha etkili kararlar alınabilecektir.

Tez çalışmasında büyük verilerden anlamlı veriye erişmemizi sağlayan veri madenciliği algoritmaları kullanılarak kalp hastalığı teşhisinde kullanılabilecek etkili algoritma ya da algoritmaların tespiti ile karar destek sistemlerinin geliştirilmesine katkı sağlamak amaçlanmaktadır.

Tez çalışmasında hastalara ait verilerin sağlık kuruluşlarından elde edilmemesi bu çalışmanın sınırlılıklarındandır. Bu nedenle uygulamada kullanılan veri seti daha önceki yıllarda oluşturulan hasta kayıtlarına ilişkin bilgilerden oluşmaktadır.

Tez çalışmasında; mevcut durumda kullanılan programlar yerine internet alt yapısına ihtiyaç duymaksızın mobil cihazlarda çalışabilen android tabanlı bir uygulama geliştirilmiştir ve ilgili çalışmalar bu uygulama üzerinden gerçekleştirilmiştir.

2. LİTERATÜR TARAMASI

Literatürde kalp hastalığının veri madenciliği algoritmaları ile tespit edilmesi konusunda çok sayıda çalışma yapılmıştır. Bu çalışmalarda farklı veri setleri kullanılmış ve algoritma ya da algoritmaların farklı kombinasyonlarından yararlanılarak birbirinden farklı sonuçlar elde edilmiştir.

H. Ayetullahi ve arkadaşlarının[3] yaptığı çalışmada R kullanılarak üç eğitim hastanesinde yatan koroner arter hastalığı olan hastaların tıbbi kayıtları ile yapay sinir ağları ve destek vektör makinelerinin hastalığı tespit etmedeki başarısı karşılaştırılmıştır. Çalışma sonucunda destek vektör makinesinin yapay sinir ağlarından daha başarılı olduğu tespit edilmiştir.

S. Joshi ve arkadaşları[4] Naive Bayes ve lojistik regresyon algoritmalarını UCI Machine Learning Repository 'den aldıkları Cleveland, Hungarian ve Long Beach veri setine Python programlama dili kullanılarak uygulamışlardır. Yapılan çalışmalar sonucunda en yüksek doğruluk oranı veren algoritmanın lojistik regresyon olduğu görülmüştür.

D. Ananey Obiri ve arkadaşlarına[5] ait çalışmada Python kullanarak UCI Machine Learning Repository 'den aldıkları 14 özellik ve 303 hasta kaydını içeren veri seti lojistik lineer regresyon, karar ağacı sınıflandırıcı, Gaussian Naive Bayes algoritmaları ile test edilmiştir. Üç algoritmanın değerlendirmesinde kullanılan performans metrikleri, Gaussian Naive Bayes algoritmasının diğer üç sınıflandırıcıdan daha iyi olduğunu göstermiştir.

S. Anitha ve arkadaşları[6] yaptıkları çalışmada veri setinin %70 kısmını eğitim veri seti, %30 kısmını ise test veri seti olarak ayrılmıştır. Veri madenciliği sınıflandırma algoritmaları olan KNN, Naive Bayes ve destek vektör makineleri ilgili veri setine R programlama kullanılarak uygulanmıştır. Deney sonuçlarına göre Naive Bayes algoritması %86.6 doğruluk oranıyla kalp hastalığını tahmin etmiştir.

İ. Özcan ve arkadaşları[7] MATLAB 2017b kullanılarak UCI Machine Learning Repository veri bankasına ait Statlog (Heart) Data Set veri tabanından kalp hastalığı ile ilgili 170 hastaya ait veri setine yapay sinir ağları ve destek vektör makineleri algoritmaları uygulanmıştır. Bu çalışma sonucunda kalp hastalığını tahmin etmede destek vektör makineleri %91.67 oranıyla

yüksek başarı elde etmiştir.

S. Bahadur Patel ve arkadaşları[8] veri setinde bulunan 14 özelliği genetik algoritma ile 6 özelliğe indirgemıştır. İşlemeye hazır hale gelen veri sınıflandırma yöntemlerinden Naive Bayes, karar ağacı ve kümelemeye göre sınıflandırma algoritmalarıyla işlenmiş ve çalışma sonunda hastalık tespitinde en etkili algoritmanın karar ağacı olduğu gözlemlenmiştir.

M. Akhil Jabbar ve arkadaşları[9] daha etkili bir sınıflandırma oluşturmak için KNN ve genetik algoritmayı birleştirmiştir. UCI Machine Learning Repository’de bulunan 7 veri setiyle yapılan deneyde KNN, genetik algoritma ve KNN ile genetik algoritmanın birleştirilmesiyle oluşturulan algoritmalar arasında en yüksek doğruluk oranı iki algoritmanın birleşimiyle oluşan model olmuştur.

M. Shafenoor Amin ve arkadaşları[10] yaptıkları çalışmada UCI Machine Learning Repository’de bulunan Cleveland veri tabanındaki verilerle 7 veri madenciliği algoritmasının doğruluk oranlarını karşılaştırmıştır. Kullanılan algoritmalar KNN, karar ağacı, Naive Bayes, lojistik regresyon, destek vektör makinesi, sinir ağları ve Vote adı verdikleri Naive Bayes ve Lojistik Regresyon algoritmalarının birlikte kullanılmasıyla oluşan algoritma seçilmiştir. Tüm veri seti 10 alt kümeye bölünmüş ve daha sonra 10 kez işlenmiştir. 9 alt küme test kümesi olarak, kalan 1 alt küme ise eğitim kümesi olarak kullanılmıştır. Son olarak, sonuçlar her 10 yinelemenin ortalaması alınarak gösterilmiştir. Çalışma sonucunda Naive Bayes, lojistik regresyon ve Vote yüksek doğruluk gösteren algoritmalar ve araştırmada önerilen Vote algoritması en yüksek doğruluğu veren algoritma olmuştur.

A. Rajkumar, G.Sophia Reena ve arkadaşları[11] çalışmalarında Tanagra programı ile Naive Bayes, KNN ve Karar listesi algoritmalarını kalp hastalığı teşhisi bakımından karşılaştırmışlardır. 14 özelliğe sahip 3000 kayıttan oluşan veri seti kullanılmıştır. Veri seti iki bölüme ayrılarak %30 test, %70 ise eğitim seti için ayrılmıştır. Yapılan çalışmalar sonucunda, %52,33 oranıyla Naive Bayes algoritmasının diğer algoritmalara göre daha etkili olduğu gözlenmiştir.

D. Chandna[12] yaptığı çalışmada kalp hastalığı teşhisi için özellik seçimleri ve bilgi kazanımı (Information Gain) ile yapılmış ve uyarlanabilir nöro bulanık çıkarım sistemi

(Adaptive Neuro Fuzzy Inference System(ANFIS)) kullanarak WEKA ve MATLAB programları aracılığıyla yapılmıştır. UCI'dan alınan 14 özelliğe sahip veri seti kullanılmıştır. Çalışma sonucunda bu iki yöntemin birlikte kullanıldığı yaklaşımla %98,24 doğruluk elde edilmiştir.

S. Palaniappan ve arkadaşları[13] karar ağaçları, Naive Bayes ve sinir ağı olmak üzere üç veri madenciliği modelleme tekniğini kullanarak bir prototip Akıllı Kalp Hastalığı Tahmin Sistemi (IHDPs) geliştirmeyi amaçlamaktadır. Cleveland veri tabanından 909 hasta kaydı 455'i eğitim veri seti ve 454'ü test veri seti olarak ayrılarak kullanılmıştır. Veri madenciliği algoritmalarını uygulayabilmek için DMX(Data Mining Extension) seçilmiştir. Çalışma sonunda Naive Bayes'in en etkili model olduğu görülmüştür.

M. Abdar ve arkadaşları[14] UCI'dan alınan 13 özellikli 270 kayıt bulunan verilerle karar ağacı, sinir ağları, destek vektör makineleri ve KNN algoritmaları kullanılmıştır. Veri %70 eğitim, %30 test seti olarak bölünmüştür. Karar ağacı, %93.02 ile en yüksek doğruluk oranı göstermiştir. KNN %88.37, SVM %86.05, sinir ağları %80.23 şeklinde hesaplanmıştır.

Mamatha Alex P. ve arkadaşları [15] yaptığı çalışmada veri seti olarak yaklaşık 2200 hastaya ait kaydı gerçek zamanlı olarak hastalarla bire bir görüşerek ve hasta taburcu özetlerinden elde etmişlerdir. Bu yöntemlerle 20 öznitelik oluşturulmuştur. Destek vektör makinesi, rastgele orman, KNN ve yapay sinir ağları algoritmalarını bu veri setine uyguladıklarında %92.21 oranıyla en fazla doğruluk yapay sinir ağlarında elde edilmiştir.

B. Bahram ve arkadaşları[16] yaptıkları çalışmalarında İran'da bir hastaneden alınan 8 özellikli 209 hasta kaydını kullanmışlardır. İlgili veri setinde kayıp veri bulunmamaktadır. Gain Ratio Attribute Evaluation(Kazanç Oranı Nitelik Değerlendirmesi) yöntemiyle özellik seçilimi yapılmıştır. 4 özellik elenmiş ve kalan 4 özellik ile Weka üzerinden J48 karar ağacı, Naive Bayes, KNN ve Sequential Minimal Optimization(Sıralı Minimal Optimizasyon) algoritmalarının doğruluk oranları karşılaştırılmıştır. Bu karşılaştırma sonucunda karar ağacı %83.73 doğruluk oranıyla diğer algoritmalara göre daha yüksek performans göstermiştir.

N. Bhatla ve arkadaşları[17] çalışmalarında özellik seçimi için genetik algoritmaları ve etkili sınıflandırma için bulanık uzman sistemi birleştiren bir modelin geliştirmişlerdir. Özellik seçimi için Matlab kullanılmış ve UCI Machine Learning Repository'de yer alan veri

setindeki 6 özelliğin etkili olduğu gözlemlenmiştir. Weka 3.6.6 üzerinden Naive Bayes, karar ağacı, kümeleme ve sinir ağları algoritmaları karşılaştırıldığında; sinir ağlarının diğerlerine göre daha doğru sonuca ulaştığı ve karar ağacının genetik algoritması ile özellik seçimi yapıldığında iyi bir performans gösterdiği görülmüştür.

A. Ahmed ve arkadaşları[18] yaptıkları bu çalışmada UCI veri tabanında bulunan 4 farklı veri setini kullanmışlardır. Bu veri setleri 920 kayıt ve 76 özellikten oluşmakta olup, çalışmada yalnızca 13 özelliğe yer verilmiştir. İlgili veri setleri üzerinde destek vektör makinesi, genetik algoritma, kaba küme teorisi, birliktelik kuralları ve sinir ağları algoritmaları karşılaştırılmıştır. Karşılaştırdıkları algoritmalar arasında karar ağacı ve destek vektör makinelerinin en yüksek oranla sonuç verdiğini gözlemlemişlerdir.

A. Nguetilbaye ve arkadaşları[19] çalışmalarında hastanelerden aldıkları 150 sağlıklı ve 165 kalp hastasına ait 14 özellikli bir veri seti kullanmışlardır. Kullanıcı arayüzleri tasarlamak için Visual Studio kullanılırken, programlar C# diliyle yazılmış ve veri analizi Python programlama dili yapılmıştır. Geliştirdikleri bu karar destek sistemi ile Naive Bayes ve destek vektör makineleri algoritmalarını karşılaştırmışlar ve bunun sonucunda Naive Bayes algoritmasının %91.42 doğruluk oranıyla en iyi sonucu vermiştir.

V. Chaurasia ve arkadaşları[20] Macaristan Kardiyoloji Enstitüsünün veri setini kullanmışlardır. Weka programı ile Naive Bayes, J48 karar ağacı ve torbalama(bagging) algoritmaları bu 294'ten 11 özelliğe düşürülen veri setine uygulanmıştır. Bu çalışma sonucunda torbalama algoritması %85.03 doğruluk oranıyla diğer algoritmalar göre daha başarılı performans göstermiştir.

U. Shafique ve arkadaşları[21] yaptıkları bu çalışmada, UCI Machine Learning Repository'den eriştikleri 76 özellikli 303 hasta kaydını en önemli 14 özellik ile Macaristan Kardiyoloji Enstitüsünden aldıkları 294 hasta kaydını ekleyerek toplamda 597 hasta kaydına ulaşmışlardır. Weka programı ile karar ağacı, yapay sinir ağları ve Naive Bayes algoritmaları karşılaştırılmıştır. Bu deney sonucunda kalp hastalığı tahmin etmede en iyi sonucu veren algoritmanın %82.914 ile Naive Bayes olduğu görülmüştür.

B.Venkatalakshmi ve arkadaşları[22] çalışmalarında Weka programı kullanarak UCI Machine Learning Repository'den alınan 13 özellikli, 294 hasta kaydından oluşan veri seti

üzerine Naive Bayes ve karar ağacı algoritmaları karşılaştırılmıştır. Naive Bayes algoritmasının karar ağacına göre daha iyi sonuç verdiği görülmüştür.

P. Bajaj ve arkadaşları[23] çalışmalarında kalp hastalığı tahmini için hibrit bir model geliştirmişlerdir. Bu modelde genetik algoritma ile normal sinir ağları birleştirilmiş ve oluşturulan model ile sinir ağlarının performans durumları incelenmiştir. Oluşturulan bu hibrit modelin sinir ağlarına göre daha iyi sonuç verdiği gözlemlenmiştir.

N. Chaithra ve arkadaşları[24] yaptıkları çalışmada Ekokardiyografi veri tabanından 24 özniteliğe sahip 336 kayıt bulunan veri seti üzerinde J48 karar ağacı, Naive Bayes ve sinir ağı algoritmalarının kalp hastalığını tahmin etmede ne kadar başarılı oranlarını karşılaştırmışlardır. Çalışma sonucunda % 97.91 doğruluk oranıyla sinir ağları diğer algoritmalarından yüksek performans göstermiştir.

A. Taneja[25] yaptığı çalışmada Weka uygulaması ile PGI, Chandigarh'dan alınan 2008-2011 yılları arasında toplanan ve 7.339 kayıt içeren veri seti kullanılmıştır. J48 sınıflandırıcı, Naive Bayes ve çok katmanlı algılama algoritmaları karşılaştırıldığında %95,56 sınıflandırma doğruluğu ile J48 sınıflandırıcı en yüksek doğruluğu elde etmiştir.

C. S. Dangare ve arkadaşları[26] çalışmalarında Cleveland ve Statlog veri tabanlarında bulunan 303 ve 270 kayıtlı toplam 573 hasta kaydı bulunan 13 öz nitelikli veri setini kullanmışlardır. Bu veri seti eğitim ve test veri seti olarak ikiye ayrılmıştır. 303 kayıt eğitim seti, 270 kayıt test veri seti olarak seçilmiştir. Weka 3.6.6 uygulaması üzerinde Naive Bayes, karar ağacı ve sinir ağları algoritmaları ilgili veri setine uygulandığında sinir ağları algoritmasının daha yüksek oranlı doğruluğa sahip olduğu gözlemlenmiştir.

P. Singh ve arkadaşları[27] hazırladıkları çalışmada kalp hastalığı için halka açık bir veri tabanı üzerinde gerçekleştirilmiştir. Veri seti, eğitim seti (%40) ve test seti (%60) olmak üzere iki sete ayrılan toplam 303 kayıt ve 15 öz nitelik içermektedir. Deney için Weka 3.6.11 isimli bir veri madenciliği aracı kullanılmıştır. Eğitim algoritması olarak geri yayımlı (BP) çok katmanlı algılayıcı sinir ağı (MLPNN) kullanılmıştır. Eğitim setine uygulanan bu işlemlerden sonra sinir ağlarının kalp hastalığını tahmin etmede yüksek performans sergilediğini göstermiştir.

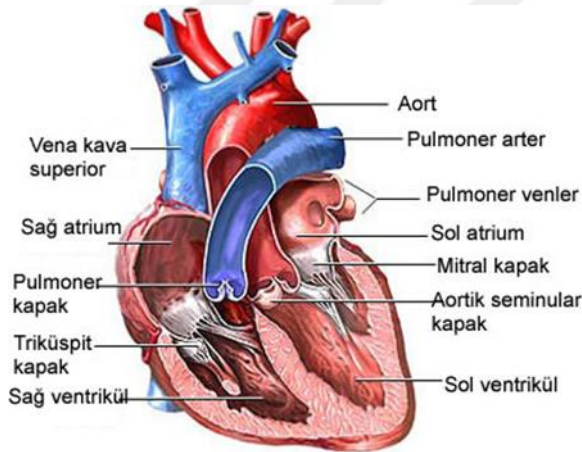


3. GENEL KAVRAMLAR

3.1. Kalp Hastalığı

Kalp bir pompa gibi kanı dokulara dağıtma görevi gören, kas ve bağ dokularından oluşan, yaklaşık kişinin yumruğu büyüklüğünde olan bir organdır. Vücudumuzda, göğüs kafesinin altında, diyaframın üzerinde akciğerlerin arasında sol tarafa daha yakın bir konumdadır. Ağırlığı yaklaşık 250 g olan kalbin içinde 200-300 ml kan bulunur[28].

Kalp çizgili bir kas tabakasından yapılmıştır ve bu kas tabakasının üzerinde kan damarları vardır. Kalp boşluğunun dört göze ayrılmış olduğu, bu gözlerden yukardakilere kulakçık, alt taraftakilere ise karıncık adı verilir. Kalbin sağ kulakçığı ile sağ karıncığı arasında üçlü kapakçık, sol kulakçığı ile sol karıncığı arasında ikili kapakçık vardır. Bu kapakçıklar, kanın kulakçıklardan karıncıklara geçişine izin verdikleri halde karıncıklardan kulakçıklara geri geçmesini engelleyecek şekilde çalışır.



Şekil 3.1. Kalbin Yapısı[29]

Kalp hastalığı, kalpte meydana gelen ve kalbe etki eden bütün bozuklukları ifade eder. Koroner arter hastalığı gibi kan damar hastalıkları, kalp ritmi ile ilgili bozukluklar ve doğuştan gelen form farklılıkları da bu başlıkta ifade edilebilir.

Kalbin çalışması için gerekli yapılar; kardiyovasküler sistem, koroner arterler ve sinir ağıdır. Kanın kalp aracılığıyla tüm vücuda dağıtılmasını sağlamak kardiyovasküler sistem denilen arterler ve damarların görevidir. Koroner arterler ise kalbin yüzeyinde yayılarak kalp kasına oksijen zengin kanı taşıyan damarlardır. Kalbin kan pompalayabilmesi için yapması

gereken kasılma ve gevşeme hareketlerini ise sinyallerle sağlayan bir sinir ağı bulunmaktadır. Dolayısıyla bu üç ana yapıdaki bozukluklar kalp hastalıklarının oluşmasına neden olabilmektedir.

Kalp hastalıkları doğuştan gelebileceği gibi sonradan yaşam tarzı, beslenme rejimi ve vücuttaki diğer bozuklukların bir etkisi olarak da ortaya çıkabilmektedir. Kalp ritim bozuklukları, koroner arter hastalıkları, kalp kapakları ve karıncıklarında ki bozukluklar, kalp kasının oksijensiz kalması gibi hastalıklar temelde görülebilecek hastalıklar olarak belirtilebilir. Koroner arter hastalığının görülme sıklığı erkeklerde yüzde 3,8 kadınlarda yüzde 2,3'tür. Erkeklerde görülme sıklığı 50-60 yaşlarında iken kadınlarda ise östrojen hormonun azaldığı dönem olan 60-70 yaş aralığındadır. Tüm dünyada ve ülkemizde en sık ölüm sebebi kalp damar hastalıklarından kaynaklanmaktadır[30].

Tansiyon, kanın atardamar duvarına yaptığı basınç olarak tanımlanabilir. Tansiyonun yüksek olması ise kalp, böbrek ve beyin gibi hayati organ hastalıklarına zemin hazırlayabilen bir bozukluktur. Yüksek kan basıncına, tansiyona maruz kalan damar duvarı hasara uğrayarak, kandaki yağ partiküllerinin damar yüzeyine sıvanmasına ve birikmesine neden olmaktadır. Zaman içinde de biriken bu partiküller damarda darlığa veya tıkanmalara neden olabilmektedir. Kalp krizi yaygın olarak bilinen bir kalp rahatsızlığıdır. Kalp damarlarının tıkanması sonucu oluşabilmektedir. Yıllık olarak binde 6 oranında yaygınlığı bulunmaktadır.

Kişinin bir kalp rahatsızlığı olabileceğine dair belirtiler de önemle takip edilmesi gereken konulardan biridir. Göğüs bölgesindeki ağrı, sıkışma gibi rahatsızlıklar, nefes darlığı oluşması, bacak ve kollardaki ağrı, uyuşukluk ve halsizlik, sırtta ağrı, baş dönmesi ve bayılma, çok yavaş veya çok hızlı kalp atışı, ten renginde ki farklılıklar, göz çevresinde ki şişkinlikler, öksürük gibi genel belirtiler mutlaka üzerinde durulması gereken konulardır.

Bu ve benzeri şikâyetler mutlaka uzman kişilerce değerlendirilmelidir. Bunun için EKO, EKG ve BT, MRI gibi nükleer tıp teknikleri kullanılarak girişimsel olmayan tetkikler yapılabileceği gibi gerekli durumlarda ise koroner anjiyografi veya holter izleme gibi girişimsel tetkikler de yapılabilmektedir.

Yapılan tetkikler sonucunda gerekli durumlarda ilaç tedavisi bunun yeterli olmadığı durumlarda ise cerrahi tedaviler uygulanabilmektedir. Yaşam tarzının ve beslenme rejiminin

ise her durumda kontrol altında olması için düşük sodyum ve az yağlı gıdalar tüketmek, düzenli egzersizlerle kalp damarlarının ve ritminin sağlıklı olmasını sağlamaya çalışmak oldukça önemlidir.

Sigara kullanımı sadece akciğerlere değil aynı zamanda kalbe de büyük zararlar verebilmektedir. Bunun nedeni, kandaki pıhtılaşma faktörlerini harekete geçiriyor olması ve tansiyonu arttırmasıdır. Ayrıca kan damarlarındaki yağdan zengin madde baloncuklarının patlama sürecini başlatmasıdır. Sigara içenler içmeyenlere oranlara kalp hastalıkları açısından 3 kat daha riskli kabul edilmektedir.

Kolestrol, kan yağlarında ki yükseklik de en önemli risk faktörlerinden biri olarak kabul edilmektedir. Kan yağları kalp damarlarında darlık oluşturarak kalp krizi riskini arttırmaktadır. Kolestrol genetik olarak yatkınlıkla da ortaya çıkabileceği gibi, büyük oranda beslenme ve kilo ile ilgili ortaya çıkmaktadır.

Ayrıca stres, depresyon ve öfke, kalp hastalıklarına zemin hazırlayan bir başka husustur. Bu nedenle olabildiğince stres ve depresyondan uzak durmaya çalışmak gereklidir.

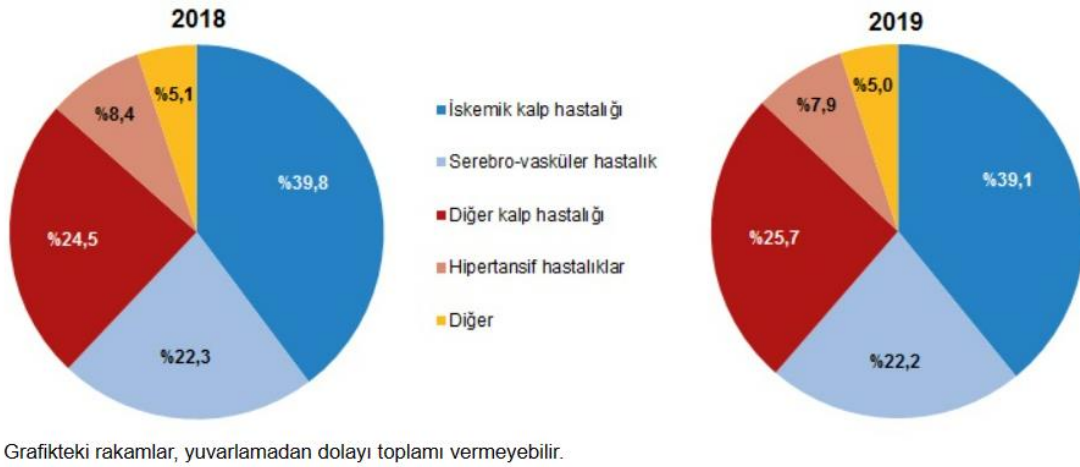
3.1.1. Kalp Hastalığı Türleri ve Ölüm Oranları

Kalp hastalıkları türleri;

- Koroner kalp hastalığı,
 - Serebrovasküler hastalık,
 - Romatizmal kalp hastalığı,
 - Konjenital kalp hastalığı,
 - Derin ven trombozu,
 - Plmoner emboli,
 - Kardiyomiyopati ve periferik arter hastalıkları
- şeklinde bilinmektedir[31].

Kalp hastalıkları dünyada ölüme neden olan hastalıkların başında gelmektedir. 2016 yılında yaklaşık 17,9 milyon kişinin kardiyovasküler hastalıklara bağlı öldüğü ve tüm ölümlerin %31'inin kardiyovasküler hastalıklara bağlı olarak gerçekleştiği tahmin edilmektedir. Kardiyovasküler hastalıklara bağlı olarak gerçekleşen ölümlerin %8'i kalp krizi ve inmeye bağlı olarak meydana gelmiştir. Kardiyovasküler hastalıklara bağlı ölümlerin $\frac{3}{4}$ ' ünden fazlası gelişmemiş ve gelişmekte olan ülkelerde meydana gelmektedir [32].

TÜİK verilerine kalp hastalıklarının neden olduğu ölüm oranları Şekil 3.2'de gösterilmiştir.



Şekil 3. 2. Dolaşım sistemi hastalıklarından ölenlerin oranı [33].

3.2. Veri Madenciliği

Veri madenciliği bilginin bu denli arttığı ve ulaşılmasının kolaylaştığı bir dönemde, veri tabanlarında mevcut bilginin keşfedilmesi olarak tanımlanabilmektedir. Büyük veriler içinden daha önceden bilinmeyen, geçerli ve uygulanabilir bilgilerin geniş veri tabanları taranarak elde edilmesi de denebilir. Bu şekilde elde edilen bilgilerle, bilgiler arasında ilişki kurmak ve geleceğe yönelik tahminlerde bulunmak gibi bilgiye işlevsellik kazandırılabilir.

İlk olarak 1960'lı yıllarda bilgisayarlar kullanılarak veri tarama işlemleri olarak başlanmıştır. Sonrası 1990'lı yıllarda artık veri madenciliği kavramı ortaya çıkmaya

başlamış ve farklı disiplinler olarak, istatistik, makine öğrenmesi, veri tabanları, otomasyon, pazarlama, araştırma bir arada kullanılmaya başlanmıştır[34].

Veri madenciliği bu haliyle, bilim, iş dünyası, arama motorları, web analizi, güvenlik, kanuni uygulamaların takibi birçok alanda kullanılabilmektedir. Çalışılan herhangi bir alanla ilgili sorulabilecek soruların cevabı veri madenciliği aracılığıyla alınmaya çalışılabilmektedir. Örneğin bankacılık sektöründe müşterilerin yönlendirilmesi, kredi tutarları, yatırım araçları vb. gibidir.

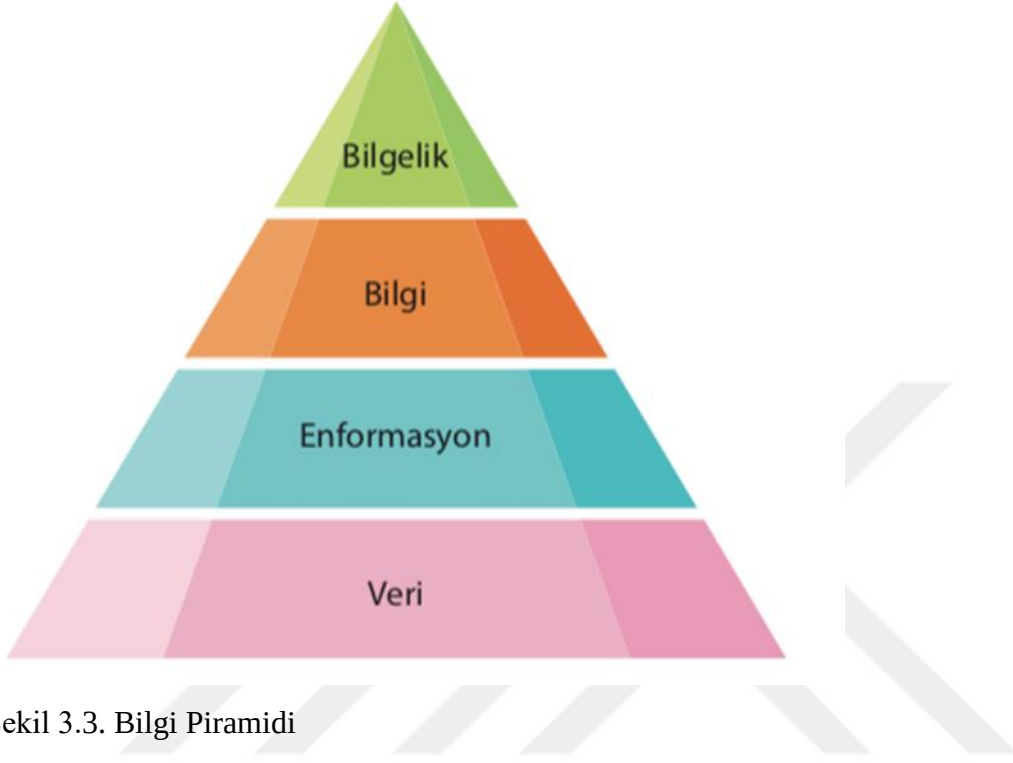
Veri madenciliğinde temel kavramlar, veri (data), enformasyon (information), bilgi (knowledge), bilgelik (wisdom) olarak söylenebilir[35]. Veri madenciliğinde veri kaynakları, veri dosyaları olabileceği gibi, ilişkisel veri tabanları da olabilmektedir. Ayrıca gelişmiş veri kümeleri olarak, duraksız veri (data stream), algılayıcı veri (sensor data), zaman serileri gibi (biyolojik veriler), sosyal ağ verileri, konumsal veriler, nesneye dayalı veri tabanları da veri kaynakları olarak değerlendirilebilmektedir.

Veri madenciliğinde farklı modeller kullanılabilmektedir. Bunlar temel olarak iki ana başlıkta tahmin edici ve tanımlayıcı olarak belirlenebilir. Tahmin edici yöntemlerde, bilinen verilerek kullanılarak bilinmeyen ve cevabı aranan veriler için öngörülerde bulunulabilmektedir. Tanımlayıcı yöntemlerde ise, mevcut verilerde ki ilişkiler kullanılarak karar mekanizmaları için gerekli veriler elde edilebilmektedir.

Bunlardan sınıflandırma, veriyi mevcut sınıflardan birine dahil ederken, zaman serileri incelemede ise, zamanla değişen verilerde verinin istenen bir zamanda ki değerini öngörebilmekteyiz. İstisna analizinde ise genel olarak uyumlu veriler içerisinde uyumsuz olan verileri belirleme için kullanılabilmektedir.

Tüm bu işlemler için öncelikle ham olan işlenmemiş kayıtlar veri olarak kabul edilmektedir. Verinin işlenmesi ve kullanılmaya hazır hale gelmiş haline ise enformasyon olarak isimlendirebilmekteyiz. Enformasyondan sonraki aşama ise bilgi olarak ifade edilmektedir. Bilgi artık kişilerde enformasyondan sonra kendi algılarının da dahil olmasıyla oluşan anlam olarak ifade edilebilmektedir. Son aşama ise bütün bu aşamalardan sonra ulaşılan bilgelik

kavramıdır. Kişinin edindiği farklı bilgiler bir araya gelerek ortaya çıkan kavram bilgeliktir. Tecrübe gibi kişisel özellikler bilgelikten kaynaklı özellikler olarak tanımlanabilmektedir.



Şekil 3.3. Bilgi Piramidi

Veri madenciliği, verilerin toplanması, temizlenmesi, düzenlenmesi, analiz edilmesi ve sonuçların yorumlanması işlemlerini içeren bir süreçtir. Genel olarak veri madenciliği adımları şöyle sıralanabilir:

Veri kaynaklarının belirlenmesi: Veri madenciliği projesinde kullanılacak verilerin nereden ve nasıl elde edileceği belirlenir.

Veri toplama: Belirlenen veri kaynaklarından veriler toplanır.

Veri temizleme: Toplanan veriler, eksik veya yanlış bilgiler içerebilir. Bu nedenle veriler temizlenir ve düzenlenir.

Veri düzenleme: Temizlenen veriler daha sonra düzenlenir ve gerekli formatta hazırlanır.

Veri analizi: Düzenlenen veriler kullanılarak analiz edilir. Bu analizler, verilerin içerisinde gizli kalabilecek bilgileri ortaya çıkarmak için yapılır.

Sonuçların yorumlanması: Analiz edilen verilerin sonuçları yorumlanır ve bu sonuçlar, işletmenin karar alma sürecinde kullanılır.

3.3. Veri Madenciliği Algoritmaları

Veri madenciliği algoritmaları, veri madenciliği sürecinde kullanılan matematiksel veya istatistiksel yöntemlerdir. Veri madenciliği algoritmaları tahmin edici modeller ve tanımlayıcı modeller olmak üzere iki ana kategoriye ayrılır. Bunlar arasında en yaygın olarak kullanılanlar aşağıda açıklanmıştır.

3.3.1. Naive Bayes

Naive Bayes, veri madenciliği algoritmalarından biridir ve genellikle sınıflandırma problemlerinde kullanılmaktadır. Bu algoritma, Bayes Teoremi'ni kullanarak verilerin sınıflandırılmasını gerçekleştirmektedir. Bayes Teoremi, olasılıklar arasındaki ilişkiyi tanımlar ve verilerin geçmişte gözlenen olasılıklarını kullanarak gelecekteki olayların olasılığını tahmin etmeye çalışır.

Naive Bayes, verilerin her bir özelliklerinin sınıflandırma için kullanılmasını varsayar ve bu özellikler arasında hiçbir ilişki olmadığını farz eder. Bu nedenle, "naive" olarak adlandırılır. Örneğin, bir e-posta sınıflandırma uygulamasında, e-postanın içeriği, göndericinin adresi ve e-postanın başlığı gibi özellikler kullanılabilir. Naive Bayes, bu özellikler arasında

hiçbir ilişki olmadığını varsayarak, e-postanın spam veya ham olduğunu tahmin etmeye çalışır.

Naive Bayes, özellikle çok sayıda özellik içeren veriler için etkili bir seçim olabilir ve çok hızlı bir şekilde çalışır. Ancak, veriler arasındaki ilişkileri gözlemlemediğinden, sınıflandırma doğruluğu diğer algoritmalara göre daha düşük olabilir.

Bayes Teoremi, olasılıklar arasındaki ilişkiyi tanımlayan bir formüldür ve aşağıdaki şekilde ifade edilir[36].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bu formülde:

$P(A|B)$: B risk faktörü kombinasyonu altında kişinin A hastalığına sahip olma olasılığı (örneğin, kişinin kalp hastalığına sahip olma olasılığı).

$P(B|A)$: A hastalığına sahip bir kişide B risk faktörü kombinasyonunun olasılığı (örneğin, kalp hastalığına sahip bir kişide 50 yaşında erkek, sigara içici ve diyabetli olma olasılığı).

$P(A)$: Kişinin A hastalığına sahip olma olasılığı (örneğin, kişinin kalp hastalığına sahip olma olasılığı).

$P(B)$: B risk faktörü kombinasyonunun gerçekleşme olasılığı (örneğin, 50 yaşında erkek, sigara içici ve diyabetli olma olasılığı).

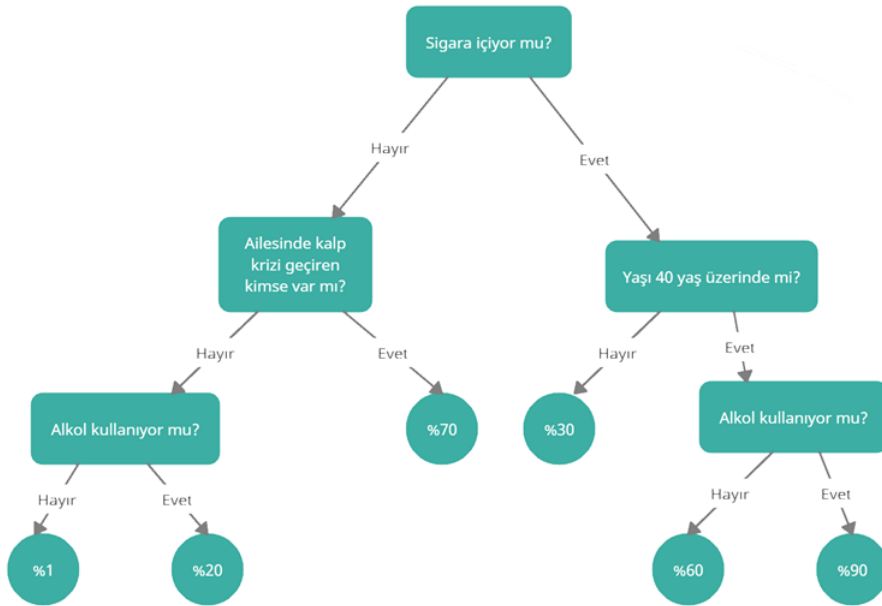
Naive Bayes algoritmasının "naive" olarak adlandırılmasının sebebi, $P(B|A)$ terimi için tüm risk faktörlerinin bağımsız olduğu varsayımıdır. Bu durumda $P(B|A) = P(\text{Age}|A) * P(\text{Gender}|A) * P(\text{Smoking}|A) * P(\text{Diabetes}|A)$.

Bu formülü kullanarak, veri setinde önceden belirlenmiş risk faktörleri verileri ile eğitilen model, verilen bir kişinin risk faktörlerine göre kalp hastalığına sahip olma olasılığını hesaplayabilmektedir.

3.3.2 Karar Ağacı

Karar ağacı algoritması, veri madenciliğinde kullanılan bir sınıflandırma veya regresyon yöntemidir. Algoritma, veri kümesinden özellikleri ve etiketleri kullanarak bir karar ağacı oluşturur. Bu ağaç, veri kümesindeki nesnelerin sınıflandırılması için kullanılan kuralları veya kararları içerir. Her düğüm, bir özellik ve bu özellik için bir değer aralığı temsil eder ve düğümün alt dalları, bu değer aralığındaki değerler için sınıflandırma kurallarını içerir. Yaprak düğümleri ise, nesnenin sınıfını temsil eder. Karar ağacı algoritması, veri kümesinde önceden bilinmeyen bir sınıflandırma yapmak için kullanılabilir ve genellikle kolay anlaşılır ve uygulanabilir sonuçlar verir.

Şekil 3.4.'te yer alan örnek karar ağacında iç düğümlerde (dikdörtgenler) çeşitli girdilere göre ağacın dallanması söz konusudur. Yapraklarda ise sonuç olarak elde edilen değerler gösterilmiştir.



Şekil 3. 4. Örnek karar ağacı[37].

Karar ağacı algoritması genellikle aşağıdaki adımları içerir:

- Veri kümesinin özellikleri ve etiketleri kullanılarak bir karar ağacı oluşturulur.
- Ağaç, veri kümesindeki nesnelerin sınıflandırılması için kullanılan kuralları veya kararları içerir.
- Her düğüm, bir özellik ve bu özellik için bir değer aralığı temsil eder.
- Düğümün alt dalları, bu değer aralığındaki değerler için sınıflandırma kurallarını içerir.
- Yaprak düğümleri ise, nesnenin sınıfını temsil eder.
- Algoritma, veri kümesinde önceden bilinmeyen bir sınıflandırma yapmak için kullanılabilir ve genellikle kolay anlaşılır ve uygulanabilir sonuçlar verir[38].

Karar ağaçlarının oluşturulmasında en yaygın kullanılan denklem, sınıflandırma veya regresyon problemi için en iyi bölünmeyi (split) bulmak için kullanılan Gini indeksi veya Entropi denklemidir. Bu denklemler, veri setindeki nesnelerin sınıfları arasındaki eşitliği ölçer. Gini indeksi, sınıflandırma problemlerinde kullanılan bir ölçüt olarak, sınıflar arasındaki eşitsizliği ölçer. 0 değeri eşit bir dağılımı, 1 değeri ise sınıflar arasında büyük bir eşitsizliği ifade eder. Gini indeksi, her bölünme için hesaplanır ve en küçük Gini indeksine sahip bölünme seçilir. Bu şekilde, bölünmeler arasında en düşük eşitsizliği sağlamak amaçlanır.

Entropi ise, veri setindeki nesnelerin sınıfları arasındaki eşitliği ölçer. Entropi 0 ile 1 arasında bir değer alır ve 0 değeri eşit bir dağılımı, 1 değeri ise sınıflar arasında büyük bir eşitsizliği ifade eder. Her bölünme için entropi hesaplanır ve en küçük entropiye sahip bölünme seçilir. Bu şekilde, bölünmeler arasında en düşük eşitsizliği sağlamak amaçlanır.

Karar ağacı algoritmasının amacı, veri setindeki özellikler arasındaki ilişkiyi belirlemek ve veri setindeki nesnelerin sınıflandırılması için kullanılacak kuralları veya kararları içeren bir karar ağacı oluşturmaktır. Bu amaç, Gini indeksi veya Entropi denklemleri ile

gerçekleştirilir ve algoritma, veri setinde önceden bilinmeyen bir sınıflandırma yapmak için kullanılabilir ve genellikle kolay anlaşılır ve uygulanabilir sonuçlar verir[39].

Karar ağacı algoritmalarının çeşitleri bulunmaktadır. Bu algoritmalar içerisinde en çok kullanılanlar; ID3, Random forest, CART, C4.5'tir.

3.3.3. K En-Yakın Komşu Algoritması

K en Yakın Komşu Algoritması (K-NN), 1950'lerin sonlarında bulunmuştur. Bu algoritma, verilen bir örnekteki özellikleri kullanarak benzer örnekleri bulmaya ve bu benzer örneklerin etiketlerini kullanarak örneğin etiketini tahmin etmeye çalışır.

Algoritmanın çalışma şekli aşağıda yer almaktadır:

1. Veri noktalarının benzerliği ölçüsü: Bu benzerlik ölçüsü, veri noktaları arasındaki mesafeyi ölçen bir fonksiyondur, örneğin Euclidean mesafesi.
2. En yakın k veri noktasının belirlenmesi: Test verisi, tüm veri noktaları arasındaki benzerlik ölçüsüne göre sıralanır ve en yakın k veri noktası seçilir.
3. Test verisinin sınıflandırması: En yakın k veri noktasının sınıfının modu (en sık tekrarlayan sınıf) test verisi için sınıflandırma olarak atanır.

Bu algoritmanın en önemli avantajı, hızlı ve basit olmasıdır. Ancak, veri kümesi çok büyükse performansı düşebilir ve veri kümesindeki gürültü etkileyebilir. Ayrıca, algoritma özellikler arasındaki ilişkileri dikkate almaz ve sadece benzerliği esas alır[40].

3.3.4. Lojistik Regresyon Algoritması

Lojistik Regresyon Algoritması, sınıflandırma problemlerinde kullanılan bir veri madenciliği yöntemidir. Bu algoritma, bir nesnenin belirli bir sınıfa ait olup olmayacağını tahmin etmek için kullanılır.

Lojistik regresyon, çıktı olarak bir olasılık değeri veren bir sınıflandırma modelidir. Model, veri setindeki nesne özelliklerine dayanarak belirli bir sınıfa ait olma olasılığını hesaplar ve olasılık değerine göre bir sınıflandırma etiketi atar.

Lojistik regresyon, diğer sınıflandırma yöntemlerine göre daha iyi performans gösterir ve basit matematiksel formüllerle uygulanabilir[41].

3.3.5. Destek Vektör Makinesi

Destek Vektör Makinesi (SVM) Algoritması, veri madenciliği ve sınıflandırma problemlerinde kullanılan bir yöntemdir. SVM, veri setindeki nesneleri en iyi ayırım çizgisi olan bir hiperdüzleme dayalı olarak sınıflandırır. Bu ayırım çizgisi, veri setindeki nesneleri en iyi şekilde ayıran ve en fazla boşluk bırakan çizgidir[42]. SVM, veri setindeki nesneleri sınıflandırma için kullanılan bir destek vektörü temel alır. Destek vektörü, veri setindeki en önemli nesnelerdir ve sınıflandırma için en önemli referans noktalarıdır. SVM, veri setinin boyutu ve veri dağılımına göre yüksek performans gösterebilir ve önceki tecrübelerle iyileştirilebilir. Örnek olarak, görsel tanıma, sınıflandırma ve metin sınıflandırma gibi uygulamalarda kullanılabilmektedir.

3.3.5. Yapay Sinir Ağları

Yapay Sinir Ağları (Artificial Neural Networks, ANN), bir bilgisayar sistemi tarafından öğrenme ve öğrenmeye benzer bir davranış sergilemeye yönelik tasarlanmıştır. ANN, gerçek sinir sistemi temel alınarak tasarlanır ve daha önce görülmüş veriler kullanılarak öğrenir[43]. ANN, giriş verilerine göre ağırlıkları ve bias değerlerini değiştirerek çıkış verilerini üretebilir. Formül olarak, her bir ağ nöronu için giriş verileri ile ağırlıkların çarpımı yapılır ve bias değeri eklenir, sonra bu değer aktivasyon fonksiyonu ile işlenir ve çıkış olarak elde edilir. Örnek olarak, görsel tanıma görevinde, ANN girdi olarak görüntü verilerini alabilir

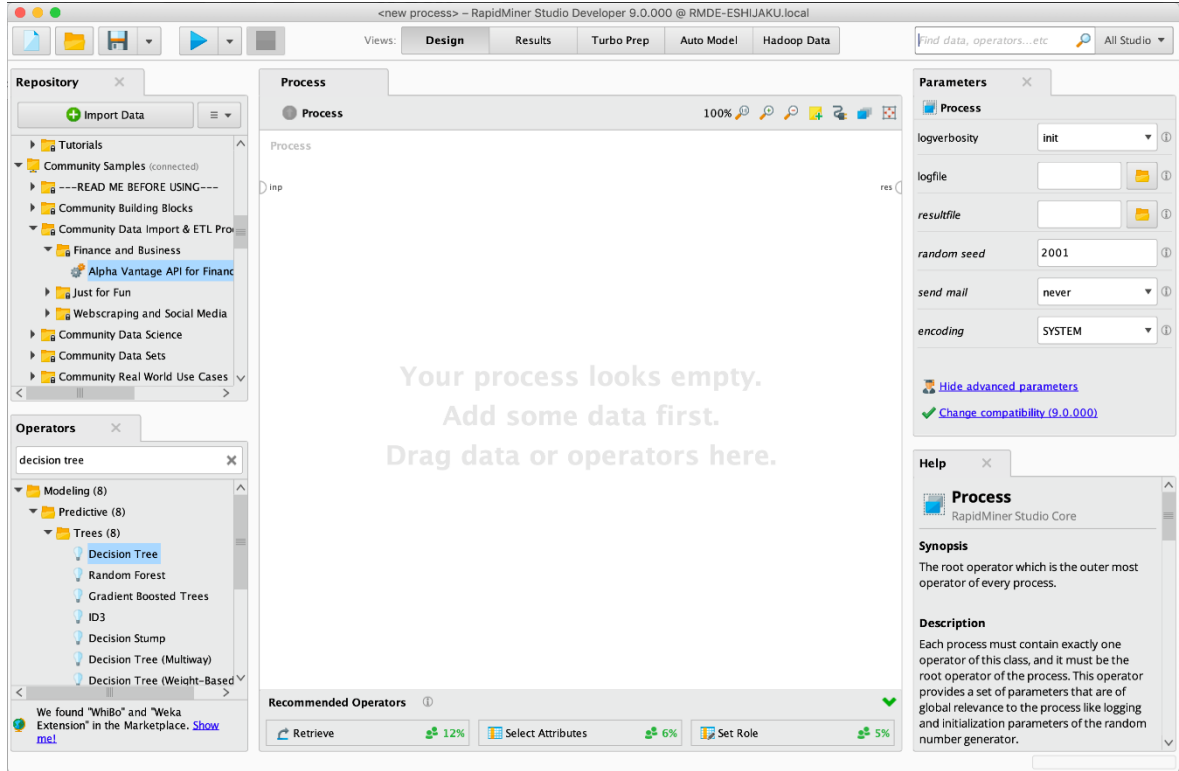
ve çıktı olarak nesnenin ne olduğunu tahmin edebilir. Ayrıca ses tanıma, metin sınıflandırma ve makine çevirisi gibi uygulamalarda da kullanılabilir.

3.4. Veri Madenciliğinde Kullanılan Programlar

Veri madenciliğinde açık kaynaklı, ticari gibi farklı seviyelerde birçok uygulama bulunmaktadır. Bu uygulamaların seçimi yapılacak çalışmanın niteliği, kullanıcının bilgi seviyesi gibi parametreler doğrultusunda yapılmalıdır. Veri madenciliğinde kullanılan başlıca programlar aşağıda yer almaktadır.

3.4.1. RapidMiner Studio

2001 yılında Dortmund Üniversitesi'nde java dili ile geliştirilen açık kaynak kodlu program daha önce Yale adıyla bilinirken 2007 yılında RapidMiner olarak değiştirilmiştir. Kullanışlı arayüzü sayesinde kullanıcılara kodlama yapmaya gerek kalmadan veri işleme imkanı sunar. Veri ön işleme ve görselleştirme, tahmine dayalı analitik ve istatistiksel modelleme, veri yükleme ve dönüştürme(ETL) gibi veri madenciliği süreçlerini sunmaktadır. Rapidminer; .aml, arff, att, bib, clm, cms, cri, csv, dat, ioc, log, mat, mod, obf, par, per, res, sim, thr, wgt, wls, xrf ve a gibi birçok dosya formatıyla kullanılabilir[44].

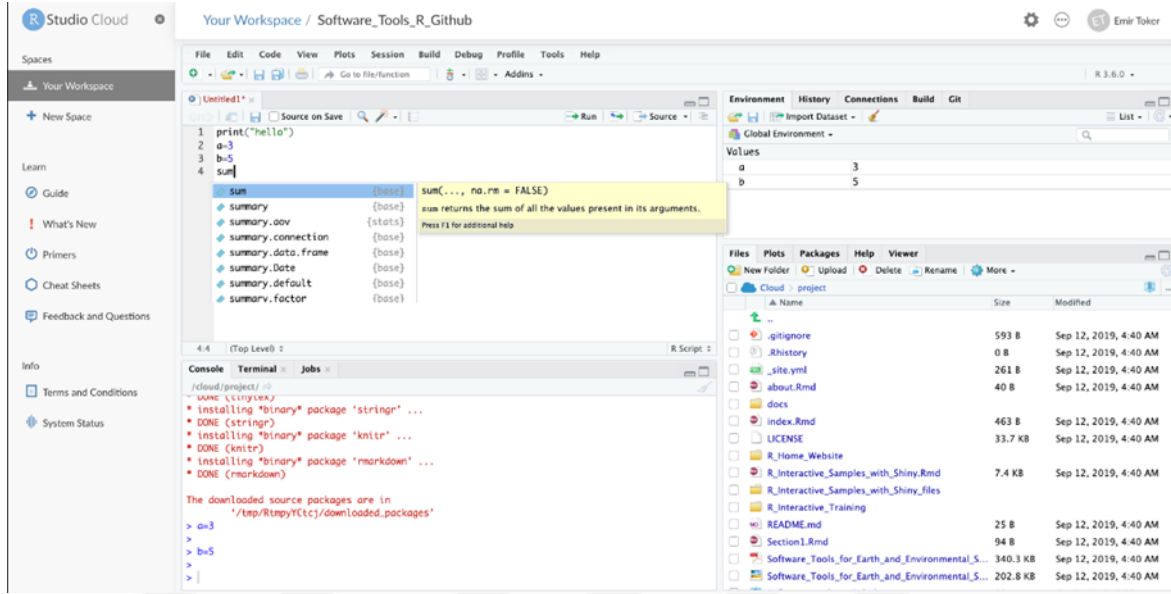


Şekil 3. 5. RapidMiner Ekran Görünümü[45].

3.4.2. R

R, S diline benzer bir programlama dilidir ve istatistiksel hesaplama ve grafikler için oluşturulmuş bir yazılımdır. 1992 yılında Auckland Üniversitesi'ndeki Ross Ihaka ve Robert Gentleman adlı istatistikçiler tarafından geliştirilmiştir[46].

R programlama dili, istatistik ve veri analitiği için tasarlanmış bir açık kaynak programlama dilidir. R, veri kümelerini inceleme, görselleştirme ve modelleme için geniş bir kütüphane yelpazesine sahiptir. R, özellikle veri bilimciler, istatistikçiler ve araştırmacılar tarafından sıklıkla kullanılmaktadır. R, diğer programlama dilleri ile birlikte kullanılabilir ve veri kaynaklarından verileri okuyarak, veri setlerini temizleyerek ve sonuçları raporlama aracı olarak kullanılabilir. R ayrıca, çeşitli grafikler ve diğer görsel araçlar ile veri görselleştirme işlemleri için de kullanılabilir.



Şekil 3. 6. R programına ilişkin ekran görüntüsü [47].

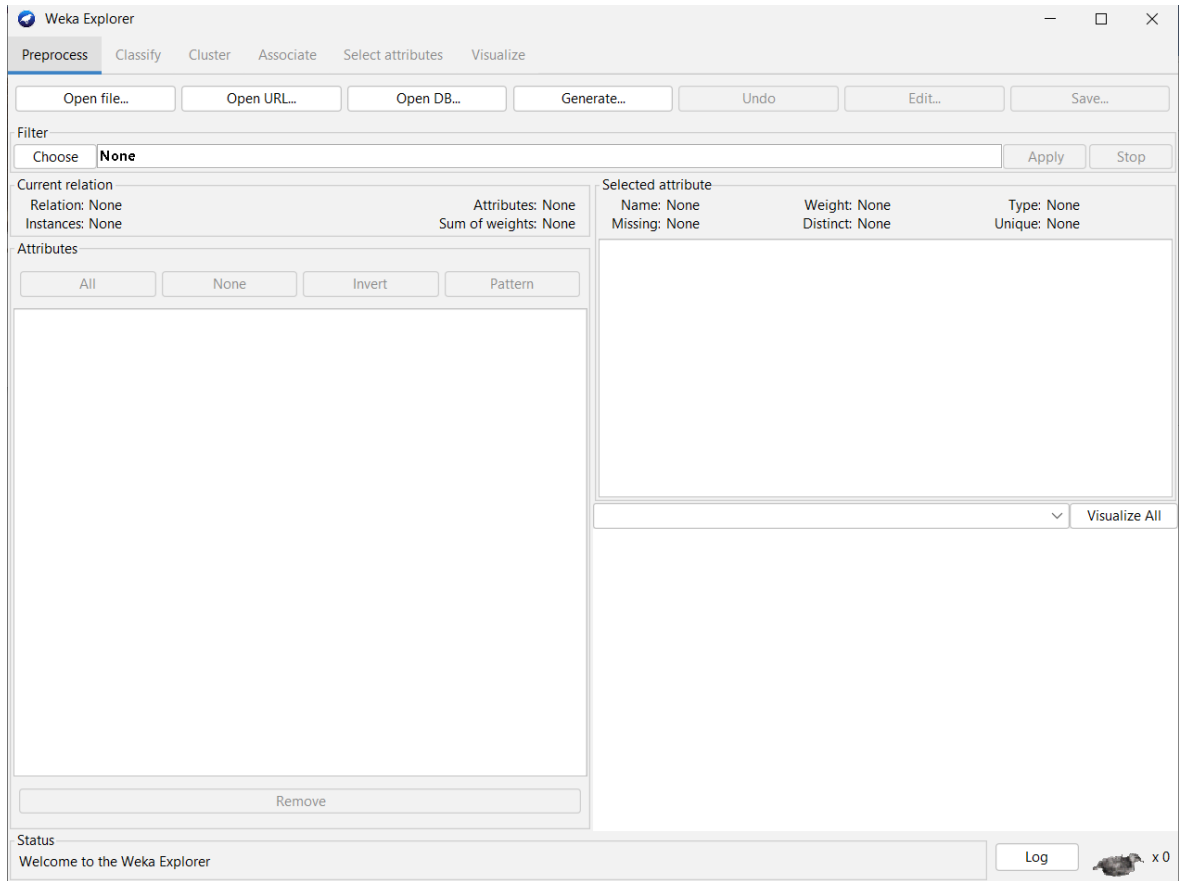
3.4.3. Spss

SPSS (Statistical Package for the Social Sciences) sosyal bilimlerde istatistiksel analiz için geniş ölçekte kullanılan IBM'e ait bir yazılım paketidir. Veri işleme, tanımlayıcı istatistikler, yargılayıcı istatistikler ve grafikler gibi çeşitli özellikleri içerir. Kullanıcı dostu bir arayüze sahip olan SPSS, veri girişi, temizleme ve analiz işlemlerini kolaylaştırır. Ayrıca, SPSS, çeşitli veri türlerini ve formatlarını destekler ve sonuçları raporlama ve sunma işlemleri için de kullanılabilir. SPSS, sosyal bilimler alanındaki araştırmacılar, istatistikçiler ve veri bilimcileri tarafından sıklıkla kullanılmaktadır[48].

3.4.4. Weka

Weka (Waikato Environment for Knowledge Analysis) ilk olarak 1990 yılında Waikato Üniversitesi'nde (Yeni Zelanda) Ian H. Witten, Eibe Frank ve Mark A. Hall tarafından geliştirilmiştir. İlk sürümü 1993 yılında yayınlanmıştır[49].

Weka, bir açık kaynaklı veri madenciliği ve öğrenme ortamıdır. Java programlama diliyle yazılmıştır ve kullanıcılarına veri madenciliği, öğrenme algoritmaları, görselleştirme ve raporlama gibi özellikleri sunar. Weka, veri setlerinin yüklenmesi, temizlenmesi, özetlenmesi ve analiz edilmesi için araçlar sunar. Ayrıca, kullanıcılar Weka aracılığıyla veri setlerinde çeşitli öğrenme algoritmalarını uygulayabilir ve sonuçlarını karşılaştırabilirler. Arff, csv formatında bulunan dosyalarla işlem yapılabilir.



Şekil 3. 7. Weka programına ilişkin ekran görüntüsü.

3.4.5. Python

Python, veri madenciliği alanında kullanılan popüler bir programlama dilidir. Python'un geniş bir kütüphaneye sahip olması ve bu kütüphanelerin veri madenciliği için gerekli olan tüm araçları sağlayabilmesi en önemli özelliklerinden biridir. Örneğin, Pandas kütüphanesi veri setlerini okuma, düzenleme ve özetleme için kullanılırken, scikit-learn kütüphanesi ise

çeşitli öğrenme algoritmalarını uygulamak için kullanılmaktadır. Ayrıca, Python'un okunaklı ve anlaşılır bir sözdiziminin oluşu veri madenciliği için kodlama yaparken kolaylık sağlamaktadır.

Python ayrıca veri görselleştirme için de kullanılabilir. Matplotlib ve Seaborn gibi kütüphaneler veri setlerini görselleştirmek için kullanılabilir. Bu sayede, veri madenciliği sürecinde elde edilen sonuçların daha iyi anlaşılması ve sunulması mümkün olur[50].

Son olarak, Python'un açık kaynak kodlu olması ve geniş bir kullanıcı topluluğu tarafından desteklenmesi nedeniyle, veri madenciliği için Python kullanmak için çok sayıda örnek, öğretici ve doküman mevcuttur. Bu sayede, Python ile veri madenciliği öğrenmek daha kolay ve hızlı hale gelir.

3.4.6. Matlab

MATLAB (Matrix Laboratory) ilk olarak 1984 yılında Cleve Moler tarafından yazılmıştır[51]. Moler, bir matematik profesörü olarak, öğrencilerine matematik derslerinde matrisleri kullanmalarını sağlamak için bu programı yazmıştır. MATLAB, ilk olarak sadece bir matris işleme aracı olarak kullanılıyordu ancak zaman içinde çok daha fazla özellik ve uygulama alanı kazanmıştır.

MATLAB, veri madenciliği alanında kullanılan bir programlama dili ve ortamıdır. MATLAB, matematiksel işlemler, öğrenme algoritmaları, görselleştirme ve raporlama gibi özelliklerin yanı sıra; veri setlerinin yüklenmesi, temizlenmesi, özetlenmesi ve analiz edilmesi için de araçlar sunar. Ayrıca, kullanıcılar MATLAB aracılığıyla veri setlerinde çeşitli öğrenme algoritmalarını uygulayabilir ve sonuçlarını karşılaştırabilirler. Böylece kullanıcılar veri setlerinden öğrenme modeli oluşturabilir ve sonuçlarını test edebilirler. MATLAB'ın veri madenciliği alanında kullanmasının avantajlarından biri de MATLAB 'ın

kendi içinde üst düzey grafiksel ara yüzünün olmasıdır. Bu sayede, veri setleri ve sonuçları kolayca görselleştirilebilir.

3.4.7. Tanagra

Ricco Rakotomalala tarafından Fransa Lumière Üniversitesi'nde geliştirilen araştırma ve akademik amaçlı ücretsiz bir yazılım paketidir. Veri madenciliği işlemleri için kullanılabilir ve kullanıcılarına çeşitli veri analitik işlemleri gerçekleştirmelerine olanak tanır. Örneğin Tanagra, kullanıcılarına sınıflandırma, kümeleme, regresyon, lojistik regresyon ve diğer çeşitli veri analitik yöntemlerini uygulamalarına olanak tanır. Ayrıca, Tanagra, veri setlerini ön işleme, temizleme ve görselleştirme gibi işlemleri gerçekleştirmek için de kullanılabilir.

Tanagra, veri madenciliği projelerinde kullanıcıların ihtiyacı olan veri analitik işlemlerini gerçekleştirmelerine olanak tanır. Python ile yazılmıştır, bu nedenle kullanıcılarına Python kodu kullanarak veri analitik işlemlerini özelleştirme ve genişletme olanağı sunar. Bu özellik sayesinde Tanagra, veri madenciliği projelerinde kullanıcıların ihtiyacı olan veri analitik işlemleri gerçekleştirmelerine olanak tanır[52].

3.4.8. Orange

Orange, bir açık kaynak kodlu veri analitik ve görselleştirme platformudur. İlk olarak 2003 yılında Ljubljana Üniversitesi (Slovenya) Bilgisayar Fakültesi ve Bilgi Bilimi tarafından geliştirildi. Orange kullanıcı dostu güçlü ve esnek görsel programlama, arama amaçlı veri analizi ve görüntüleme ve Python bağlama ve kodlama için kütüphaneler içeren bileşen tabanlı bir veri madenciliği ve makine öğrenmesi yazılım takımıdır. Veri ön işleme, özellik skorlama ve filtreleme, modelleme, model değerlendirme ve keşif teknikleri gibi geniş kapsamlı bileşen seti içerir. C++ (hız) ve Python (esneklik) 'a uygulanmıştır. Grafik kullanıcı

arayüzü çapraz platform üzerine inşa eder. Orange GPL (Genel Kamu Lisansı) altında ücretsiz olarak dağıtılmaktadır(Tekerek, 2011: 166).

3.4.9. Knime

KNIME (Konstanz Information Miner), açık kaynak kodlu bir veri analitik, raporlama ve entegrasyon platformudur. Kullanıcıların görsel olarak veri analitik işlemleri yapmasını sağlamak için tasarlanmıştır. KNIME, veri madenciliği için geniş bir araç yelpazesine sahiptir. Kullanıcılar, veri kaynaklarını toplamak, veri temizliği ve hazırlık, modeller oluşturma ve sonuçları raporlama gibi işlemleri gerçekleştirebilir. KNIME, ayrıca, veri madenciliği için popüler açık kaynak kodlu paketleri entegre edebilmektedir. Örneğin R ve Python gibi diller entegre edilebilmektedir. Bu özellikleri ile KNIME, veri madenciliği için kullanılabilir bir platformdur[54].

3.4.10. SAS

SAS (Statistical Analysis System) veri yönetimi, gelişmiş analitik ve iş zekası için yaygın olarak kullanılan bir yazılım paketidir. İçinde veri girişi, depolama, raporlama ve analitik araçlar gibi birçok araç bulunur. SAS veri madenciliği için de birçok araç sunar. Öncelikle, SAS veri setlerini kolayca yükleyebilir ve manipüle edebilir. Bu, veri madenciliği işlemleri için gerekli olan veri temizliği ve hazırlık adımlarını kolaylaştırır. SAS, ayrıca, veri madenciliği teknikleri için çeşitli modeller ve algoritmalar sunar. Örneğin, kurumsal analitik paketi, lojistik regresyon, karar ağaçları, k-means, clustering gibi klasik veri madenciliği yöntemlerini içerir. Ayrıca, SAS Visual Data Mining and Machine Learning isimli uygulaması, kullanıcıların görselleştirme ve keşif analitikleri kullanarak veri madenciliği

işlemleri gerçekleştirmelerine olanak tanımaktadır. Bu araçlar, SAS veri madenciliği için kullanılabilir[55].

3.5. Veri Setleri

Yapılan araştırmalarda kullanılacak veri setleri hastanede bulunan hasta kayıtlarının anonimleştirilmesiyle, hastaların rızasının alınarak veri toplanması şeklinde yapılabilmektedir. Bu amaçla hastalardan veri bilgisinin alınması uzun ve zahmetli olduğundan genellikle mevcutta bulunan veri setleri tercih edilmektedir. En çok tercih edilen veri setlerinin bulunduğu alanlar;

- UCI Machine Learning Repository
- Kaggle Datasets
- Amazon Datasets
- Google's Datasets Search Engine

şeklinde. Bu veriler açık veri kayıtları olarak yer almaktadır. Farklı türden veri setlerinin bulunduğu bu alanlarda birçok araştırmaya konu olacak veri setleri yer almaktadır.

Kalp hastalığı ile ilgili verilere daha çok UCI Machine Learning Repository ve Kaggle'da rastlanmaktadır. UCI Machine Learning Repository'de kalp hastalığına ilişki 4 ayrı veri tabanı bulunmaktadır. Bunlar; Cleveland, Hungary, Switzerland ve VA Long Beach veri tabanlarıdır.

Kalp hastalığı tahmin etme ile ilgili yapılan araştırmalarda genellikle UCI Machine Learning Repository'de bulunan Cleveland veri tabanı tercih edilmiştir. Bu veri tabanında yer alan veri seti 76 özellik ve 303 hasta kaydından oluşmaktadır[56]. Araştırmacılar bu özelliklerden önemli olanları birçok yöntemlerle belirlemektedirler. Ancak bununla birlikte yayınlanan deneylerin çoğunda 14 özelliğin kullanılması yönünde atıfta bulunmaktadır. Bu özelliklerin bulunduğu veri seti örneği Çizelge 4.1'de gösterilmiştir.

Çizelge 3.1. Veri seti örneği

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
58	0	3	150	283	1	0	162	0	1	2	0	2	1
50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
58	0	2	120	340	0	1	172	0	0	2	0	2	1
66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
42	1	0	140	226	0	1	178	0	0	2	0	2	1
61	1	2	150	243	1	1	137	1	1	1	0	2	1
40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
53	1	2	130	197	1	0	152	0	1.2	0	0	2	1



4. YÖNTEM

Tez çalışmasında Android Studio Dolphin 2021.3.1 programı üzerinden Android tabanlı bir uygulama geliştirilmiştir. Mobil cihazlarda çalışacak olan bu uygulama ile 13 özniteliğe sahip ve 303 hasta kaydı bulunan veri setine veri madenciliği algoritmalarından Naive Bayes, KNN ve lojistik regresyon algoritmaları seçilerek doğruluk oranları tespit edilmiştir.

4.1. Veriler ve Toplanması

UCI Machine Learning Repository'den alınan ve Kaggle web sitesinde bulunan "Cleveland Heart Disease Dataset" adlı veri seti, 1988 yılında yapılan bir çalışmadan oluşturulmuştur. Tez çalışmasında bu veri seti kullanılmıştır. Veri seti 303 kayıt ve 13 öznitelikten oluşmaktadır. Veri setine ilişkin özellikler ve açıklamalar Tablo 5.1'de verilmiştir[57].

Çizelge 4.1. Öznitelik tablosu

Sıra No	Öznitelik	Açıklama
1	Yaş (age)	Nümerik
2	Cinsiyet (sex)	İkili Sayı
3	Göğüs Ağrısı Tipi (cp)	Nominal
4	İstihrahat Basıncı (trestbps)	Nümerik
5	Serum Kolesterol (chol)	Nümerik
6	Açlık Kan Şekeri (fbs)	İkili Sayı
7	İstihrahat Elektrokardiyografik Sonuç (restecg)	Nominal
8	Maksimum Kalp Hızı (thalach)	Nümerik
9	Egzersize Bağlı Anjin (exang)	İkili Sayı
10	Egzersize Bağlı ST Depresyonu (oldpeak)	Nümerik
11	ST Segmentinin Eğimi (slope)	Nominal
12	Floroskopi ile Renklendirilmiş Damar Sayısı (ca)	Nominal
13	Defekt Tipi (thal)	Nominal
14	Num	İkili Sayı

Veri setinde yer alan num değişkeni tahmin değişkenidir. Bu değer '0' ise hasta, '1' ise hasta değil anlamına gelmektedir. Ayrıca veri setinde yer alan eksik değerler içeren alanlar çıkarılmıştır. Bu durumda 294 hasta kaydı elde edilmiştir.

4.2. Model Performans Ölçütleri

Veri madenciliği model performans ölçütleri, bir modelin başarısının nasıl ölçüldüğünü açıklar. Bunlar arasında en yaygın olarak kullanılanlar aşağıda açıklanmıştır.

Doğruluk (Accuracy)

Modelin, tüm veri kümesinde doğru tahmin yaptığı oranıdır.

$$\text{Doğruluk} = \frac{TP + TN}{TP + FP + FN + TN}$$

Kesinlik (Precision)

Modelin, pozitif olarak sınıflandırdığı verilerin gerçekten pozitif olduğu oranıdır.

$$\text{Kesinlik} = \frac{TP}{TP + FP}$$

Duyarlılık (Recall)

Modelin, gerçekten pozitif olan verileri pozitif olarak sınıflandırdığı oranıdır.

$$\text{Duyarlılık} = \frac{TP}{TP + FN}$$

F Ölçütü

Kesinlik(K) ve duyarlılığın(D) harmonik oranıdır[58].

$$F = \frac{2DK}{D + K}$$

Kappa İstatistiği

Etiketleme veya sınıflandırma problemlerinde hakemler arasındaki uyumun ölçülmesinde kullanılmaktadır. Örneğin, bir veri kümesi içindeki örneklerin belirli bir kategoride sınıflandırılması gerektiğinde, birden fazla hakemin verileri etiketlemesi ve sonrasında kappa istatistiği kullanılarak hakemler arasındaki uyumun ölçülmesi mümkündür. Kappa, hakemler arasındaki anlaşma düzeyini 0 ile 1 arasında bir skorla ölçer ve 1 tam anlaşma, 0 ise tam uyuşmazlık anlamına gelir.

ROC Eğrisi

Modelin pozitif olarak sınıflandırdığı verilerin gerçekten pozitif olma olasılığını ve gerçekten negatif olan verilerin yanlış pozitif olarak pozitif olarak sınıflandırma olasılığını gösterir.

AUC (Area Under the ROC Curve)

ROC eğrisinin altındaki alandır.

RMSE (Root Mean Squared Error)

Tahmin edilen değerler ile gerçek değerler arasındaki hata ortalamasının kareköküdür.

MAE (Mean Absolute Error)

Tahmin edilen değerler ile gerçek değerler arasındaki hata ortalamasıdır. Her veri kümesi için en uygun performans ölçütü farklı olabilir ve modelin tasarımına göre değişebilir.

Confusion Matrix (Karışıklık Matrisi)

Sınıflandırma problemlerinde bir modelin doğru ve yanlış tahminlerini görsel olarak raporlamak için kullanılan bir tablo şeklidir. Aşağıda yer alan değerler bulunmaktadır.

True Positives (TP)

Modelin, bir verinin pozitif olduğunu doğru bir şekilde tahmin etmesidir.

False Positives (FP)

Modelin, bir verinin negatif olduğunu pozitif olarak tahmin etmesidir.

True Negatives (TN)

Modelin, bir verinin negatif olduğunu doğru bir şekilde tahmin etmesidir.

False Negatives (FN)

Modelin, bir verinin pozitif olduğunu negatif olarak tahmin etmesidir.

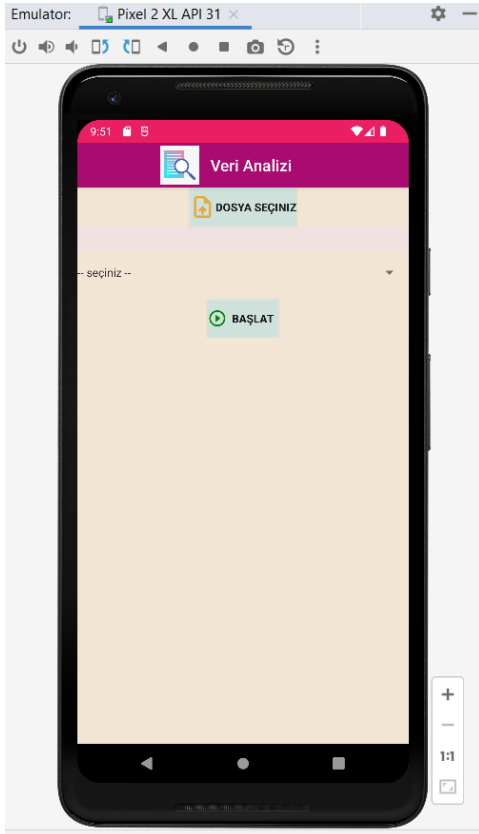
Karışıklık matrisi şu şekildedir:

	Tahmin Edilen Sınıf		
	Sınıf	Pozitif	Negatif
Gerçek Sınıf	Pozitif	DP	YN
	Negatif	YP	DN

Bu değerler, bir modelin performansının ölçülmesinde kullanılabilir ve doğruluk, kesinlik, duyarlılık, F ölçütü gibi metrikler hesaplanabilir.

5. UYGULAMA

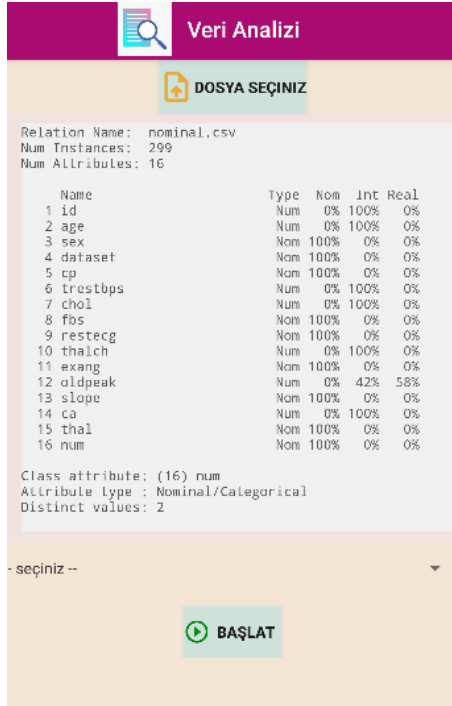
Bu tez çalışmasında veri setleri ile veri madenciliği algoritmalarının karşılaştırılması amacıyla mobil cihazlarda kullanılabilecek bir uygulama geliştirilmiştir. Uygulama Android işletim sisteme sahip cihazlar üzerinde çalışacak şekilde tasarlanmış ve programlama dili olarak JAVA dili seçilmiştir.



Şekil 5.1. Veri analizi dosya seçme ara yüz görüntüleri

Bu uygulama ile kullanıcılar veri setini Şekil 5.1’de yer alan ara yüz ekranında görüldüğü gibi “Dosya Seç” butonu ile mobil cihazdan dosya seçimi yapılır ve Şekil 5.2’de yer alan ekranda yer aldığı gibi seçilen veri setine ilişkin bilgiler kullanıcıya gösterilir. Aynı ekranda yer alan kısımdan uygulanmak istenen algoritma seçilerek “Başlat” tuşuyla veri setine ilişkin

değerler hesaplanır. Bu hesaplama işlemlerinde Weka kütüphanesinden yararlanılmıştır. Yapılan hesaplamalarda cross validation folds değeri default olarak 10 seçilmiştir.



Şekil 5.2. Veri seti bilgisinin ara yüz görünümü

Hesaplamalara ilişkin bilgiler 5.3'te yer alan ekranda yer almaktadır. Ayrıca detay bilgileri farklı ekranda gösterilmektedir. Detay bilgilerinde ilgili algoritmaya ait karışıklık matrisi, kesinlik, duyarlılık, F ölçütü, ROC eğrisi hesaplamalarına yer verilmiştir.



Şekil 5.3. Hesaplanan değerlerin ara yüzde gösterilmesi

Seilen veri setinin uygulama zerinde alıřtırılmasıyla elde edilen bulgular ařağıdaki gibidir.

izelge 5.1. Algoritma Performansları

Algoritma	Doğruluk Oranı
KNN	%77.59
Naive Bayes	%82.60
Lojistik Regresyon	%82.94

izelge 5.2. Karışıklık Matrisi

	a (kalp hastalığı yok)	b (kalp hastalığı var)
a (kalp hastalığı yok)	TP	FN
b (kalp hastalığı var)	FP	TN

TP (True Positive – Doğru Pozitif): Gerçekte doğruyken doğru olarak sınıflandırılan kayıt sayısı

FN (False Negative – Yanlış Negatif): Gerçekte doğruyken yanlış olarak sınıflandırılan kayıt sayısı

FP (False Positive – Yanlış Pozitif): Gerçekte yanlışken doğru olarak sınıflandırılan kayıt sayısı

TN (True Negative – Doğru Negatif): Gerçekte yanlışken yanlış olarak sınıflandırılan kayıt sayısı

izelge 5.3. KNN Algoritması Karışıklık Matrisi

	a	b
a	128	32
b	35	104

izelge 5.4. Naive Bayes Algoritması Karışıklık Matrisi

	a	b
a	137	23
b	29	110

izelge 5.5. Lojistik Regresyon Algoritması Karışıklık Matrisi

	a	b
a	138	22
b	29	110



6. SONUÇ

Bu tez çalışmasında, java ile geliştirilen mobil bir uygulama aracılığıyla kalp hastalığı tahmin etme problemi ele alınmıştır. İlk olarak, uygulamada kullanılmak üzere bir veri seti seçilmiş ve bu veri seti eğitim ve test veri seti olarak ikiye ayrılmıştır. Veri madenciliği algoritmalarında KNN, Naive Bayes ve lojistik regresyon algoritmaları veri seti üzerine uygulanmıştır. Algoritmaların performansları mobil uygulama aracılığıyla değerlendirilmiş ve en yüksek doğruluk oranına sahip olan algoritma bu şekilde belirlenmiştir. Sonuç olarak, Naive Bayes algoritmasının %82.60, lojistik regresyon algoritmasının %82.94 ve KNN algoritması %77.59 olarak hesaplanmıştır. Sonuç lojistik regresyon algoritmasının en yüksek doğruluk oranına sahip olduğu görülmüştür. Mobil uygulama üzerinden elde edilen değerler ile Weka 3.9.6 programı ile elde edilen değerler karşılaştırılmış ve aynı sonuçlar bulunmuştur. Bu çalışma, veri madenciliği algoritmalarının kalp hastalığı tahmin etmede kullanılabileceğini ve farklı algoritmaların performanslarının karşılaştırılmasının önemini vurgulamıştır.

Uzman doktorlar ile yapılacak çalışmalar doğrultusunda kalp hastalığı teşhisinde kullanılabilecek yeni özniteliklerin belirlenmesi ve farklı veri setleri üzerinde çalışmaların yapılması klinik karar destek sistemlerinin altyapısını oluşturulmasında önemli bir rol oynayacağı düşünülmektedir.



KAYNAKLAR

1. Internet: World Health Organization, URL: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1 , Son Erişim Tarihi: 03.01.2022.
2. Internet : Data Portal For Statistics, <https://data.tuik.gov.tr/Bulten/Index?p=Death-and-Causes-of-Death-Statistics-2019-33710> , Son Erişim Tarihi: 24.06.2020.
3. Ayatollahi, H., Gholamhosseini, L. and Salehi, M. (2019). Predicting coronary artery disease: a comparison between two data mining algorithms. *BMC Public Health* 19, 448.
4. Joshi, S., Sasanapuri, A., Anand, S., Nandi, S. and Nemade, V. (2018). Predictive Analysis using Data Mining Techniques for Heart Disease Diagnosis. *International Journal of Engineering & Technology*, 7 (3.1), 166-170.
5. Obiri, D. A. and Sarku, E. (2020). Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning Algorithms. *International Journal of Computer Applications*, 176(11), 975-8887.
6. Anitha, S. and Sridevi, N. (2019). Heart Disease Prediction Using Data Mining Techniques. *Journal of Analysis and Computation*, 3(2).
7. Özcan, İ., Taşar, B., Tatar, A. B. ve Yakut O., (2019). Destek Vektör Makinesi Algoritması ile Kalp Hastalıklarının Tahmini. *Anatolian Journal of Computer Science*, 4(2), 74-79.
8. Patel, S. B., Yadav, P. K. and Shukla, D. P. (2013). Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. *IOSR Journal of Agriculture and Veterinary Science*, 4(2), 61-64.
9. Jabbar, M.A., Deekshatulua B.L. and Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, *Procedia Technology* 10, 85 – 94.
10. Amin, M. S., Chiama, Y. K. and Varathan, K. D. (2019). Identification Of Significant Features And Data Mining Techniques in Predicting Heart Disease. *Telematics and Informatics*, 36, 82-93.
11. Rajkumar, A. and Reena, G. S. (2010). Diagnosis Of Heart Disease Using Datamining Algorithm. *Global Journal of Computer Science and Technology*, 10(10), 38.
12. Chandna, D. (2014). Diagnosis of Heart Disease Using Data Mining Algorithm. *International Journal of Computer Science and Information Technologies*, 5 (2) , 2014, 1678-1680.
13. Palaniappan, S. and Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques, *IEEE/ACS International Conference on Computer Systems and Applications*, 108-115.

14. Abdar, M., Kalhori, S. R. N., Sutikno, T., Subroto, I. M. I. and Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering*, 5(6), pp. 1569-1576.
15. Alex P, M. and Shaji, S. P. (2019). Prediction and Diagnosis of Heart Disease Patients using Data Mining Technique. 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 0848-0852.
16. Bahrami, B. and Shirvani, M. H. (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques. *Journal of Multidisciplinary Engineering Science and Technology*, 2(2).
17. Bhatla, N. And Jyoti, K. (2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques. *International Journal of Engineering Research & Technology (IJERT)*, 1(8).
18. Ahmed, A. and Hannan, S. A. (2012). Data Mining Techniques to Find Out Heart Diseases: An Overview. *International Journal of Innovative Technology and Exploring Engineering*, 1(4).
19. Nguetilbaye A., Lei, L. and Wang, H. (2016). Comparative Study of Data Mining Techniques on Heart Disease Prediction System: a case study for the Republic of Chad. *International Journal of Science and Research (IJSR)*, 5(5), 6-391.
20. Chaurasia, V. and Pal, S. (2013). Data Mining Approach to Detect Heart Diseases. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, 2(4), 2013, 56-66.
21. Shafique, U., Majeed, F., Qaiser, H., and Mustafa, I.U. (2015). Data Mining in Healthcare for Heart Diseases, *International Journal of Innovation and Applied Studies*, 10(4), 1312-1322.
22. Venkatalakshmi, B. and Shivsankar, M.V. (2014). Heart Disease Diagnosis Using Predictive Data mining. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(3).
23. Bajaj, P. and Gupta, P. (2014). Review on Heart Disease Diagnosis Based on Data Mining Techniques. *International Journal of Science and Research (IJSR)*, 3(5).
24. Chaithra, N. and Madhu, B. (2018). Classification Models on Cardiovascular Disease Prediction Using Data Mining Techniques, *Journal of Cardiovascular Diseases & Diagnosis* 6(6).
25. Taneja, A. (2013). Heart Disease Prediction System Using Data Mining Techniques. *Oriental Journal Of Computer Science & Technology*, 6 (4), 457-466.
26. Dangare, C. S. and Apte, S. S. (2012). Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Applications*, 47(10), 2012.

27. Singh, P., Singh, S., Pandi-Jain, G. S. (2018). Effective heart disease prediction system using data mining techniques. *International Journal of Nanomedicine*, 2018,13,121–124.
28. Malmivuo, J. and Plonsey, R. (1995). *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetis Fields* (Press 3). UK: Oxford University Press,119–130.
29. İnternet: Kalbin Yapısı, URL: <http://evrensemizel.com/kategori/kalbin-yapisi>, Son Erişim Tarihi: 01.06.2014.
30. Türkiye Kalp ve Damar Hastalıkları Önleme ve Kontrol Programı Eylem Planı (2021-2026), Sağlık Bakanlığı, 1215, Ankara, 2021.
31. Uçar, E.ve Erzurum Alim, N. (2020). Kardiyovasküler Hastalıklarda Diyet Posasının Rolü. *Türkiye Sağlık Bilimleri ve Araştırmaları Dergisi*, 3(1), 1-10.
32. Türkiye Kalp ve Damar Hastalıkları Önleme ve Kontrol Programı Eylem Planı (2021-2026), Sağlık Bakanlığı, 1215, Ankara, 2021.
33. İnternet: İstatistik Veri Portalı, URL: <https://data.tuik.gov.tr/Bulten/Index?p=Death-and-Causes-of-Death-Statistics-2019-33710> , Son Erişim Tarihi: 24.06.2020.
34. Gullo, F. (2015). From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Physics Procedia*, 62,18-22.
35. Cheng, J. (2017). Data mining Research in Education. International School of Software, Wuhan University.
36. Arslan, Necati. (2018). Özelleştirilmiş Naive Bayes Algoritması. *Data Mining and Knowledge Discovery*.
37. İnternet : Şadi Evren Şeker, Karar Ağacı Öğrenmesi. URL: <https://bilgisayarkavramlari.com/2012/04/11/karar-agaci-ogrenmesi-decision-tree-learning/> , Son Erişim Tarihi: 11.04.2012.
38. Kuzey, C. (2012). Veri Madenciliğinde Destek Vektör Makinaları ve Karar Ağaçları Yöntemlerini Kullanarak Bilgi Çalışanlarının Kurum Performansı Üzerine Etkisinin Ölçülmesi ve Bir Uygulama, Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
39. Yılmaz, Ş. (2009). Veri Madenciliği: Karar Ağacı Algoritmaları ve İMKB Verileri Üzerine Bir Uygulama, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 14(1), 31-52.
40. Taşcı, E. ve Onan, A. (2017). K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi.
41. Kaya, Y. ve Yeşilova, A. (2011). İki Durumlu Karışımli Lojistik Regresyona İlişkin Bir

42. Serdarer Kuzu, B., Giray Yakut, S. (2020). Destek Vektör Makineleri Yardımıyla İmalat Sanayisinde Mali Başarısızlık Tahminlerinin Teknoloji Yoğunluğuna Göre İncelenmesi, *Osmaniye Korkut Ata Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 4(2), Aralık 2020, 36-54.
43. Gültekin, S. U. (2017). Veri Madenciliği: Yapay Sinir Ağı Ve Doğrusal Regresyon Yöntemleri İle Fiyat Tahmini, Yüksek Lisans Tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü, Denizli.
44. Pynam, V., Spanadna, R. R. and Srikanth, K. (2018). An Extensive Study of Data Analysis Tools (Rapid Miner, Weka, R Tool, Knime, Orange), *SSRG International Journal of Computer Science and Engineering (SSRG – IJCSE)*, 5(9).
45. İnternet :The Design View. URL: <https://docs.rapidminer.com/9.9/studio/getting-started/design-view.html>, Son Erişim Tarihi: 2023.
46. Tekerek, A. (2011). Veri Madenciliği Süreçleri Ve Açık Kaynak Kodlu Veri Madenciliği Araçları, *Akademik Bilişim*, 11, 2-4.
47. İnternet : R and RStudio, URL: https://web.itu.edu.tr/~tokerem/r_and_rstudio.html, Son Erişim Tarihi: 2021.
48. Dogan, O. (2017). Free Tools For Data Mining And A Research On Their Recognition In Turkey, *Ege Stratejik Araştırmalar Dergisi*, 8(1).
49. Aksu, G. ve Doğan, N. (2019). Veri Madenciliğinde Kullanılan Bir Analiz Programı: WEKA, *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 2019, 10(1),80-95.
50. Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., Bennie, E. and Tyson, M. (2017). MatLab vs. Python vs. R, *Journal of Data Science* 15(2017),355-372.
- 51.İnternet:MATLAB Programming, URL: https://en.wikibooks.org/wiki/MATLAB_Programming Son Erişim Tarihi: 17.11.2022.
52. Naika, A. and Samant, L. (2016). Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*,85, 662-668.
53. Kaya, M. ve Özel, S. A. (2014). Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması, *Akademik Bilişim’14 - XVI. Akademik Bilişim Konferansı Bildirileri*, 5 - 7 Şubat 2014 Mersin Üniversitesi.
54. Dwivedi, S., Kasliwal, P. and Soni, S. (2016). Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). 2016 Symposium on Colossal Data Analysis and Networking (CDAN), India, 1-8.
55. Han, J., Kamber M., and Pei J. (2012). *Data Mining Concepts and Techniques* (Second Edition). USA: CRC Press,1-4.

56. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
57. İnternet: Heart Disease Dataset, David Lapp, URL: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset> ,Son Erişim Tarihi: 2019.
58. Bilgin, M. (2017). Gerçek Veri Setlerinde Klasik Makine Öğrenmesi Yöntemlerinin Performans Analizi, 19.Akademik Bilişim Konferansı, Aksaray, 1-6.





ÖZGEÇMİŞ**Kişisel Bilgiler**

Soyadı, adı : NART, Ayşegül
Uyruğu : T.C.

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi / Sağlık Bilişimi	-
Lisans	Erciyes Üniversitesi / Bilgisayar Mühendisliği	2013
Lise	Kayseri Lisesi	2008

İş Deneyimi

Yıl	Yer	Görev
2016-Halen	Gelir İdaresi Başkanlığı	Bilgisayar Mühendisi
2014-2016	Mantis Yazılım	Yazılım Mühendisi

Yabancı Dil

İngilizce

Yayınlar

Determination of Effective Algorithm with Data Mining Techniques in Predicting Heart Disease(Tezden üretilmiştir), 8. Uluslararası Değişen Dünyada Fen, Sosyal Ve Sağlık Bilimleri Kongresi Bildiri Kitabı(2023).



GAZİLİ OLMAK AYRICALIKTIR.

