T.C.
İSTANBUL ÜNİVERSİTESİ-CERRAHPAŞA
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

# DOKTORA TEZİ

# LOJİSTİK REGRESYON ANALİZİ ve MAKİNE ÖĞRENMESİ ALGORİTMALARI ile PROSTAT KANSERİ RİSK FAKTÖRLERİNİN BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

**Selman AKTAŞ**

**DANIŞMAN**
**Prof. Dr. Murat KİRİŞCİ**

**Biyoistatistik Anabilim Dalı**

**Biyoistatistik ve Tıbbi Bilişimi Programı**

**Aralık, 2022**

Babama ithafen…

# BÜTÇE DESTEKLERİ

## LOJİSTİK REGRESYON ANALİZİ ve MAKİNE ÖĞRENMESİ ALGORİTMALARI ile PROSTAT KANSERİ RİSK FAKTÖRLERİNİN BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

Bu tez çalışması için herhangi bir kurumdan bütçe desteği alınmamıştır.

# TEŞEKKÜR

Kendiyle tanıştığım günden bugüne dek benden desteğini ve yardımını eksik etmeyen Danışmanım Prof. Dr. Murat KİRİŞCİ' ye Değerli dostlarım Dr.Öğr.Üyesi Kürşad Nuri BAYDİLİ, Doç.Dr. Eray Metin GÜLER ve Serdar POLAT' a Kıymetli eşim Eda AKTAŞ' a hayatta olduğum günden bu zamana kadar desteklerini biran olsun benden esirgemeyen annem Yetgin AKTAŞ' a baba yarısı abilerim Cem ve Vehip AKTAŞ' a teşekkür ederim

Aralık 2022 Selman AKTAŞ

# İÇİNDEKİLER

**Sayfa No**

**BEYAN** ..... **iii**
**BÜTÇE DESTEKLERİ** ..... **v**
**TEŞEKKÜR** ..... **vi**
**İÇİNDEKİLER** ..... **vii**
**ŞEKİL LİSTESİ** ..... **ix**
**TABLO LİSTESİ** ..... **x**
**SİMGE VE KISALTMA LİSTESİ** ..... **xi**
**ÖZET** ..... **xii**
**ABSTRACT** ..... **xiv**
**1. GİRİŞ** ..... **1**
**2. KAVRAMSAL ÇERÇEVE** ..... **3**
2.1. Literatür İncelemesi ..... 3
2.2. Lojistik Regresyon ..... 5
2.2.1. Giriş ..... 5
2.2.2. İki (Binary) Kategorili Lojistik Regresyon ..... 6
2.2.3. Regresyon Katsayılarının Kestirimi ..... 8
2.2.4. Katsayıların Anlamlılığı Testleri ..... 8
2.2.5. Bağımsız Değişkenlerin Modele Alınması ..... 9
2.2.6. Model Uyum İyiliği ..... 11
2.2.7. Açıklayıcılık Kat Sayıları ($R^2$) ..... 12
2.2.8. Modelin Yeterliliği ..... 14
2.2.9. Katsayıların Yorumlanması ..... 15
2.2.10. Uygun model seçimi ..... 16
2.3. Makine Öğrenmesi ..... 17
2.3.1. Öğrenme Türleri ..... 18
**3. YÖNTEM** ..... **26**
**4. BULGULAR** ..... **27**
**5. TARTIŞMA** ..... **43**

**6. SONUÇ VE ÖNERİLER ..... 45**
**KAYNAKLAR ..... 48**
**EKLER ..... 55**
**İNTİHAL RAPORU İLK SAYFASI ..... 58**
**ETİK KURUL İZİN YAZISI ..... 59**
**KURUM İZNİ YAZILARI ..... 61**
**ÖZGEÇMİŞ ..... 62**

# ŞEKİL LİSTESİ

**Sayfa No**

Şekil 2-1: Denetimli Öğrenme Modeli Süreci ..... 18

Şekil 2.2 Hyperplance ve Destek Vektörleri ..... 21

Şekil 4-1: Verinin programa tanıtılması ..... 32

Şekil 4-2: Eğitim verisinin oluşturulması. ..... 32

Şekil 4-3: Test verisinin oluşturulması ..... 33

Şekil 4-4: SVM eğitim ve deneme veri setine uygulanması. ..... 33

Şekil 4-5: SVM analizi ile oluşturulan modelin değişkenleri. ..... 34

Şekil 4-6: KNN analizi ile oluşturulan modelin değişkenleri ..... 35

Şekil 4-7: CHAİD analizi ile oluşturulan modelin değişkenleri ..... 36

Şekil 4-8: CHAİD analizi karar ağacı. ..... 38

Şekil 4.9 C5.0 analizi karar ağacı ..... 40

Şekil 4-10: C5.0 analizi karar ağacı ..... 40

Şekil 4-11: Lojistik regresyon ve oluşturulan modellerin karşılaştırılması. ..... 41

Şekil 4-12: LR, SVM, KNN, CHAİD, C5.0 analizlerinin gainlerinin karşılaştırılması. ..... 41

# TABLO LİSTESİ

**Sayfa No**

Tablo 2-1: Tahmin edilen frekanslar ( $E_{kj}$) ve gözlemlerin kontenjans tablosu ($O_{kj}$).....12

Tablo 2-2: Kontenjans tablosu.....15

Tablo 2-3. Dizayn değişkeni.....16

Tablo 4-1. Katılımcılara ait demografik bilgiler. .....27

Tablo 4-2. Katılımcılara ait demografik bilgiler (devamı). .....28

Tablo 4-3. Katılımcıların şikayetleri. .....29

Tablo 4-4. Bağımsız değişkenlerin seçimi. .....30

Tablo 4-5. İkili Lojistik regresyon analizi sonuçları. .....31

Tablo 4-6. Sınıflandırma tablosu.....31

Tablo 4-7: SVM modelinin sınıflandırma tablosu.....34

Tablo 4-8: SVM öğrenmiş modelin sınıflandırma tablosu.....35

Tablo 4-9: KNN modelinin sınıflandırma tablosu.....36

Tablo 4-10: KNN ile eğitilmiş modelin sınıflandırma tablosu.....36

Tablo 4-11: CHAİD sınıflandırma tablosu.....37

Tablo 4-12: CHAİD ile eğitilmiş modelin sınıflandırma tablosu.....37

Tablo 4-13: C5.0 sınıflandırma tablosu.....39

Tablo 4-14: C5.0 ile eğitilen modelin sınıflandırma tablosu.....39

Tablo 4-15: Modellerin risk faktörleri.....42

Tablo 4-16: Modellerin sınıflandırma yüzdeleri. .....42

# SİMGE VE KISALTMA LİSTESİ

| Kısaltmalar | Açıklama |
|---|---|
| **AIC** | **:** Akaike'nin bilgi kriteri |
| **BIC** | **:** Bayesian bilgi kriteri |
| **CHAID** | **:** Otomatik ki-kare etkileşim belirleme |
| **Gain** | **:** Bilgi kazancı |
| **KNN** | **:** K-En yakın komşu |
| **LR** | **:** Lojistik regresyon |
| **OR** | **:** Odds Oranı |
| **SVM** | **:** Destek vektör makinesi, Makine destek vektörü |

# ÖZET

## DOKTORA TEZİ

### LOJİSTİK REGRESYON ANALİZİ ve MAKİNE ÖĞRENMESİ ALGORİTMALARI ile PROSTAT KANSERİ RİSK FAKTÖRLERİNİN BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

**Selman AKTAŞ**

**İstanbul Üniversitesi-Cerrahpaşa**

**Lisansüstü Eğitim Enstitüsü**

**Biyoistatistik Anabilim Dalı**

**Biyoistatistik ve Tıbbi Bilişimi Programı**

**Danışman: Prof. Dr. Murat KİRİŞCİ**

Bağımlı değişkenler ile bağımsız değişkenler arasındaki neden-sonuç ilişkileri regresyon modelleri ile çözümlenmektedir. Regresyon modelleri kullanılan veri yapıları ve istenilen ön koşullar ile çeşitlenmektedir. Açıklanan (bağımlı) değişkenin niteliksel veri türünde olduğu durumlarda Lojistik Regresyon analizleri kullanılmaktadır. Lojistik regresyon analizleri (LR) bağımlı değişkeninin kategori sayısına göre çeşitlenmektedir. Bu çalışmada sadece bağımlı değişkenin kategori sayısının iki olduğu durumlarda kullanılan ikili (binary) lojistik regresyon analizi kullanılmıştır. İkili lojistik regresyon analizi özellikle tıp alanında hastalığa etki eden çevresel risk faktörlerini belirlemek, hasta-sağlıklı bireyleri sınıflandırmak ve tahminlemek amacıyla sıklıkla kullanılmaktadır.

Bilgi çağı olarak nitelendirilen içinde bulunduğumuz çağ internet ve bilgisayar alanındaki gelişmeler ile yeni kavramları hayatımıza taşımıştır. Şüphesiz bu kavramların başında özellikle matematik, istatistik ve veri madenciliği alanlarının konusuna giren makine öğrenmesi kavramı gelmektedir. Bu kavram Alan. M. Turing tarafından 1950 yılında ortaya

atılan makinelerin insanlar gibi düşünebilir mi? sorusundan türemiştir. Makine öğrenmesi yöntemleri günümüzde finans, mühendislik, otomotiv ve tıp gibi çeşitli alanlarda kullanılmaktadır. Makine öğrenmesinde; matematiksel ve istatistiksel yöntemleri temel alarak oluşturulan algoritmalar kullanılır. Çeşitli algoritmalarla eğitilen modeller, öğrenme işlemini tamamlayarak denemeler yapar. Makine öğrenmesinde temel olarak üç çeşit öğrenme vardır. Bu çalışmada girdi (bağımsız) değişkenleri ile çıktı (bağımlı) değişkenlerinin olduğu makine öğrenmesi türü olan denetimli (gözetimli) öğrenme kullanılmıştır. Makine öğrenmesinde temel amaç bağımsız değişkenler ile bağımlı değişkenin tahminini yaparak öğrenmeyi sağlamaktır.

Bu çalışma da makine öğrenme yöntemleri ve lojistik regresyon analizini kullanarak prostat kanseri risk faktörlerini belirleyerek bu yöntemlerden en iyi sınıflandırmayı sağlayan yöntemi bulmayı amaçlamaktadır.

Türkiye ve dünya genelinde erkeklerde yaygın olarak görülen prostat kanseri hastalığının risk faktörleri belirlenmiştir. 248'i hasta 253'ü sağlıklı 501 kişi ile analizler gerçekleştirilmiştir. Hastalığa ait risk faktörleri makine öğrenmesi yöntemlerinden olan destek makine vektörü (SVM), K-en yakın komşuluğu (KNN), Otomatik ki-kare etkileşim belirleme (CHAİD), C5.0 algoritmalarıyla ve LR ile ayrı ayrı belirlenmiştir. Oluşturulan modellerde yaş, sigara kullanımı ve ailede kanser hastalığının varlığı ortak risk faktörleri olarak bulunmuştur. Bunun yanında modellerde menide veya idrarda kan görülmesi, idrara çıkma sıklığı, gün içindeki yaşam şeklinin risk faktörü olduğu tespit edilmiştir. Doğru sınıflama yüzdesi LR'nin %92,2 SVM'nin %89,92 KNN'nin %88,48 CHAİD'in %91,36 ve C5.0'ın %88,48 dir. Prostat kanseri hastalarını sınıflandırmada en iyi sonucu İkili Lojistik Regresyon analizinin verdiği görülmüştür. Çalışmanın örneklem büyüklüğü yeterli olsa da yapılacak çalışmalarda gerçek veriler üzerinden daha büyük örneklemlerde analizlerin yapılması tavsiye edilmektedir.

Aralık 2022 , 77. sayfa.

**Anahtar kelimeler:** Lojistik regresyon analizi, Makine öğrenmesi, Sınıflama, Prostat Kanseri

# ABSTRACT

## Ph.D. THESIS

### A STUDY ON THE DETERMINATION OF PROSTATE CANCER RISK FACTORS BY LOGISTIC REGRESSION ANALYSIS and MACHINE LEARNING ALGORITHMs

**Selman AKTAŞ**

**İstanbul University-Cerrahpaşa**

**Institute of Graduate Studies**

**Department of Biostatistics**

**Biostatistics and Medical Informatics Program**

**Supervisor: Prof. Dr. Murat KİRİŞCİ**

Cause-effect relationships between dependent variables and independent variables are analyzed with regression models. Regression models vary with the data structures used and the desired preconditions. Logistic Regression analyzes are used in cases where the explained (dependent) variable is of qualitative data type. Logistic regression analyzes (LR) vary according to the number of categories of the dependent variable. In this study, binary logistic regression analysis was used, which is used only when the number of categories of the dependent variable is two. Binary logistic regression analysis is frequently used, especially in the field of medicine, to determine the risk factors affecting the disease and to classify and predict patients and healthy individuals.

The age we live in, which is described as the information age, has brought new concepts to our lives with the developments in the field of internet and computer. Undoubtedly, the first of these concepts is the concept of machine learning, which is especially encountered in the fields of mathematics, statistics and data mining. This concept is

Field. Can machines, introduced by M. Turing in 1950, think like humans? derived from the question. Machine learning methods are used today in various fields such as finance, engineering, automotive and medicine. In machine learning, algorithms based on mathematical and statistical methods are used. Models trained with various algorithms complete the learning process and experiment. There are basically three types of learning in machine learning. In the study, supervised (supervised) learning, which is a type of machine learning with input (independent) variables and output (dependent) variables, was used. The main purpose is to provide learning by estimating the independent variables and the dependent variable.

The study aims to determine the prostate cancer risk factors using machine learning methods and logistic regression analysis and to find the method that provides the best classification from these methods.

Risk factors for prostate cancer, which is common in men in our country and around the world, have been determined. Analyzes were carried out with 501 people, 248 of whom were patients and 253 were healthy. The risk factors for the disease were determined separately by machine learning methods such as support machine vector (SVM), K-nearest neighbor (KNN), Automatic chi-square interaction detection (CHAID), C5.0 algorithms and LR. In the models created, age, smoking and the presence of cancer in the family were generally obtained as risk factors. In addition, it has been determined that blood in the semen or urine, frequency of urination, and daily life styles are risk factors in the models. The correct classification percentage is 92.2% of LR, 89.92% of SVM, 88.48% of KNN, 91.36% of CHAID and 88% of C5.0. It was seen that Binary Logistic Regression analysis gave the best results in classifying prostate cancer patients. Although the sample size of the study is sufficient, it is recommended to conduct analyzes in larger samples based on real data in future studies

December 2022, 77. pages.

**Keywords:** Logistic regression analysis, Machine Learning, Classification, Prostat Cancer

# 1. GİRİŞ

Tıp alanındaki çalışmalar, bağımlı değişkene etkisi olan bağımsız değişkenlerin ayrı ayrı ve tümel olarak etkisini araştırmak istemektedir(Atakurt,1999). Bağımlı değişkenin üzerinde görülen etkinin bağımsız değişkenler tarafından açıklanmasını sağlayan teknikler çok değişkenli analiz yöntemleri ile gerçekleştirilir. Ancak bu yöntemleri kullanabilmek için veri türlerinin, kullanılacak olan yöntemlere uygun olması gerekmektedir. Açıklanan değişkenin nitel olması ve iki ya da ikiden çok kategorisi olması durumunda lojistik regresyon yöntemi ile analiz gerçekleştirilmektedir (Alpar, 2017).

Lojistik regresyon analizi diğer çok değişkenli yöntemler gibi herhangi bir ön koşulda bulunmaz (Bewick ve ark.,2005). Bağımsız değişkenlerin normal dağılım göstermesi, bağımlı değişkenle doğrusal ilişki içinde olması ya da grupların eşit varyans içinde olması varsayımları yoktur (Tabachnick ve Fidell, 2015). Ön koşulların olmaması uygulama kolaylığı sağlaması bu analizin popüler hale gelmesini sağlamıştır. 1985-1989 yılları arasında yapılan bir çalışma lojistik regresyon analizinin en sık kullanılan çok değişkenli analiz yöntemi olduğunu belirtmiştir (Concato ve ark.,1993). Lojistik regresyon analizi; bağımsız değişkenlerin, bağımlı değişkenler üzerinde olan etkilerini dikkate alarak bağımlı değişkenin kategorilerine göre kategorilerin sınıflandırılmasını da sağlar (Özdamar, 2011). Analiz gerçekleştirilirken kurulan model sayesinde veri kümesine eklenecek yeni gözlemlerin hangi gruba dahil olacağı tahminini de yapar (Başarır, 1990). Lojistik regresyon analizi sınıflama ve tahmin özellikleriyle sağlık alanında etkin bir yöntem olarak kullanılmaktadır.

Gelişen bilgisayar teknolojileri ile bilgiye ulaşım, işleme ve depolama hususlarında da büyük gelişmeler sağlanarak hayatımıza "Büyük Veri" kavramı eklendi. Büyük veri, devletlerin kurum ve kuruluşların farklı veri yapılarını birleştirip istatistik ve diğer veri madenciliği yöntemleriyle bilgiler edinmesi olarak tanımlanır (Rubinstein, 2013). Verinin boyut, hacim ve hız açısından artması ile 1950'li yıllarda ortaya çıkan "Makine Öğrenmesi" yöntemleri de trend hale geldi. Çünkü bu yöntemler büyük veri de daha az maliyet ve daha iyi sonuçlar elde edilmesini sağladı (Başer ve ark., 2021).

Makine öğrenmesi, bir problemi o probleme ait veriye göre modelleyen bilgisayar algoritmalarının genel adıdır (Atalay ve Çelik, 2017). Makine öğrenme yöntemleri için alan yazında çeşitli sınıflandırmalar mevcuttur. Genel olarak makine öğrenmesi üç ana kategoriye ayrılmaktadır bunlar; gözetimli (denetimli) öğrenme, gözetimsiz (denetimsiz) öğrenme ve Güçlendirmeli (Takviyeli) Öğrenme (Saravanan ve Sujatha, 2018).

Gözetimli (denetimli) öğrenme bağımlı değişkenin bilindiği ve bağımsız değişkenler ile bağımlı değişkeni açıklayabilen algoritmalardır. Bu algoritmalar sınıflandırma ve grup ataması işlemlerini yapar. Gözetimli öğrenme yöntemlerinde Naive Bayes, makine destek sistemleri, diskriminant analizi ve karar ağaçları gibi algoritmaları kullanır (Taiwo,2010). Gözetimsiz (denetimsiz) öğrenimi türü, analizi denetleyebilecek bir bağımlı değişkeni olmadığı için gözetimsiz (denetimsiz) olarak nitelendirilir (James ve ark, 2013). Güçlendirmeli öğrenme gözetimli ve gözetimsiz öğrenme yöntemlerinin birleşimidir (Mahesh, 2018).

Çalışmada, tıpta hastalık risk faktörlerini belirleme de yaygın kullanımı olan Lojistik Regresyon analizi ile makine öğrenme algoritmalarının risk faktörü belirleme, sınıflama ve tahmin yetenekleri incelecektir. Lojistik regresyon analizi ve makine öğrenme yöntemlerinin kıyaslandığı çalışmalar farklı alanlarda mevcuttur. Ancak literatür incelendiğinde prostat kanseri üzerine bunun benzeri bir çalışma mevcut değildir.

Amerika da yapılan bir araştırmada prostat kanserinin erkeklerde en çok görülen ikinci kanser türü olduğu ve bu kanser türünün erkeklerde kansere bağlı ölümlerde ikinci sırada olduğu tespit edilmiştir. (Greenlee, 2000). Türkiyede de benzer çalışmalar incelendiğin bu kanser türünün erkeklerde yaygın olduğu görülmüştür. Görülme sıklığı bu denli fazla olan prostat kanseri özelinde; gelişmekte olan makine öğrenmesi yöntemleri ve lojistik regresyon analizi ile risk faktörlerini belirlemek ve sınıflama yaparak daha sonra sonuçları karşılaştırmak amaçlanmıştır.

# 2. KAVRAMSAL ÇERÇEVE

## 2.1. Literatür İncelemesi

Literatür incelendiğinde tıp alanında Lojistik regresyon analizi ve makine öğrenmesi algoritmaları kullanılarak hastalık tespiti ile ilgili birçok araştırma yapıldığı görülmüştür.

Hagan R. ve arkadaşları (2021); kalp hastalığını tahmin edebilmek için UCI ve Kaggle'dan aldıkları veriler ile makine öğrenmesi algoritmalarından SVM, sinir ağlarını ve toplu öğrenmeyi kullanarak çalışmalarını gerçekleştirmiştir. UCI veri e ile SVM algoritmasıyla %92 doğruluk elde ederken, Kaggle verileri için %72 doğruluk elde edilmiştir. Çok katmanlı sinir ağları ile UCI verileri için %74 doğruluk elde ederken, Kaggle verileri için %71 elde edilmiştir. Toplu öğrenme ile UCI verilerinde %90'ın üzerinde doğruluk elde ederken, Extra Tree ile %96 doğruluk elde edilmiştir. Çalışma sonucunda karar ağaçlarının az maliyet ile yüksek doğruluk oranı sağladığını tespit etmişlerdir.

Shah ve Ark., (2020); UCI 'den aldıkları veriler ile kalp hastalığı tahmin edebilmek için WEKA yazılımı ile makine öğrenmesini kullanmıştır. Makine öğrenmesi algoritmalarından Naive Bayes, k-en yakın komşu, karar ağacı ve rastgele ormanı kullanarak çalışma sonunda en yüksek doğruluğu k=7 ile k-en yakın komşu algoritmasından elde etmişlerdir.

İşeri (2014); mamogram görüntülerinden makine öğrenmesi yöntemleri ile mikro kireçlenme (MC) tespitini CAD sistemler ile yaparak, meme kanserini tahmin etmeye çalışmıştır. Çok katmanlı ileri beslemeli yapay sinir ağı sınıflandırıcısı ile SVM makine öğrenme algoritması kullanılmıştır. Karar sürecinde test ayırt ediciliği, alıcı işlem karakteristikleri eğrisi (ROC) ile yapılmıştır. Mikro kireçlenme tespitinde SVM ile en iyi sonuç EWD2 ve GLCM yöntemleriyle elde edilmiştir. CAD sistemlerinin farklı kanser türleri için geliştirilmesi gerektiği savunulmuştur.

Yıldız (2019), makine öğrenmesi yöntemleri ile tiroit hastalığının teşhisi için iki ayrı veri grubu ve tahlil sonuçları kullanılmıştır. Çok katmanlı algılayıcı ağı ile sınıflama yapılmış, Knn algoritması kullanılmış ve performans bakımından değerlendirilmiştir. Çalışma

sonucunda çok katmanlı yapay sinir ağları sonuçlarının doğruluk oranı KNN ile yapılan çalışmanın sonuçlarının doğruluk oranından yüksek olduğu tespit edilmiştir.

Örenç (2019), KOAH hastalığının tespiti için PSG cihazına alternatif olarak makine öğrenmesinden yararlanmıştır. KNN algoritması, SVM ve Olasılıksal Yapar Sinir Ağları sınıflandırma algoritmalarından yararlanılmıştır. 8 KOAH hastası, 6 sağlıklı bireyin verileri kullanılmıştır. PPG sinyalinin makine öğrenmesi yöntemleri ile kullanıldığında hastalığın teşhis edilebileceği ve sistemin geliştirilerek hastalığın erken teşhisinde etkin olarak kullanılabileceği savunulmuştur.

Yahyaoui (2017), göğüs hastalıklarının erken teşhisi için SVM, uyarlamalı destek vektör makine (ASVM), KNN, NB makine öğrenme algoritmalarını kullanmıştır. Diyarbakır Göğüs Hastalıkları Hastanesi'nden elde edilen veriler literatürdeki diğer sinir ağları yöntemleri ile kıyaslanmıştır. Çalışma sonucunda ASVM metodunun yüksek doğruluk oranına sahip olduğu, göğüs hastalıklarının tespitinde yararlanılabileceği sonucuna varılmıştır.

Tanık (2021), polinöropati hastalığının erken teşhisi için EMG analizi ile 72 sağlıklı ve 28 polinöropati verilerini makine öğrenmesi algoritmalarına uygun şekilde yapılandırmıştır. EMG verileri verimliliği arttırmak için, sağlık hizmeti sunucularından destek alarak, işlem öncesinde veri ön işleme, öznitelik çıkarma ve gruplandırma işlemlerinden geçirilmiştir. Makine öğrenme sınıflandırma algoritmalarından Karar Ağaçları NB, K-NN, SVM ve Rastsal Orman sınıflandırma ile sınıflandırma yapılmıştır. Yapılan çalışma sonucunda Naive Bayes algoritmasının, diğer sıralama algoritmalarına göre hastalık tespitinde daha yüksek doğruluk oranında olduğu saptanmıştır.

Yılmaz, (2022), PACS arşivinden alınan bilgisayarlı tomo ervikal, torakal, lumbal ve sakral MPR ile ortogonal düzleme getirilerek T1, T11 ve T12 vertebrasından 15 parametre ölçülmüştür. Cerrahi müdahale görmemiş 50 kadın, 50 erkek çalışmaya dahil edilmiştir. Çalışma sonucunda cinsiyet tahmininin vertabralardan makine öğrenme algoritmaları ile yüksek doğruluk oranında olduğu saptanmıştır.

Okal, (2019), Meme kanseri ön tetkiklerinde termal görüntüler aracılığıyla makine öğrenmesi DVM algoritması ile sınıflandırarak teşhis edilmesi amaçlanmıştır. Çalışma sonucunda ön tetkik işleminde termal görüntüleme ile meme kanseri tespitinde ümit verici bir çalışma olduğu saptanmıştır.

Ayaz, (2021). Covid-19 tespiti için tam kan sayımı yapılan veri setlerinden SARS-CoV-2 pozitif olanların makine öğrenmesi algoritmaları ile tespit edilebilmesi için yapılmıştır. Veriler normal ve genel olarak ikiye ayrılarak KNN, NB, SVM makine öğrenme algoritmaları ile incelenmiştir. SVM algoritmasının kullanılan diğer algoritmalara oranla sonuçlarının yüksek doğruluk oranında olduğu ve hastalık teşhisi için erken tanı imkânı sunabileceği öngörülmüştür.

Oğuztürk, G. (2018). Diyabet hastalığının tespiti için makine öğrenmesi yöntemlerinden kNN, karar ağacı, Random Forest, Neural Network, NB, Logistic Regression, CN2 rule inducer algoritmaları ile incelenmiştir. Neural Network algoritmasının diğer algoritmalara oranla sonuçlarının yüksek doğruluk oranında olduğu ve hastalık teşhisi için erken tanı imkânı sunabileceği öngörülmüştür.

## 2.2. Lojistik Regresyon

### 2.2.1. Giriş

Lojistik regresyon (LR) analizi temeli 19. yüzyıla dayanan istatistikte yaygın olarak kullanılan bir analiz çeşididir. Analizin ortaya çıkışının nüfus artışı hesaplaması çalışmaları sırasında Malthus'un (1789) "kendi haline bırakılan bir nüfus geometrik olarak artacaktır" ifadesi üzerine ortaya çıktığı düşünülmektedir. Aslen gökbilimci olan Belçika uyruklu Aphonse Quetelet bu ifadeden yola çıkarak çalışmalarda bulunmuş üstel olarak hesaplamalar yapmayı denemiştir. Quetelet'in öğrencisi olan Pierre François Verhulst' da hocasının isteği üzerine bu konuda çalışmalara başlamıştır. 1830-1850 yılları arasında çalışmalarından üç makale yazan Verhulst, böylelikle Lojistik Regresyon Analizinin temelini atmış oldu (Bacaer, 2011). 1920'li yıllarda Verhulst'un çalışmaların dan haberleri olmadığını belirten ABD'li bilim adamları Pearl ve Reed tarafından ikinci kez keşfedildi (Pearl ve Red, 1920).

LR analizi sağlık bilimlerinde, bağımlı değişkenin iki ya da daha çok düzey içerdiği, bağımsız değişkenlerin ise hem sürekli hem de süreksiz veri türünde olduğunda gözlemelerin ait oldukları gruplara atama yapmak ve hastalıklara ilişkin risk faktörlerini belirleyecek modeli kurmak amacıyla sıklıkla başvurulan bir yöntemdir. (Elhan, 1997). Ek olarak lojistik regresyon analizinde bağımsız değişkenler kolayca yorumlanabildiğinden dolayı da sıklıkla tercih nedeni olmaktadırlar (Cox ve Snell,1989).

Lojistik model veya logit model olarak adlandırılan lojistik regresyon, bağımsız değişkenler ile kategorik bir bağımlı değişken arasındaki ilişkiyi inceleyip, verileri bir lojistik eğriye uydurarak bir olayın meydana gelme olasılığını tahmin eder (Park, 2013). Lojistik regresyon kısacası LR bağımlı değişkenin üzerinde etkili olan bağımsız değişkenlerin etkilerini (Odds Ratio) ortaya koyan bir analizdir (Özdamar, 2011). LR bir fenomenin ortaya çıkmasına neden olan etkenin risk faktörü olmasını belirlememizde ve bu risk faktörlerinin fenomene etkilerini ölçmemizi sağlar (Agresti, 1996).

LR'de diğer regresyon analizlerinde olduğu gibi bağımsız değişkenlerin varlığında kestirim yapmayı da sağlar (Karabulut ve Alpar, 2017). Ayrıca LR bağımsız değişkenlerle, bağımlı değişkenleri hesaplayarak sınıflama ve gruplara atama işlemi yapmamızı da sağlar (Press ve Wilson, 2012).

Bağımlı değişkenin veri türüne göre farklı LR yöntemleri geliştirilmiştir. Bağımlı değişkenin iki kategorili olması durumunda İkili (Binary) LR, kategori sayısı üç ve üzerinde olduğunda Çok Kategorili LR ve bağımlı değişken ordinal olduğunda Sıralı (Ordinal) LR yöntemi kullanılmaktadır (Liu, 2015). Çalışmamız da sadece İki kategorili LR üzerinde durulacaktır.

### 2.2.2. İki (Binary) Kategorili Lojistik Regresyon

Bağımlı değişkenin kategori sayısının iki olduğu durumlarda kullanılan LR şeklidir. Bağımlı değişken kategorileri risk var- yok, hastalık var-yok şeklinde isimlendirilir. Kodlama yapılırken genellikle maruziyetin olduğu durum 1 olmadığı durum 0 ile kodlanır. LR analizinde bağımsız değişkenler, sürekli veya süreksiz değişken türünde olabilir. Ayrıca sürekli ve süreksiz veri türünde olan bağımsız değişkenler bir arada da modelde bulunabilir (Şeneş ve Alatlı,2014). Bunun yanında LR analizinde bağımsız değişkenlerin dağılımına ilişkin de herhangi bir önkoşul yoktur (Aktaş, 2009).

Lojistik regresyonda bağımlı değişkenin değerini tahmin etmek amaçlanmaz, bağımlı değişkenin maruziyetin olması olasılığı tahmin edilmeye çalışılır (Alpar, 2017). Lojistik model, ortaya çıkacak riski 0 ile 1 arasında herhangi bir değer olarak tahmin etmeye yarar (Hosmer ve Lemeshov, 1980). Bu modelle ede edilen tahmin değeri 1'in üstünde veya 0'ın altında bir risk değeri almamaktadır.

***Bir bağımsız değişken ile oluşturulan LR modeli;***

$$E\left(Y \mid X = x\right) = \beta_0 + \beta_1 x \tag{2-1}$$

$$\pi(x) = P\left(Y = 1 \,/\, X = x\right) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{2-2}$$

***Birden çok bağımsız değişken ile oluşturulan LR modeli;***

Eşitlik 2.3'te T bağımsız değişkenleri ifade ederken, s ise değişken sayısını ifade etmektedir.

$$\text{T}= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + \beta_s X_s \tag{2-3}$$

$$P(Y) = \frac{e^t}{1+ e^t} = \frac{1}{1+e^{-t}} \tag{2-4}$$

Bağımsız değişkenlerin, bağımlı değişken üzerindeki etkileri, bağımlı değişkenin iki düzeyinden herhangi birine karşı diğerinin olma olasılığı kullanılarak belirlenmektedir (Arabacı, 2002). Bu durum Odds oranı (OR) ile açıklanmaktadır. OR bir olayın gerçekleşmesi olasılığı ile olayın gerçekleşmemesi olasılığının oranı birbirine oranı ile hesaplanır. LR analizinde genellikle kat sayılar en çok olabilirlik oranı ile tahmin edilir (Berenson ve Levine, 1996). En çok olabilirlik oranı da odds oranının doğal logaritması alınarak hesaplanır (Karagöz, 2021).

$$\frac{P(Y)}{1-P(Y)} = \frac{\frac{e^t}{1+ e^t}}{1-(\frac{e^t}{1+ e^t})} = e^t = e^{\beta_0+\beta_1 X_1 +\beta_2 X_2 \ldots\ldots\ldots\ldots+\beta_s X_s} \tag{2-5}$$

Eşitlik 2-5 'in doğal logaritması alındığında kat sayılar hesaplanmış olacaktır.

$$Ln\frac{P(Y)}{1-P(Y)} = ln e^{\beta_0+\beta_1 X_1 +\beta_2 X_2 \ldots\ldots\ldots+\beta_s X_s} = \beta_0 + \beta_1 X_1 \ldots\ldots\ldots + \beta_s X_s \tag{2-6}$$

Eşitlik 2-6 ile lojistik regresyon analizi denkleminin katsayıları tahmin edilir. Bu işlem aynı zaman da lojit model ile bağımsız değişkenler arasındaki ilişkinin de doğrusal hale gelmesini sağlamaktadır (Alpar, 2017). Doğrusal hale gelen lojit model, doğrusal regresyonda istenilen özelliklerin çoğunu sağlamaktadır. Sadece farklılık doğrusal regresyonda artıklar normal dağılım gösterirken LR de artıkların binom dağılımı göstermesidir (Hosmer ve Lemeshow, 2000).

### 2.2.3. Regresyon Katsayılarının Kestirimi

Regresyon modelleri ile LR'nin bir diğer farkı katsayıların kestiriminde en küçük kareler (EKK) yöntemi yerine, en çok olabilirlik (maximum likelihood) yöntemi kullanmasıdır (Çokluk,2010). Çünkü en küçük kareler yönteminde bağımlı değişkenin normallik varsayımı bulunmaktadır. En çok olabilirlik yöntemi olayın gerçekleşme olasılığını en az yapmak yerine, bir olayın olma gerçekleşme olasılığını en çok yapmayı amaçlar (Hair ve ark., 2006). En çok olabilirlik yöntemi haricinde kullanılan yöntemlerde bulunmaktadır. Ağırlıklandırılmış EKK, yeniden ağırlıklandırılmış EKK ve diskriminant kat sayıları ile de katsayı kestirimi yapılmaktadır (Tatlıdil, 2002). İstatistik tabanlı yazılımlar LR'nin katsayı kestiriminde sıklıkla en çok olabilirlik yöntemini kullanmaktadırlar.

En çok olabilirlik yöntemi örneklem büyüklüğünden etkilenmektedir. Bu nedenle alan yazında çeşitli görüşler bulunmaktadır, her bağımsız değişken için en az 50 gözlemin alınması gerektiği yaygın görüşlerden biridir (Raymond, 1994). Başka bir görüş ise örneklem büyüklüğünün açıklayıcı değişken sayısının en az 10 katı kadar olması gerektiği görüşüdür (Akgül ve Çevik, 2003). Bu görüşlerin haricinde eğer veri setindeki bağımsız değişkenler arasında yüksek korelasyon varsa ya da bağımlı değişkenin kategorilerinden sadece birinde yığılma varsa örneklemin büyük olması beklenir (Şahin, 2011).

### 2.2.4. Katsayıların Anlamlılığı Testleri

LR de bağımsız değişkenlerde kestirim yapıldıktan sonra yapılan kestirimlerin anlamlılığı denetlenir. Denetleme işleminde amaç bağımlı değişkeni açıklamak için kurulan modelde bağımsız değişkenlerin modele katkı sağlayıp sağlayamadığını belirlemektir. Bu sınama işlemleri olabilirlik oran testleri, Wald ve Skor testleri ile gerçekleştirilir (Hosmer ve Lemeshow, 1989).

#### *2.2.4.1. Olabilirlik Oran Testi*

Bütün değişkenlerin dahil olduğu model ile kestirim yapılan modelin olabilirlik oranlarının farklarının ki -kare dağılacağı varsayılarak model geçerliliğini araştırır (Elhan, 1997). Olabilirlik oran testi likelihood veya G ile ifade edilir ve eşitlik 2-7'deki gibi hesaplanır.

$$G = -2\ln\left[\frac{Değişkenin\ olmadığı\ olabilirlik}{Değişkenin\ olduğu\ olabilirlik}\right] \qquad (2\text{-}7)$$

Sapma (deviance) testi olarak da adlandırılan bu test doğrusal regresyondaki hata kareler toplamına denk gelmektedir (Hosmer-Lemeshow,1989). Olabilirlik oran testinde G

değerinin büyük olması modele eklenen değişkenin lojit'in tahminine katkı sağladığı düşük çıkması katkı sağlamadığı anlamına gelmektedir (Alpay, 2011).

#### *2.2.4.2. Wald Testi*

Bu analiz regresyon analizinde bulunan z ve t testlerine benzemektedir (Wald, 1943). Wald test istatistiği, regresyon katsayısının en çok olabilirlik kestiriminin standart hatasına oranıdır (Brant, 1990). Wald istatistiği eşitlik 2-8 formülüyle elde edilir, formülde $\hat{\beta}_j$ , $\hat{\beta}_1$'in en çok olabilirlik tahmini $S(\hat{\beta}_j)$ ise , $\hat{\beta}_1$'in standart hatasıdır.

$$W = \frac{\hat{\beta}_j^2}{S(\hat{\beta}_j)^2} \tag{2-8}$$

$H_0$: $\beta_1$= 0 'in hesaplanmasında normal dağılımdan faydalanılır (Hosmer ve Lemeshow, 1989).

#### *2.2.4.3. Skor Testi*

Score testi genel olarak, matris hesaplamalarını gerektiren çok değişkenli bir testtir (Hosmer ve Lemeshow,1989). Skor testi hesaplama açısından kolaylık sağlar çünkü en çok olabilirlik kestiriminin hesaplamasına gerek duymaz (Alpar,2017).

$$ST = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{\sqrt{y_i}-(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-9}$$

Eşitlik 2-9 ile hesaplanan skor testi normal dağılıma uygunluk göstermektedir.

### 2.2.5. Bağımsız Değişkenlerin Modele Alınması

#### *2.2.5.1. Niceliksel (Sayısal) Değişkenlerin Modele alınması*

Bilindiği üzere lojistik regresyon analizinde bağımsız değişkenin türü ile ilgi olarak bir kısıtlama yoktur (Şeneş ve Alatlı, 2014). Sayısal değişkene herhangi bir işlem yapılmadan analize dahil etme açısından bir sorun yoktur. Sayısal değişkenlerin katsayıları, değişkenin 1 birim değişim göstermesi lojit'te ne kadarlık bir değişim sağladığını gösterir (Karagöz, 2021). Değişimin geçerli olması için bağımsız değişken ile lojit doğrusal bir ilişki içinde olması gerekmektedir. Bağımsız değişkenin lojitle doğrusal ilişkisini incelemek için Box-Tidwell testi kullanılmaktadır (Bircan, 2004; Alpay, 2017). Box -Tidwell testinin yanında saçılım grafiği ile de lojit ile sayısal değişken arasındaki doğrusal ilişki incelenebilmektedir.

### *2.2.5.2. Niteliksel (Kategorik) değişkenlerin Modele alınması*

Niteliksel değişkenlerin lojistik regresyona dahil edilirken dikkat edilmesi gereken hususlar mevcuttur. İki kategorili (binary)bağımsız değişkenler tıpkı niceliksel değişkenler gibi analize dahil edilir, ancak kodlama yapılırken yorumlamayı kolaylaştırmak adına etken maruziyeti olan ya da riskli görülen grubun büyük kodun verilmesidir. Çalışmamızdaki verilerden örnek verilecek olursa, sigara içenler 1 içmeyenler 0 olarak kodlamamız yorumlama kolaylığı sağlayacaktır. Çünkü katsayı kestiriminde bahsettiğimiz gibi odds oranın (olma olasılığı /olmama olasılığının) doğal logaritması bize lojit deki değişim miktarını vererek risk olan grup hakkında yorum yapmamızı sağlayacaktır.

Niteliksel değişkenler sırasız (nominal) veri türünde olduğu zaman modele dahil edilirken referans kategorilerin belirlenmesi için çeşitli yöntemler mevcuttur (SPSS, 2021);

**İndicator:** 0 değerini alan kategorinin referans kategorisi olması istendiği durumlarda kullanılır.

**Simple/Last:** Referans kategorinin, en son kategori veya kodlanan en büyük kategori olması istendiği durumlarda kullanılır.

**Simple/First:** Bağımsız değişkenin kodlamadaki en küçük değeri veya ilk kategorisi referans olması istenilen durumlarda kullanılır.

**Difference:** Bağımsız değişkenin her bir kategorisini kendinden önceki kategorilerinin etkilerinin ortalamasıyla karşılaştırılması istenen durumlarda kullanılır.

**Helmert:** Bağımsız değişkenin her bir kategorisini kendinden sonraki kategorilerinin etkilerinin ortalamasıyla karşılaştırılması istenen durumlarda kullanılır.

**Repeated:** İlgili değişkenin son kategorisi hariç diğer her kategorisinin bir önceki kategoriyle karşılaştırılması istenilen durumlarda kullanılır.

**Deviation:** İlgili değişkenin her bir kategorisinin etkisinin, tüm kategorilerinin ortalama etkisiyle karşılaştırılması istenilen durumlarda kullanılır.

**Polynomial:** Bağımsız değişkenin kategorileri arası etkileşimlerin doğrusallığının incelenmesi istenildiği durumlarda kullanılır.

Bağımsız değişkenlerin sıralanabilir (ordinal) nitel veri türünde olduğu durumlarda modele eklenirken, sırasız (nominal) nitel veri türünde olan prosedürler uygulanabilmektedir.

Bunun yanında sırasız (nominal) nitel veri türünde olduğu gibi referans kategoriler belirlenip de modele dahil edilebilirler. Sıralanabilen (ordinal) nitel veri türünde kategoriler isimlendirmeler veya sahip oldukları düzeylere göre kodlandığından dolayı niceliksel (sayısal) verilerde olduğu gibi doğrudan modele eklenebilir. Modele doğrudan ya da referans kategori seçip eklenmesi durumları bakımında sıralanabilir (ordinal) bağımsız değişkenler için bir şart bulunmamaktadır. Ancak kategori sayısı az olduğu durumlarda referans kategori seçimi yaparak modele eklenmesi tavsiye edilmektedir (Alpay, 2017).

### 2.2.6. Model Uyum İyiliği

Uyum iyiliği testleri genellikle, uygun modelin gözlemlenen verilerden genel olarak sapmasını değerlendiren genel testlerdir (Archer ve Lemeshow, 2006). Genel olarak uyum iyiliği testlerinin amacı, uygun modelin verilerde gözlemlenen sonucun ne oranda açıklayıp tanımlamadığını belirlemektir (Hosmer ve Lemeshow 2000). Model uyum testleri sonucunda verinin modele uygunluğu hakkında bilgi sahibi olmamızda sağlanır. Pearson-ki kare, sapma (deviance) istatistikleri, Hosmer-Lemeshow testi model uyum iyiliğini belirlemede sıklıkla kullanılan testlerdir (Dolgun ve Saraçbaş, 2015).

#### *2.2.6.1. Pearson-ki kare ve Sapma (deviance) istatistikleri*

Sapma (deviance) ve Pearson Khi-kare uyum istatistikleri lojistik regresyonda en yaygın kullanılan uyum istatistikleridir (Pregion, 1981). Pearson'ın ki-kare testi, gözlemlenen ve beklenen gözlemler arasındaki kare farklarının toplamını, standart hatasına bölen ve kovaryant deseni başına inceleyen böyle bir uyum iyiliği testidir (Archer ve Lemeshow, 2006). Pearson ve Sapma istatistikleri hesaplama yaparken artıkların toplamını kullanarak test sonuçlarını elde ederler (Alpay,2017).

$$\chi^2 = \sum_{i=1}^{j} r_i^2 \qquad (2\text{-}10)$$

$$D = \sum_{i=1}^{j} d_i^2 \qquad (2\text{-}11)$$

Eşitlik 2.10 da Pearson Ki-Kare istatistiği, 2-11 de ise Sapma (deviance) istatistiği verilmiştir. Analizlerin sonucunda $H_0$'ın yani verinin modele uygunluğu p değerine bakarak yorumlanır.

#### *2.2.6.2. Hosmer – Lemeshow Testi*

1980 ve 1982'de David W. Hosmer ile Stanley L. Lemeshow in ortak yazdıklardı çalışmalarda uyum iyiliğini tespit etmek için geliştirdikleri testtir (Archer ve Lemeshow, 2006). Hosmer-Lemeshow adıyla modelin veriye uyumluluğunu denetlemek için kullanılır.

Sonuçlarının kolay yorumlanması ve istatistik yazılımlarında tümünde bulunmasından dolayı uyum iyiliğini belirlemek için sıklıkla kullanılan bir testtir (Fagerland ve Hosmer, 2016).

Test uygulanırken grup sayısını belirlemek çok önemlidir. İkili lojistik regresyon analizinde test istatistiği yapılırken genelde 10'lu gruplama kullanılmaktadır. Aşağıda verdiğimiz eşitlik 2.12 test istatistiğinin hesaplama denklemidir.

**Tablo 2-1:** Tahmin edilen frekanslar ( $E_{kj}$) ve gözlemlerin kontenjans tablosu ($O_{kj}$).

| Grup | Y=0 | | Y=1 | | …….. | Y= c-1 | |
|---|---|---|---|---|---|---|---|
| 1 | $O_{10}$ | $E_{10}$ | $O_{11}$ | $E_{11}$ | …… | $O_{1,c-1}$ | $E_{1,c-1}$ |
| 2 | $O_{20}$ | $E_{20}$ | $O_{21}$ | $E_{21}$ | …… | $O_{2,c-1}$ | $E_{2,c-1}$ |
| . | | | | | …… | | |
| . | | | | | …… | | |
| . | | | | | …… | | |
| g | $0_{g0}$ | $E_{g0}$ | $O_{g1}$ | $E_{g1}$ | …… | $O_{g,c-1}$ | Eg,c-1 |

$$\hat{C}_g = \sum_{k=0}^{1} \sum_{j=1}^{10} \frac{(O_{kj}-E_{kj})^2}{E_{kj}} \qquad (2\text{-}12)$$

Tablo 2.1 yardımıyla ve Eşitlik 2-13 ile, belirlenen bir kesim (sabit) noktası için Hosmer-Lemeshow test istatistiğini hesaplama denklemi verilmiştir.

$$\hat{H}_g = \sum_{k=0}^{g} \sum_{j=0}^{c-1} \frac{(O_{kj}-E_{kj})^2}{E_{kj}} \qquad (2\text{-}13)$$

### 2.2.7. Açıklayıcılık Kat Sayıları ($R^2$)

Açıklayıcılık kat sayıları doğrusal regresyonda bağımlı değişkendeki değişimin miktarının bağımsız değişkenlerin ne oranla açıkladığını gösteren bir kat sayıydı. Fakat LR de bu durumu EKK tekniği kullanılmadığından dolayı mümkün değildir. LR de modelin uyumu iyiliğini değerlendirmek amacıyla yalancı (pseudo) $R^2$ istatistikleri geliştirilmiştir. Yalancı (pseudo) $R^2$ değerleri bağımsız değişkenlerin, bağımlı değişkenleri yüzdelik olarak ne kadar açıkladığını gösterir (Karagöz, 2021). Yüzdeler açıklanan varyans oranını değil, boş bir modele göre oluşturulan modelin olabilirlik oranın temsil eder (Hemmert ve ark,2016). Hesaplanan yalancı $R^2$ değerleri küçük değerler alma eğiliminde olduklarından model uyumunu değerlendirmekten ziyade, farklı modellerin performanslarını incelemede kullanılmalıdırlar (Alpar, 2017). Bağımlı değişkendeki açıklama yüzdesini veren istatistiksel yazılımlarda sıklıkla kullanılan yalancı (pseudo) $R^2$ analizleri şunlardır; McFadden'in $R^2$, Cox ve Snell $R^2$, Nagelkerke $R^2$, (Hu ve ark, 2006).

#### *2.2.7.1. McFadden'in $R^2$*

1973 yılında McFadden tarafından açıklanmış Olabilirlik oran indekisi olarak da nitelendirilen yöntemdir. Bağımsız değişkenleri barındırmayan sadece sabitin olduğu model ile tüm değişkenleri içeren modelin karşılaştırmıştır (McFadden, 1973).

$$R^2_{McF} = 1 - \frac{lnL(M_\beta)}{lnL(M_0)} \quad (2\text{-}14)$$

Eşitlik 2-14'te McFadden $R^2$ hesaplama formülü verilmiştir. Burada $M_\beta$ tüm değişkenlerin içinde olduğu modelin olabilirlik oranı, $M_0$ sadece sabitin olduğu modelin olabilirlik oranı. McFadden $R^2$ değeri hiçbir zaman 1 değerini almaz ancak tüm eğimler 0 ise 0 değerini alır (Hu ve ark, 2006). McFadden $R^2$ değei 0,20–0,40 arasında güçlü bir açıklayıcılığı işaret eder (McFadden, 1979).

#### *2.2.7.2. Cox-Snell $R^2$*

1983 yılında G.S. Madala tarafından geliştirilen açıklayıcılık katsayısıdır. Lineer regresyondan üretilen bir testdir (Madala, 1983).

$$R^2_{CS} = 1 - \left\{\frac{lnL(M_\beta)}{lnL(M_0)}\right\}^{\frac{2}{n}} = 1 - e^{\left(-G^2/n\right)} \quad (2\text{-}15)$$

Eşitlik 2-15 de gösterilen formül ile hesaplanır. Cox-Snell $R^2$'nin sayım verileri için negatif binom regresyonu veya survival verileri için Weibull regresyonu gibi en çok olabilirlik yöntemi ile tahmin edilen diğer regresyon türlerine genişletilebilir (Allison, 2014). En küçük değeri 0 iken en büyük değeri 1'in altında 0.775 civarındadır.

#### *2.2.7.3. Nagelkerke $R^2$*

Cox-Snell $R^2$ değerinin üst sınır hesaplamasından kaynaklı olarak 1'in altında olması yorumlama güçlüklerine neden olmuştur. Katsayının sınırlarını 0-1 arasında tutmak üzere Cox-Snell $R^2$ hesaplamasına bir düzenleme eklenmiştir (Nagelkerke, 1991). Nicolaas Jan Dirk Nagelkerke bu düzeltmeyi uygularken Cragg ve Uhler' in (1970) yaptığı çalışmadan ilham aldığından Cragg ve Uhler Pseudo $R^2$ olarak da bilinmektedir.

$$R^2_{NAG} = \frac{R^2_{CS}}{R^2_{\text{ÜST}}} = \frac{1-\left\{\frac{L(M_0)}{L(M_\beta)}\right\}^{\frac{2}{n}}}{1-L(M_0)^{\frac{2}{n}}} \quad (2\text{-}16)$$

Eşitlik 2-16 da Nagelkerke $R^2$ değerinin hesaplama formülü verilmiştir.

### 2.2.8. Modelin Yeterliliği

LR'de açıklayıcılık katsayılarının incelenmesinin yanında, artıkların ve etkili gözlemlerin incelemesinde model yeterliliğini gösteren ölçütlerdir. (Alpar, 2017).

#### *2.2.8.1. Artıkların İncelenmesi*

Regresyon modellerin tümünden artıklar, gözlenen değerler ile regresyon denkleminden elde edilen değerler arasındaki fark olarak tanımlanır. Genelleştirilmiş regresyon modellerinde artıklar modelin uyumu hakkında bilgi veren ölçütlerdir. (Cordeiro, 2004). Artıklar, uygun olmayan ya da kestirimde normal etkiden daha fazla olan gözlemleri belirler (Menard, 2002). Böylelikle de veriler modelle uyumlumu ya da yeterli bir modell mi sorularına cevap bulunur. LR'de sıklıkla Pearson Artıkları, standartlaştırılmış Pearson artıkları ve sapma artıkları kullanılmaktadır (Alpar, 2017). Pearson artığı ve sapma artıklarının bileşenleridir (Pearson artıkları varyansın karekökü ile düzeltilmiş ham artıklardır (Hilbe, 2009).

**Pearson artığı($r_i$)**

$$r_i = \frac{y_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}} \qquad (2\text{-}17)$$

Pearson artığı(ri) eşitlik 2-17 de verilen formül ile elde edilir. İstatistiğin sonucunda değer -2 ile 2 aralığında değilse artık değer varlığı söz konusudur (Christensen, 1997).

**Standardize Pearson artığı ($r_{std}$)**

Yapılan çalışmalarda pearson artıklarının varyans değerlerindeki hatalarını düzeltmek için önerilmiş olan artık türüdür (Alpay, 2017). Eşitlik 2-14 yardımıyla hesaplanır.

$$\boldsymbol{r_{std}} = \frac{r_i}{\sqrt{1-h_{ii}}} \qquad (2\text{-}18)$$

**Sapma Artıkları ($d_j$)**

$$d_j = sign(y_i - \pi_i)\sqrt{-2[(-y_i ln(\pi_i) - (1-y) ln(1-\pi_i)]} \qquad (2\text{-}19)$$

Doğrusal regresyondaki hata kareler toplamı ile benzerdir. Sapma değeri en düşük olan model diğerlerine göre daha iyi model olarak kabul edilir (Işığıçok, 2003). Sapma artıklarının k parametre sayısı ve k – (n+1) serbestlik derecesi ile ki kare dağılması beklenmektedir (Albayrak, 2006).

#### *2.2.8.2. Etkili Gözlemlerin Belirlenmesi*

Gözlem uzaklığı (Leverage): gözlemin diğer gözlemlerden ne kadar uzakta olduğunu belirler. 2p/n değerinden büyük değerleri uzak değereler olarak kabul eder.

Cook Uzaklığı: Cook tarafından 1979 tarafında önerilmiştir. Belli bir gözlemin veri setinden çıkarılması durumunda geri kalan artıkların ne kadar değiştiğini ölçer (Alpar, 2017).

$$D_j = \frac{z_j^2 h_j}{1-h_j} \qquad (2\text{-}20)$$

Eşitlik 2.20 yardımıyla hesaplanır. Dj istatistiğine ilişkin kritik değer $F_{0.50,p,n-p}$ olarak verilmiştir. Dj > $F_{0.50,p,n-p}$ olması halinde etkili gözlem olduğu söylenir (Cook, 1977).

### 2.2.9. Katsayıların Yorumlanması

#### *2.2.9.1. Nicel değişkenlerin Yorumlanması*

İkili LR'de nicel (sayısal) değişkenlerin modele doğrudan dahil olduğu bilinmektedir. İkili LR denklem katsayılarının hesaplanması eşitlik 2-6 da gösterilmiştir. Eşitlikte yer alan $\beta_s$ değerleri denklem kat sayılarıdır ve bu katsayıların exponansiyeli alındığında $\beta_s$ kat sayılarının lojite ne kadar değişim yaptığı hesaplanmış olur. $e^{\beta_s}$ 'in değeri, nitel değişkendeki 1 birimlik artışın, nicel bağımsız değişkendeki risk artışının miktarına gösteren sayısal değeridir (Alpar,2017).

#### *2.2.9.2. Nitel Değişkenlerin Yorumlanması*

Nitel bağımsız değişkenlerin tahmin edilen katsayılarının yorumlanması, tahmin edicinin referans seviyesine göredir (Minitab, 2022). Nitel bağımsız değişkenin kategori sayısı iki olduğu durumlar hesaplama odds oranı üzerinden yapılır çünkü denklem katsayısının doğal logaritması odd oranına eşittir (Minitab,2022). Aşağıdaki tablodan hareketle önce $\beta_i$ katsayısını hesaplayacak olursak;

**Tablo 2-2:** Kontenjans tablosu

| | **Bağımlı değişken (1)** | **Bağımlı değişken (0)** |
|---|---|---|
| Bağımsız değişken (1) | A | b |
| Bağımsız değişken (0) | C | d |

Tablo 2-2'den odds oranı $\frac{ad}{bc}$ olarak hesaplanır ve $e^{\beta_i} = \frac{ad}{bc}$ elde edilir. Eğer güven aralıkları 1 değerini içermiyorsa; Odds oranı 1'den büyük olduğunda risk faktörü olduğu

belirtilerek yorum yapılır Odds oranı eğer 1'den küçükse de azaltıcı riski azaltıcı faktör olarak yorumlama yapılır (Özdamar, 2011).

Bağımsız nitel değişkenin kategori sayısı ikiden fazla olduğu durumlarda kukla (dummy) değişkenleri kullanılır, literatürde desen yada dizayn değişkenleri yöntemi olarak da geçmektedir. Bu yöntemde kategori sayısının 1 eksiği kadar dizayn değişkeni oluşturulur. Referans kategorisi olarak belirlenen kategori 0 olarak kodlanır. 5 kategoriye sahip bir nitel değişken için dizayn değişkeni aşağıdaki gibidir.

**Tablo 2-3.** Dizayn değişkeni

| | $D_1$ | $D_2$ | $D_3$ | $D_4$ |
|---|---|---|---|---|
| **1. Kategori** | 0 | 0 | 0 | 0 |
| **2. Kategori** | 1 | 0 | 0 | 0 |
| **3. Kategori** | 0 | 1 | 0 | 0 |
| **4. Kategori** | 0 | 0 | 1 | 0 |
| **5. Kategori** | 0 | 0 | 0 | 1 |

p terimli bir Modelde j'inci terim, t kategorisi olan nitel bağımsız değişken olsun. Bu terim için t-1 adet dizayn değişkeni oluşturulur ve dizayn değişkeni "$D_{ju}$" (u= 1, 2,…, t-1) olur.

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{u=1}^{t_{j-1}} \beta_{ju} D_{ju} + \beta_p x_p \qquad (2\text{-}21)$$

Eşitlik 2.21 de j'ninci değişkeni kesikli olan model için lojit şekli verilir. Bu şekilde referans kategoriye diğer kategoriler iki kategorili değişkenlerdeki gibi yorumlanır.

### 2.2.10. Uygun model seçimi

Veri setini en iyi açıklayan modelin seçiminde, Akaike'nin bilgi kriteri (AIC) ve Bayesian bilgi kriteri (BIC) en çok kullanılan model uyum kriterleridir (Wang ve Putterman, 1998). Akaike tarafından 1973 yılında geliştirilmiş bir yöntemdir ve 2-22 eşitliğindeki gibi hesaplanır. AIC Kullback-Leibler ayrışmasını temel alır. Bu ayrışım beklenen logartima olasılıkları arasındaki fark olarak iki dağılım arasındaki mesafeyi alan bir görüştür (Kullback ve Leibler, 1951). Bayesian Bilgi kriteri (BIC), AIC fikrinde ortaya çıkan problemleri çözümlemek için 1978 yılında Schwarz tarafından ortaya atılmıştır (Acquah, 2018).

$$AIC = -2\log(L) + 2p \qquad (2\text{-}22)$$

$$BIC = -2\log(L) + p\log(n) \qquad (2\text{-}23)$$

Eşitlik 2-22 AIC, eşitlik 2.23'te BIC' nin hesaplama formülleri verilmiştir. Burada, L: Model parametre tahinlerini maksimum yapan log olabilirlik olasılığıdır, p: Parametre sayısı n: Örnek büyüklüğü olmaktadır.

Başka bir yaklaşımda bağımsız değişkenler ile bağımlı değişkenin tek tek LR ile analiz edilmesidir (Alpay, 2017). Bunun yanında bağımlı değişken ile bağımsız değişken arasındaki ilişkilerin ve ortalamalar arasındaki farkların uygun testler ile tek tek incelenmesi görüşü de mevcuttur (Alpay,2017). Tek değişkenli analizler sonucunda anlamlılığı olmayan değişkenler modele dahil etmeden model oluşturulur.

Model seçiminde kullanılacak başka bir model seçim stratejisi ileriye ve geriye dönük seçimlerdir. İleriye yönelik seçimlerde sadece sabit ile kurulan boş model ile başlanılır. Başlangıçta oluşturulan modele en çok katkı sağlayan değişkenler eklenerek model oluşturulur (Agresti, 1996). Geriye dönük yöntemde ise tüm değişkenler modele dahil edilerek etkisi olmayan değişkenler aşama aşama modelden atılır. Değişken ekleme ve çıkarma yapılırken olabilirlik oranı ve Wald istatistiklerine göre işlemler yapılmaktadır.

## 2.3. Makine Öğrenmesi

Öğrenmek, bilgi edinmek, beceri kazanmak, bellemek anlamlarına gelmektedir. (Türk Dil Kurumu, 2022). İnsan düşünebilen bir varlıktır fakat doğuştan içgüdüleri oldukça azdır, çevreye uyum sağlayabilmek ve doyum alabilmek için yaşam boyu öğrenmek zorundadır (Selçuk, 1999). İnsani bir özellik olan öğrenmenin makinelere duyarlamak zor görünse de yapay zekâ, istatistik ve matematik alanlarının çalışmaları ile mümkün hale gelmiştir (Kartal,2015).

Yapay zekâ, öğrenme, problem çözme vb. durumlar karşısında insan zekasını taklit edebilen bilgisayar sistemleri olarak tanımlanabilir. Makine öğrenmesi ise yapay zekâ kapsamında ortaya çıkan bir yöntemdir. Yapay zekâ bu amaçla programlara yönelikken, makine öğrenmesi bunun için yazılan algoritmalardır (Nacar ve Erdebilli, 2021). Makine öğrenmesi, istatistik, olasılık vb. matematik alanları ile ilişkilidir.

Büyük ve karmaşık veriler ile veri analizlerinin klasik yöntemlerle yapılması hem zaman alıcı olmakta hem de hata payını arttırmaktadır. Geleneksel yazılım geliştiricilerde; verilerden kalıplar çıkarılmaya çalışılması, verilerden elde edilen sonuçların gerçekçiliğini düşürmektedir. Bu sebeple makine öğrenmesi ile kalıp algoritmalar oluşturmadan, geçmiş

verilerden hareketle ve değişen durumlara uyumlu olarak kendi algoritmasını geliştirmesini sağlayarak, verilerin analizinde daha hızlı ve güvenilir tahminlerde bulunulabilmesi sağlanabilmektedir (Alpaydin,2020). Bundan dolayı makine öğrenmesi tıp alanında, kanser tanı ve teşhislerinde de sıklıkla kullanılmaktadır (Cruz ve Wishart, 2006).

### 2.3.1. Öğrenme Türleri

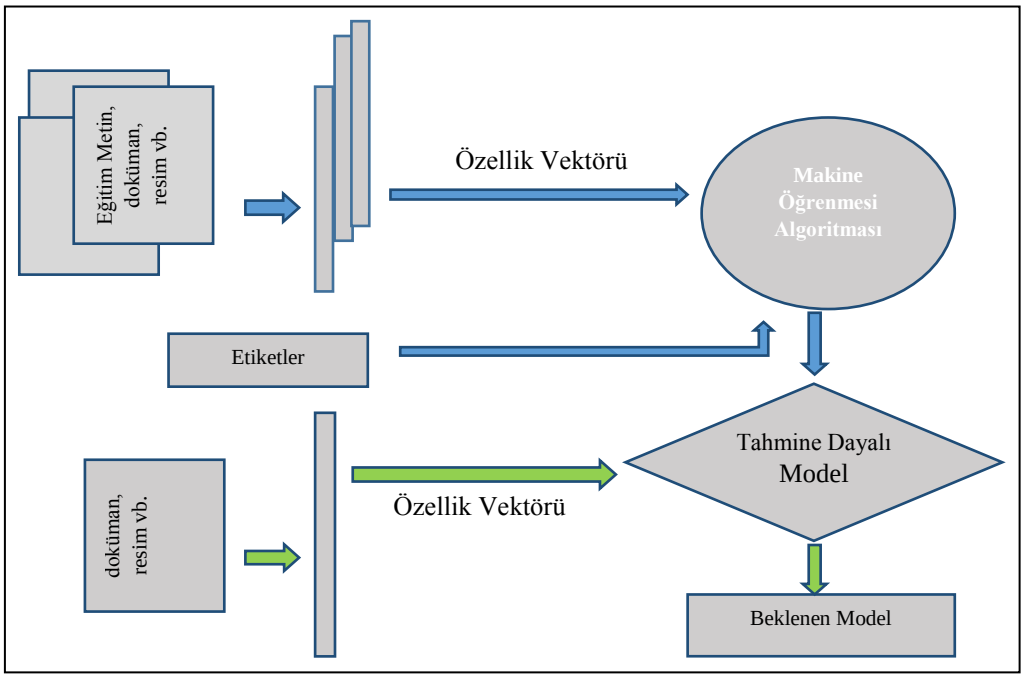
Makine öğrenmesi için öğrenme stratejileri vardır, öğrenme stratejilerini;

- Denetimli Öğrenme (Supervised Learning/ Predictive Model)
- Denetimsiz Öğrenme (Unsupervised Learning / Descriptive Models
- Güçlendirme /Takviyeli Öğrenme (Reinforcement Learning)

olarak sıralamıştır (Alpaydın, 2010).

#### *2.3.1.1. Denetimli Öğrenme (Supervised Learning/ Predictive Model)*

Bu modelde giriş verileri ve bu verilerin çıkış verileri (etiketleri) verilir. Veriler giriş ve çıkış arasında bir ilişki oluşturması için verilir. Makine öğrenmesinde giriş verisi bağımsız değişkenleri, çıkış verisi bağımlı değişkeni temsil etmektedir. Temel amaç sonuçları bilinen veri setinden yapılan sınıflandırmadan hareketle sonuçları bilinmeyen veri setine dair etkili tahminler yapabilmektir (Aydın ve Özkul, 2015). Denetimli öğrenme Şekil 1'de ki gibi olmaktadır (Afrin ve Nahar, 2015).




**Şekil 2-1:** Denetimli Öğrenme Modeli Süreci

Bilgisayarlar karmaşık problemlerin çözümü için kullanıldığında probleme yönelik net çözüm tanımlanmamış olabilir veya pahalı çözümlemeler olabilmektedir. Örneğin protein türlerinin sıralaması vb. Bu tarz sıralamalarda girdi verisinden hareketle çıktı verisini hesaplayamadığından klasik yöntemler yetersiz kalabilmektedir. Bu sebeple bilgisayarlara girdi verilerinden elde edilen çıktı verileri tanımlanarak, geçmiş verilerden çıkarım yapabilmesini kısacası öğrenmeyi öğrenmesi sağlanabilir.

Denetimli öğrenme, girdi ve bu verilerin çıkış verileri (etiketleri) nin dış bir mekanizma (insan) ile yapılmasıdır. Böylece sistem girdi ve çıktı arasındaki ilişkiyi öğrenir. Dış bir mekanizma tarafından yapıldığından dolayı denetimli öğrenme ismini almaktadır (Sebe ve ark., 2005).

Girdi ve çıktı ilişkisini öğrenen bilgisayarın, giriş değerlerinden hareketle elde edilen çıkış verisi ile istenilen doğruluktaki çıkış verisi elde edilinceye kadar bilgisayarın öğrenmesine ve eğitim işlemine devam ettirilir. İstenilen doğrulukta çıkış verisi elde edildiğinde ise öğrenme ve eğitim işlemi tamamlanmış olur (Elmas, 2003).

Denetimli öğrenme üçe ayrılır:

- Sınıflandırma Öğrenmesi (Classification Learning)
- Tercihli Öğrenme (Preference Learning
- Fonksiyon Öğrenme (Function Learning)

Denetimli öğrenme Modellerinde en sık kullanılan algoritmaları şu şekilde sıralayabiliriz (Caruana ve Niculescu-Mizil, 2006):

- Destek Vektör Makinesi (Support Vector Machine, SVM)
- Yapay Sinir Ağları
- Basit Bayes
- Multinom Basit Bayes
- K-En Yakın Komşu (K-Nearest Neighbor, KNN)
- Rastgele Orman (Random Forest)
- Karar Ağaçları (Decision Trees)

***Sınıflandırma Öğrenmesi:***

Sınıflama denetimli öğrenme stratejileri içerisinde yer alır. Eski verilerden hareketle, yeni verinin hangi sınıftan olduğunun saptanmasıdır. Örneklerden çıkarım olarak da

adlandırılabilen sınıflandırmanın hedefi, kavram tanımı elde edildikten sonra, önceden algoritmaya tanımlanmamış yeni veriyi en az hata ile etiketleyerek sınıflandırıcıyı geliştirmektir (Çakır, 2005).

Makine öğrenmesinde, birçok sınıflama algoritması vardır. Tez çalışmasında ise makine öğrenmesindeki bu sınıflama algoritmalarından faydalanılacaktır;

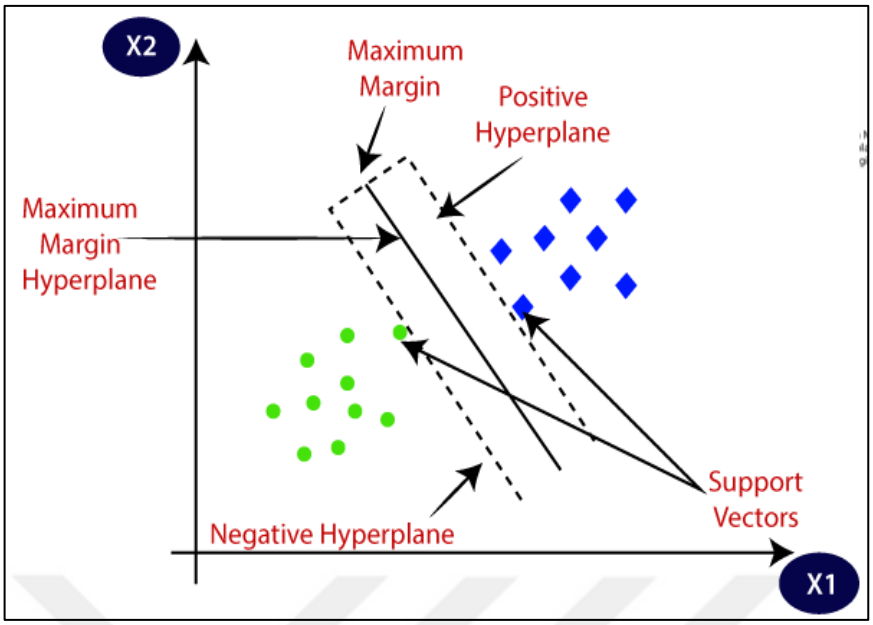
- Destek Vektör Makinesi (Support Vector Machine, SVM)
- K-En Yakın Komşu (K-Nearest Neighbor, KNN)
- Otomatik Ki-Kare Etkileşim Belirleme (CHAID)
- C5.0

***Destek Vektör Makinesi (Support Vector Machine, SVM)***

SVM, sınıflandırma ile regresyon problemleri için uygulanmaya başlanılmış bir yöntemdir. SVM, problemi karesel probleme çevirerek yerel çözümlere takılmadan çözümler. Bu özelliği ile diğer yöntemlerden avantajlı olmaktadır (Eray, 2008).

İlk çıktığında ikili sınıflandırma olarak kullanılmış sonradan ise çok sınıflı sınıflandırmalar yapabilecek şekilde geliştirilmiştir. D- öznitelikli veriler için SVM tarafından D- boyutlu sanal uzay meydana gelir (Hearst ve ark.,1998).

SVM, iki sınıfa da en uzak olan, sınıfları ayıran özel çizginin (hyperplane) bulunmasını amaçlar. Bu çizginin belirlenmesini sağlayan ve sınıfların birbirine yakın noktalarından sınırlar geçmektedir. Bu sınırlara destek vektörleri denilmektedir (Burges,1998).




**Şekil 2.2** Hyperplance ve Destek Vektörleri

SVM' de ki temel amaç Şekil 'de ki uzaklığı maksimize ederek, en uygun Maximum Margin Hyperplane ile gösterilen en uygun Hyperplance bularak, destek vektörlerinin ortalayarak yerleştirmektir.

SVM etiketli verilerin öznitelik değerleri uzaydaki boyutuyla eşlenerek dağıtılır. Sınıflandırıcı veri etiketine uygun, tüm uzayı ufak her bölgenin tek veri sınıfı örneğine sahip olacağı şekilde bölgelere ayıran hiperdüzlemler oluşturulmaya çalışılır. Bazı veriler öznitelik olarak bir sınıfta görülmesine rağmen başka sınıftan olabilir veya yakın değerlerde olan farklı sınıflardan verilerde olabilir. Bu gibi durumlarda sınıflayıcı, sınıfların veri örneklerini bölgede çoğunluk oluşturacak şekilde yapmaya çalışır (Meyer ve Wien, 2001).

***K-En Yakın Komşu (K-Nearest Neighbor, KNN)***

Öncelikle veri seti test ve eğitim olarak ikiye ayrılır. Test verisinin sınıflaması için, eğitim verisindeki gözlemlerindeki her biri için uzaklıkları hesaplanır. Sınıflandırmalardan elde edilen verilerden, sınıflandırılmak istenen verinin k tanesine yakınlığına bakılır. En fazla örneğe sahip sınıf, test verisinin sınıfı olur.

kNN uygulamasının uygulanabilmesi için temelde;

- Kayıtlı veriler
- Kayıtlı verilerdeki mesafeyi belirlemek için uzaklık hesaplaması
- k değeri, komşuların sayısı gerekmektedir.

Test verisini sınıflandırmak için temelde;

- Eğitim verilerine mesafe hesaplanır
- k tane en yakın komşu seçilir
- Test verilerinin sınıfını belirlemek için belirlenen k' ların sınıf etiketleri kullanılır

İki nokta arasındaki uzaklık hesaplaması için Öklidyen (http://www.ttbilgin.com/2018-2019-guz/veri-bilimine-giris/hafta6/siniflandirma2.pdf):

$$d(i, j)=\sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2\text{-}24)$$

K değeri önceden belirlenir, değerinin çok büyük olması durumunda benzemeyen noktaların sınıflanmasına, küçük olması ise aynı sınıflamadan olduğu halde farklı sınıfların açılmasına neden olabilmektedir (Silahtaroğlu, 2008). Bu sebeple 3,5 ve 7 gibi sayıların seçilmesi önerilmektedir.

KNN sınıfı bilinmeyen veri için en yakın k komşu belirlenerek, en çok tekrar eden sınıfın belirlenmesi mantığındadır fakat, seçilen sınıf *k* 'yı kısas aldığından uygun olmayabilir. Bu durumda sık tekrarlanan sınıfı seçmektense, ağırlıklı oylama yöntemi uygulanabilir (Özkan, 2008). Ağırlıklı oylama:

$$d(i, j)^1=\frac{1}{d(i,j)^2)} \quad (2\text{-}25)$$

d(i, j) Öklid uzaklığıdır. Her sınıf için uzaklıklar toplanarak, ağırlıklı oylama değeri hesaplanır.

***Otomatik Ki-Kare Etkileşim Belirleme (CHAID)***

CHAID metodu, karar ağacı algoritmalarındandır. CHAİD analizi ile birden çok ağaç oluşmaktadır. Bölümlendirme amacıyla kullanılan, bölünmelerin tespiti için ise ki- kare istatistiğinden yararlanan bir yöntemdir. Veriler homojen alt gruplara ayrılmaya çalışılır. Bu sebeple CHAID metodunda diğer değişkenleri ve verileri bağımlı değişkene uyumlu bir şekilde ele almak için alt gruplara ayırır. Tahmin edicilerden meydana gelen bu alt gruplar, sonraki analizlerde bağımlı değişkenin tahmin edilebilmesi için kullanılır. Karmaşık veri ilişkilerinde gizli ilişkilerin tespiti için kullanılır (Hoare, 2004).

CHAİD analizi, birden çok kategorili değişkenlerin olduğu büyük veri kümesini, emsal kategorileri bir araya getirip, önemli olan değişkenlere göre ayırmaktır. Her açıklayıcı değişken, kategorilerin anlamlı olacak şekilde bir araya getirilmesinden sonra, bağımlı

değişkene uygun kontenjans tabloları oluşturulur. P değerleri ile χ2 istatistik değerleri hesaplanır. Açıklayıcı değişkenler kıyaslanarak, en küçük Bonferronni p değerindeki kategoriye uygun olacak şekilde veriler alt gruplara ayrılır. CHAID analizi ki-kare istatistiklerini, Bonferroniyi ve kategori birleştirme algoritmalarıyla araştırmacının ağaç diyagramını, açıklayıcı değişkenleri ve bağımlı değişkenle ile olan ilişkilerini bulmasını sağlar (Pehlivan, 2006).

Bonferroni yaklaşımı her grubun ortalama vektörlerinin, genel ortalama vektörü ( $\bar{x}$ ) çıkarıldığında sıfır olup olmaması durumuna göre değerlendirme mantığındadır.

Genel ortalama vektörü $\bar{x}$ ve her grubun i. Ortalama vektörü $\bar{x}_{.g}$ şu şekildedir:

$$\bar{X}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix} \quad \bar{x}_1=\begin{bmatrix}\bar{x}_{11}\\ \bar{x}_{21}\\ \vdots\\ \bar{x}_{p1}\end{bmatrix} \quad \bar{x}_2=\begin{bmatrix}\bar{x}_{12}\\ \bar{x}_{22}\\ \vdots\\ \bar{x}_{p2}\end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots \bar{x}_g=\begin{bmatrix}\bar{x}_{1g}\\ \bar{x}_{2g}\\ \vdots\\ \bar{x}_{pg}\end{bmatrix} \qquad (2\text{-}26)$$

Her bir veri grubunun genel ortalama vektöründen farkı;

$$d_1=\bar{x}_1-\bar{x}=\begin{bmatrix}\bar{x}_{11}\\ \bar{x}_{21}\\ \vdots\\ \bar{x}_{p1}\end{bmatrix}-\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots d_g=\bar{x}_g-\bar{\bar{x}}=\begin{bmatrix}\bar{x}_{1g}\\ \bar{x}_{2g}\\ \vdots\\ \bar{x}_{pg}\end{bmatrix}-\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix} \qquad (2\text{-}27)$$

şeklinde hesaplanır.

k. grup ile j. grup i. değişken ortalamaları farklarına ait 1-α güven aralığı,

$$\left(d_{ki}-d_{ji}\right)=\left(\bar{x}_{ki}-\bar{x}_{ji}\right)\mp t\left(\frac{1}{pg(g-1)},N-g\right)\sqrt{\left(\frac{1}{n_k}\frac{1}{n_j}\right)\frac{w_{ii}}{N-g}} \qquad (2\text{-}28)$$

formülü ile hesaplanmaktadır. N=$n_1+n_2+\cdots+n_g$, g grup, p değişken sayılarını; $w_{ii}$ matrisin köşegen elemanlarını göstermektedir.

W gruplar için değişimi ifade etmekte olup;

$$w=\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)\left(x_{ij}-\bar{x}_i\right)^1 \qquad (2\text{-}29)$$

g gup sayısı, $n_i$ i. gruptaki sayıdır. Değişkenler için gruplar ikişerli ele alınır ve eşitlik (2-29) ile i için bulunan aralıkta sıfır değeri olup olmadığı kontrol edilir. Eğer aralıkta sıfır varsa incelenen gruplar arasında manidar farklılık olmadığı, sıfır değeri yer almıyorsa manidar farklılık olduğu çıkarımı yapılabilir (Özdamar, 2004).

***C5.0***

C5.0, karar ağaçlarından parametrik olmayan bir sınıflama algoritmasıdır. Algoritma ikili veya daha fazla dallı ağaç ile düğümleri ayırmak için bilgi kazancı (Entropy) ve budama yöntemi binom güven sınırını kullanır (Nguyen, 2020).

X değişkeninde k sayıda olasılıklar $P1$, $P2$, $P3$,…$Pk$ olarak ifade edilir. X değişkeni ya da niteliği için şu şekilde hesaplama yapılır (Larose, 2005);

$$\mathrm{H(x)} = -\sum_{j=1}^{k} p_j \log_2(p_j) \qquad (2\text{-}30)$$

"Eğitim seti içerisinde yer alan X niteliğine bağlı olarak hedef niteliği T alt kümelerinin,$T_1$ $T_2$… $T_k$ olarak alt bölmelere ayrıldığı varsayılmaktadır. Her bir T'nin sınıfını belirlemek için gerekli olan bilgilerin ağırlıklı ortalaması entropileri ağırlıklandırılmış toplamları'' olmaktadır. Bilgilerin ağırlıklı ortalaması denklem eşitliğinde verilmiştir:

$$H_S(T) = \sum_{\mathrm{i}=1}^{k} p_{\mathrm{i}}\, H_S(T_{\mathrm{i}}) \qquad (2\text{-}31)$$

Bölme için bilgi kazancı bulunmalıdır. Böylece ''C5.0 algoritması her bir karar verme düğümünde en büyük bilgi kazancına sahip olan ayırma ölçütünü belirleyerek optimal ayırma işlemini gerçekleştirmektedir''. Bilgi kazancı şu şekilde formüle edilir (Larose, 2005):

Bilgi kazancı (Information Gain-S) = $H(T)$ - $H_S(T)$ İle hesaplanır.

***2.3.1.2. Denetimsiz Öğrenme (Unsupervised Learning / Descriptive Models)***

Bu modelde giriş verisi verilir, çıktı verisi verilmeden veriler arasında ilişki bulunmak istenmektedir. Bu model çoğunlukla kümeleme, olasılık kestirimi ve boyut indirgemede kullanılmaktadır. Algoritmalar: K-Ortalama/Kümeleme Algoritması (Goodfellow ve ark., 2016). Girdilerin belli olduğu, çıktıların belli olmadığı sistemin öğrenme performansını ölçmek zordur (Mohri ve ark, 2012).

Denetimsiz öğrenmede denetimli öğrenmedeki gibi dış bir mekanizma yoktur. Sadece girdi verileri vardır ve sistem doğru çıktı verilerinin ne olduğunu bilmeden girdi- çıktı verileri arasındaki ilişkiyi bulmaya çalışır.

Birçok araştırmacı denetimsiz öğrenmeyi, yapay zekânın insan beyni gibi geliştirilebilmesini sağlayabilecek model olarak görmektedir (Gürsakal, 2021).

Denetimsiz öğrenme modellerinde en sık kullanılan algoritmaları şu şekilde sıralayabiliriz (Géron, 2019):

- Kümeleme (Clustering)

  - K-Ortalamalar (K-Means) Yoğunluğa Dayalı Uzamsal Kümeleme (Density-Based Spatial Clustering-DBSCAN)
  - Hiyerarşik Kümeleme Analizi (Hierarchical Cluster Analysis)
- Anomali ve Yenilik Tespiti (Anomaly detection and novelty detection)
  - Bir Sınıflı SVM (One Class SVM)
  - İzolasyon Ormanları (Isolation Forest)
- Görselleme ve Boyut İndirgeme (Visualization and dimensionality reduction)
  - Temel Bileşen Analizleri (Principal Component Analysis-PCA)
  - Çekirdek Temel Bileşen Analizleri (Kernel PCA)
  - Yerel Doğrusal Gömme (Locally Linear Embedding-LLE)
  - t-dağılımlı Stokastik Komşu Gömme (t-distributed Stochastic Neighbor Embedding)
- İlişkilendirme Kuralı Öğrenimi (Association rule learning)
  - Apriori
  - Eclat

#### *2.3.1.3. Güçlendirme /Takviyeli Öğrenme (Reinforcement Learning)*

Bu modelde etiketlenmemiş veri, etiketlenmiş veriden fazladır. Bu sebeple denetimli ve denetimsiz öğrenme stratejileri yetersiz olabilmektedir. Model dinamikliği algılayarak, hedefe doğru kararlar ile ulaşmaya çalışır. Sisteme ödül ve ceza verilir. Böylece sınanan durumların doğru veya yanlış olduğu saptanır (Alpaydın, 2010). Algoritmalar: Markov Karar Süreci.

Destekleyici öğrenmede denilen model ödülü arttırmak için ne yapılması gerektiğinin nasıl ilişkilenmesi gerektiğinin makineye öğretilmesidir (Herbrich, 2002). Bu modelde denetimli öğrenmedeki gibi bir denetçi vardır, fakat denetçi girdi için çıktı verisini girmez. Bunun yerine sistemin girdiye karşılık çıktı verisi oluşturmasını ister ve çıktının doğruluğuna göre sinyal verir (Öztemel, 2012). Ödülü arttırmak için daha önceden uygulanmış algoritmadan durum uzayını bulmalı ve ödül almış olan eylemleri bulmalıdır (Herbrich, 2002).

# 3. YÖNTEM

Çalışma prostat kanserine etki eden risk faktörlerini Lojistik Regresyon analizi ve sınıflayıcı makine öğrenme yöntemleri ile belirlemeyi amaçlar. Kullanılacak olan sınıflayıcı makine öğrenme algoritmaları; SVM, KNN, CHAİD ve C5.0'dır. SVM, KNN, CHAİD, C5.0 ve Lojistik regresyon analizinin prostat kanserini tespit etmek için kullandıkları risk faktörleriyle oluşturulan modellerden elde edecekleri; doğru sınıflama ve gain değerleri ile yöntemler karşılaştırılacaktır.

Verilerin analizi IBM' firmasının geliştirdiği iki farklı program kullanarak gerçekleştirilecektir. Sınıflayıcı makine öğrenmesi algoritmaları IBM SPSS Modeller programı ile Lojistik regresyon ve diğer tanımlayıcı analizler SPSS 25 programında gerçekleştirilecektir.

T.C. Sağlık Bilimleri Üniversitesi Hamidiye Bilimsel Araştırmalar Etik Kurulundan, 19.03.2021 tarihli 21/125 kayıt numarası ile alınan izin sonucunda veriler toplanmaya başlanmıştır. Veriler Ek-1 de yer alan bilgi toplama formu ile toplanmıştır. Veri toplama formu 20 sorudan oluşmaktadır. Sorular prostat kanserinde kabul gören risk faktörleri ve uzman görüşleri doğrultusunda hazırlanmıştır. Ek-1'de bilgi toplama formu tüm maddeleriyle verilmiştir. Haziran 2021–2022 Nisan tarihleri arasında, anketler bireylere ile yüz yüze görüşmeler sonucunda doldurulmuştur.

Lojistik regresyon analizinde örneklem büyüklüğü minimum bağımsız değişken sayısının 10 katı olmalıdır (Alpar, 2017; Akgül ve Çevik, 2003). Örneklem büyüklüğü için belirtilen bu görüş ile 501 kişi çalışmaya dahil edilmiştir. Toplamda 546 birey ile yapılan görüşmelerde 45 bireyden eksi bilgiler alınmasından dolayı çalışmaya dahil edilmemiştir. 248 prostat kanseri tanısı almış (hasta), 253 prostat kanseri tanısı almamış (sağlam) bireyler ile gerçekleştirilmiştir.

Lojistik regresyon analizinden elde edilen modellerle hastalığa ait risk faktörleri belirlenecektir. Daha sonra denetimli makine öğrenme yöntemleri SVM, KNN, CHAİD ve C5.0 prostat kanseri verileri ile modellere birer birer öğretilecektir. Öğrenme işlemlerinden sonra modeller tekrar analize alınıp test edilecektir. Böylelikle makine öğrenme yöntemleri ile de hastalığa ait risk faktörlerini belirlenecektir. En son olarak modellerin doğru tahminleme ve gain değerleri karşılaştırılacaktır.

# 4. BULGULAR

**Tablo 4-1.** Katılımcılara ait demografik bilgiler.

| Değişkenler | | Grup n (%) | |
|---|---|---|---|
| | | **Hasta** | **Sağlam** |
| Medeni Durum | Bekar | 50(%20) | 148(%59) |
| | Evli | 176(%71) | 79(%31) |
| | Diğer | 20(%9) | 26(%10) |
| Sigara kullanımı | Hayır | 32(%12,9) | 189(%74,7) |
| | Evet | 216(%87,1) | 64(%25,3) |
| Eğitim düzeyi | Okur-yazar | 13(%5,24) | 5(%1,97) |
| | İlkokul | 78(%31,4) | 29(%11,4) |
| | Ortaokul | 51(%20,5) | 47(%18,5) |
| | Lise | 81(%32,6) | 80(%31,6) |
| | Üniversite | 25(%10,0) | 92(%36,3) |
| Alkol kullanımı | Evet | 86(%34,7) | 67(%26,5) |
| | Hayır | 162(%65,3) | 186(%73,5) |
| Meslek | Hayır | 22(%8,9) | 191(%75,5) |
| | İşçi | 23(%9,27) | 16(%6,32) |
| | Serbest meslek | 53(%21,3) | 42(%16,6) |
| | Öğrenci | 0(%0) | 37(%14,6) |
| | Akademik personel | 14(%5,64) | 6(%2,37) |
| | Devlet memuru | 37(%14,9) | 45(%17,7) |
| | Çalışmıyor | 18(%7,25) | 7(%2,76) |
| | Emekli | 50(%20,1) | 36(%14,2) |
| | Sağlık personeli | 30(%12,0) | 28(%11,0) |
| | Öğretmen | 23(%9,27) | 36(%14,2) |
| Ailenizde kanser hastası olan kişi varlığı | Hayır | 135(%54,4) | 208(%82,2) |
| | Evet | 113(%45,6) | 45(%17,8) |
| Doğduğunuz şehir | İl | 136(%54) | 129(%51) |
| | İlçe | 112(%46) | 124(%49) |
| Yaşadığınız şehir | İl | 173(%70) | 187(%74) |
| | İlçe | 75(%30) | 66(%26) |

Hasta: Prostat kanseri tanısı almış bireyler.
Sağlam: Prostat kanseri tanısı almamış bireyler.

**Tablo 4-2.** Katılımcılara ait demografik bilgiler (devamı).

| Değişkenler | | Grup n (%) | |
|---|---|---|---|
| | | Hasta | Sağlam |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Anne] | Hayır | 215(%86,7) | 240(%94,9) |
| | Evet | 33(%13,3) | 13(%5,1) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Baba] | Hayır | 220(%88,7) | 241(%95,3) |
| | Evet | 28(%11,3) | 12(%4,7) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Kardeş] | Hayır | 227(%91,5) | 250(%98,8) |
| | Evet | 21(%8,5) | 3(%1,2) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Dede] | Hayır | 233(%94) | 247(%97,6) |
| | Evet | 15(%6) | 6(%2,4) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Nine] | Hayır | 246(%99,2) | 247(%97,6) |
| | Evet | 2(%0,8) | 6(%2,4) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Dayı] | Hayır | 231(%93,1) | 248(%98) |
| | Evet | 17(%6,9) | 5(%2) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Amca] | Hayır | 230(%92,7) | 243(%96) |
| | Evet | 18(%7,3) | 10(%4) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Teyze] | Hayır | 244(%98,4) | 246(%97,2) |
| | Evet | 4(%1,6) | 7(%2,8) |
| Kanser hastası olan bireyle akrabalık düzeyiniz [Hala] | Hayır | 244(%98,4) | 245(%96,8) |
| | Evet | 4(%1,6) | 8(%3,2) |
| Ağırlık olarak Beslenme şekli | Kırmızı et ağırlıklı | 103(%41,5) | 87(%34,4) |
| | Beyaz et ağırlıklı | 36(%14,5) | 46(%18,2) |
| | Sebze ağırlıklı | 37(%14,9) | 61(%24,1) |
| | Bakliyat ve tahıl ağırlıklı | 26(%10,5) | 59(%23,3) |
| | Hepsi | 46(%18,5) | 0(%0) |
| Günlük ortalama su tüketimi | 1 litre | 116(%46,8) | 126(%50,2) |
| | Ortalama 2.5 lt | 98(%39,5) | 94(%37,5) |
| | 2.5 lt ve üzeri | 34(%13,7) | 31(%12,4) |
| İdrarı yapma pozisyonu | Genelde ayakta | 95(%38,3) | 62(%24,6) |
| | Genelde oturarak | 67(%27) | 103(%40,9) |
| | Hep oturarak | 30(%12,1) | 10(%4) |
| | Hep ayakta | 56(%22,6) | 77(%30,6) |
| Kronik hastalık varlığı | Var | 122(%49,2) | 42(%16,7) |
| | Yok | 126(%50,8) | 209(%83,3) |
| Ne sıklıkla idrara çıkarsınız | Günde 5 kez ve aşağısı | 25(%10,1) | 173(%68,4) |
| | Günde 5-10 kez | 61(%24,6) | 71(%28,1) |
| | Günde 10 ve üzeri | 162(%65,3) | 9(%3,6) |
| Genel anlamda gün içinde yaşam şekli | Hareketsiz | 158(%63,7) | 65(%25,8) |
| | Spor yaparım | 21(%8,5) | 26(%10,3) |
| | Spor yapmam ama hareketliyim | 69(%27,8) | 161(%63,9) |
| Yiyeceklerde kullanılan yağ | Bitkisel yağ | 89(%35,9) | 86(%34,4) |
| | Hayvansal yağ | 124(%50) | 164(%65,6) |
| | Her ikisi | 35(%14,1) | 0(%0) |

Hasta: Prostat kanseri tanısı almış bireyler.
Sağlam: Prostat kanseri tanısı almamış bireyler.

**Tablo 4-3.** Katılımcıların şikayetleri.

| **Değişkenler** | | **Grup n (%)** | |
|---|---|---|---|
| | | **Hasta** | **Sağlam** |
| [İdrar yapma güçlüğü] | Hayır | 118(%47,6) | 237(%93,7) |
| | Evet | 130(%52,4) | 15(%7,7) |
| [İdrar akışında kuvvet azalması] | Hayır | 158(%63,7) | 222(%87,7) |
| | Evet | 90(%36,3) | 31(%12,3) |
| [Menide ya da idrarda kan görülmesi] | Hayır | 171(%69) | 243(%96) |
| | Evet | 77(%31) | 10(%4) |
| [Sertleşme bozukluğu] | Hayır | 215(%86,7) | 232(%91,7) |
| | Evet | 33(%13,3) | 21(%8,3) |
| [Boşalma esnasında ağrı ] | Hayır | 237(%95,6) | 249(%98,4) |
| | Evet | 11(%4,4) | 4(%1,6) |
| [Kasık bölgesinde rahatsızlık hissi] | Hayır | 186(%75) | 240(%94,9) |
| | Evet | 62(%25) | 13(%5,1) |
| [Kemik ağrıları] | Hayır | 205(%82,7) | 243(%96) |
| | Evet | 43(%17,3) | 10(%4) |
| [Yukarıdakilerden birini yaşıyorum] | Hayır | 7(%2,8) | 42(%20) |
| | Evet | 241(%97,2) | 211(%80) |
| **Nicel değişkenler** | | **Ortalama ± standart sapma** | |
| | | **Hasta** | **Sağlam** |
| Yaş | | 72±8,74 | 46±9,92 |
| Boy | | 175±7,01 | 178±7,32 |

Hasta: Prostat kanseri tanısı almış bireyler.
Sağlam: Prostat kanseri tanısı almamış bireyler.

Prostat kanseri tanısı almış bireylerin yaş ortalaması 72±8,74 iken en düşük yaşa sahip 45 iken en yüksek 90 yaşında hasta olduğu görülmüştür. Prostat kanseri tanısı almamış bireylerin yaş ortalaması 46±9,92 iken en düşük yaşa sahip 19 iken en yüksek 78 yaşında bireyin olduğu görülmüştür.

Bağımsız değişkenlerin bağımlı değişken ile tek tek analiz edilmesi sonucunda aşağıdaki tablo 4-4'teki sonuçlar elde edilmiştir.

**Tablo 4-4.** Bağımsız değişkenlerin seçimi.

| | B | S.E. | Wald | Sig. | OR (95%Cl) |
|---|---|---|---|---|---|
| Yaşınız | 0,261 | 0,028 | 87,949 | **,001** | 1,298(1,229-1,371) |
| Boy | -1,096 | 0,013 | 18,04 | 0,272 | 0,334(0,924-0,971) |
| Sigara kullanıyor musunuz (1) | 2,992 | 0,238 | 157,662 | **,001** | 19,92(12,495-31,801) |
| Alkol (1) | -0,388 | 0,195 | 3,947 | 0,077 | 0,678(0,463-0,995) |
| Ailede kanser varlığı (1) | 1,353 | 0,208 | 42,289 | **,001** | 3,869(2,573-5,817) |
| [Anne] (1) | -0,842 | 0,341 | 4,349 | 0,132 | 0,430(1,453-5,525) |
| [Baba] (1) | -0,938 | 0,357 | 6,895 | 0,064 | 0,391(0,194-0,788) |
| [Kardeş] (1) | 0,042 | 0,624 | 7,714 | 0,057 | 1,042(2,269-26,191) |
| [Dede] (1) | 0,375 | 0,492 | 3,931 | 0,072 | 1,454(1,011-6,946) |
| [Nine] (1) | -1,095 | 0,821 | 1,775 | 0,183 | 0,334(0,067-1,674) |
| [Dayı] (1) | -0,295 | 0,517 | 5,275 | 0,092 | 0,744(1,325-10,053) |
| [Amca] (1) | 0,643 | 0,405 | 2,519 | 0,112 | 1,902(0,86-4,206) |
| [Teyze] (1) | -0,551 | 0,633 | 0,758 | 0,384 | 0,576(0,167-1,993) |
| [Hala] (1) | -0,689 | 0,619 | 1,239 | 0,266 | 0,502(0,149-1,689) |
| Ağırlık olarak Beslenme şekliniz | 0,072 | 0,065 | 1,228 | 0,268 | 1,074(0,946-1,221) |
| Günlük su tüketimi | 0,098 | 0,128 | 0,583 | 0,445 | 1,102(0,858-1,418) |
| İdrarınız hangi pozisyonda yaparsınız | -0,158 | 0,077 | 4,209 | 0,207 | 0,853(0,734-0,993) |
| Kronik bir hastalığınız var mı? | -1,572 | 0,211 | 2,278 | 0,113 | 0,207(0,137-0,314) |
| Ne sıklıkla idrara çıkarsınız | 2,305 | 0,182 | 161,139 | **,001** | 10,02(7,022-14,308) |
| Gün içindeki yaşam şekli | -0,867 | 0,103 | 70,964 | **,001** | 0,420(0,343-0,514) |
| Kullanılan yağ | 0,372 | 0,155 | 5,737 | 0,067 | 1,450(1,07-1,967) |
| [İdrar yapma güçlüğü] | 0,249 | 0,281 | 5,432 | 0,238 | 1,282(8,614-25,951) |
| [İdrar akışında kuvvet azalması] | 0,406 | 0,233 | 6,466 | 0,143 | 1,500(2,585-6,438) |
| [Menide ya da idrarda kan görülmesi] | 4,181 | 0,722 | 33,499 | **,001** | 65,43(15,884-269,647) |
| [Sertleşme bozukluğu] | 0,528 | 0,295 | 3,21 | 0,073 | 1,695(0,952-3,022) |
| [Boşalma esnasında ağrı] | 1,061 | 0,591 | 3,224 | 0,073 | 2,889(0,907-9,199) |
| [Kasık bölgesinde rahatsızlık hissi] | -2,843 | 0,32 | 4,183 | 0,107 | 0,058(3,285-11,529) |
| [Kemik ağrıları] | 0,178 | 0,364 | 7,057 | 0,091 | 1,194(2,499-10,396) |
| [Hayır hiçbirini yaşamıyorum] | -4,081 | 0,405 | 2,569 | 0,074 | 0,016(0,008-0,037) |

Lojistik regresyonda modeli belirlemek üzere yapılan analizler sonucunda Yaş, sigara, ailesinde kanser varlığı, idrara çıkma sıklığı, gün içindeki yaşam şekli ve menide kan görülmesi bağımsız değişkenleri anlamlı bulunmuştur (p<,05).

Bağımsız değişkenler ile iki kategorili lojistik regresyon analizi gerçekleştirilmiştir. Analiz sonucunda oluşan model aşağıda tablo 4-5'te verilmiştir.

**Tablo 4-5.** İkili Lojistik regresyon analizi sonuçları.

| | B | S.E. | Wald | Sig. | OR (95% CI) |
|---|---|---|---|---|---|
| Yaşınız | 0,098 | 0,012 | 70,142 | ,001 | 1,103(1,078-1,128) |
| Sigara (1) | 1,727 | 0,365 | 22,425 | ,001 | 5,624(2,752-11,494) |
| Ailede kanser varlığı (1) | 0,923 | 0,383 | 5,824 | ,016 | 2,517(1,189-5,329) |
| Ne sıklıkla idrara çıkarsınız | | | 8,898 | ,012 | |
| Ne sıklıkla idrara çıkarsınız (1) | 0,91 | 0,418 | 4,741 | ,029 | 2,484(1,095-5,637) |
| Ne sıklıkla idrara çıkarsınız (2) | 1,325 | 0,472 | 7,876 | ,005 | 3,763(1,491-9,496) |
| Gün içindeki yaşam şekli | | | 12,841 | ,002 | |
| Gün içindeki yaşam şekli (1) | 0,983 | 0,556 | 8,111 | ,004 | 2,672(1,638-14,487) |
| Gün içindeki yaşam şekli (2) | -0,381 | 0,387 | 0,973 | ,324 | 0,683(0,32-1,457) |
| Menide kan görünmesi | 2,436 | 0,724 | 11,31 | ,001 | 11,432(2,763-47,289) |
| Sabit | -3,826 | 0,701 | 29,771 | ,001 | |

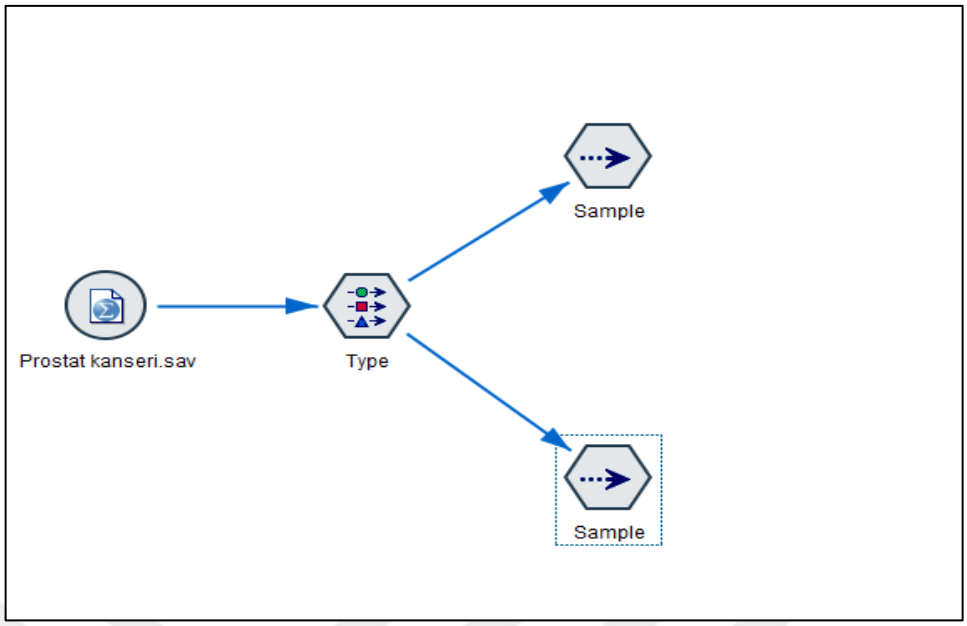
Çok değişkenli regresyon analizi sonucunda Hosmer - Lemeshow testi ($x^2$ =12,112 df = 8 ve p = ,146) sonucunda verinin modele uygun olduğu görülmüştür (p>,05). -2log likehoold =228,115, Cox-Snell $R^2$=,706 ve Nagelkerke $R^2$= ,808 değerleri elde edilmiştir. Cox-Snell $R^2$ değeri modelin %71 bağımlı değişkeni açıkladığı ve Nagelkerke $R^2$ değeri %81 oranda modelin yüksek açıklayıcılığı olduğunu göstermektedir.

Analizler sonucunda yaş değişkenin 1 birimlik artışının prostat kanserinde 1,103 kat risk faktörü olduğu tespit edilmiştir (p<,05). Sigara kullanımının 5,624 kat (p<,05), ailesinde kanser hastası olanların 2,517 kat (p<,05) ve menide kan gelme şikayetinin 11,432 kat (p<,05) risk faktörü olduğu tespit edilmiştir. Spor yapmayanların, spor yapanlara göre 2,672 kat risk altında olduğu tespit edilmiştir (p<,05). İdrara 10 kez ve üzeri çıkanların, en fazla 5 kez çıkanlara göre 3,763 kat ve 5- 10 kez arası çıkanların, en fazla 5 kez çıkanlara oranla 2,484 kat risk altında olduğu tespit edilmiştir.

**Tablo 4-6.** Sınıflandırma tablosu

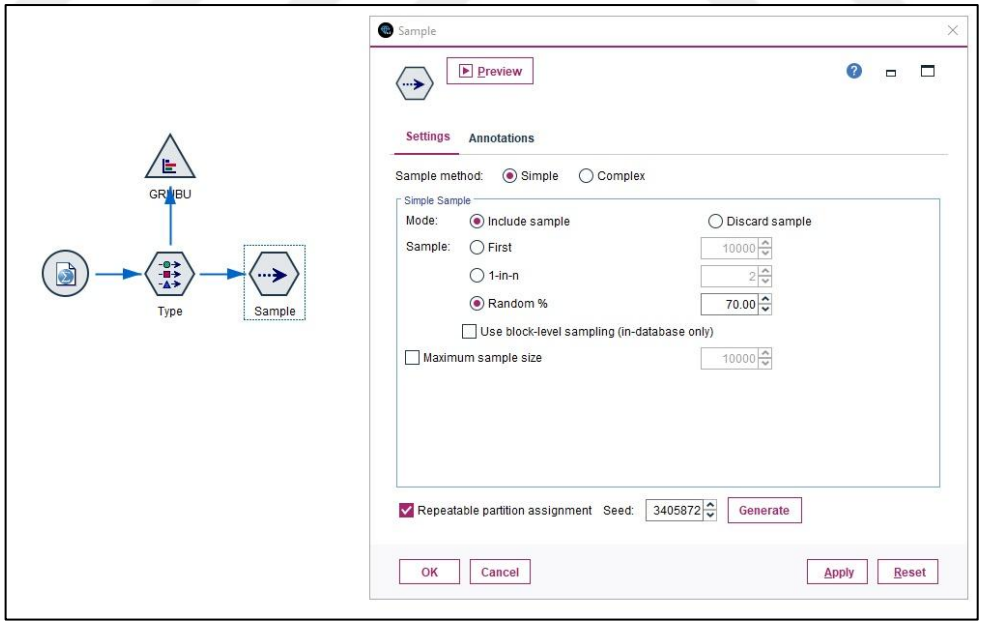
| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 233 | 20 | 92,1 |
| Prostat kanseri olmayan | 19 | 229 | 92,3 |
| | | | 92,2 |

Tablo 4-6'da oluşturulan modelin sınıflandırma tablosu verilmiştir. LR ile oluşturulan model prostat kanseri olmayan bireyleri %92,1 oranla doğru tahmin yapılmıştır. Prostat kanseri olan bireyler %92,3 oranla doğru tahmin yapılırken modelin toplamda doğru tahmin yüzdesi %92,2 olarak hesaplanmıştır.



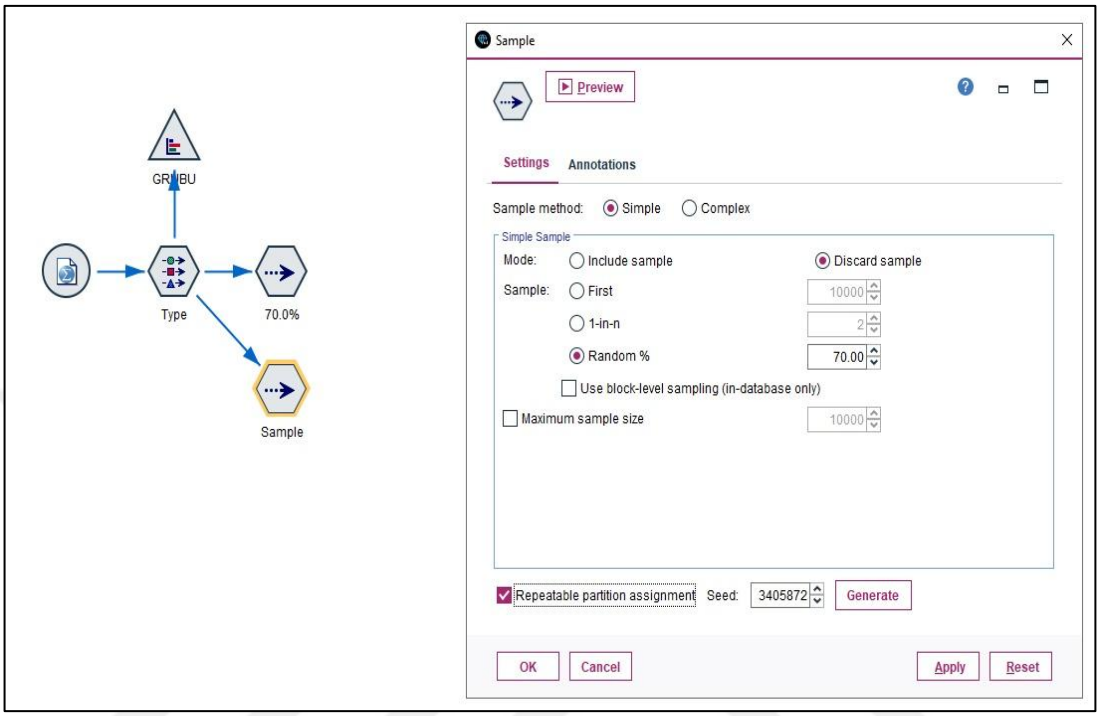

**Şekil 4-1:** Verinin programa tanıtılması

SPSS dosyası olarak hazırlanan veriler SPSS Modeller programına import edilir. Veri seti programa tanıtıldıktan sonra veri test ve eğitim olarak bölünür.



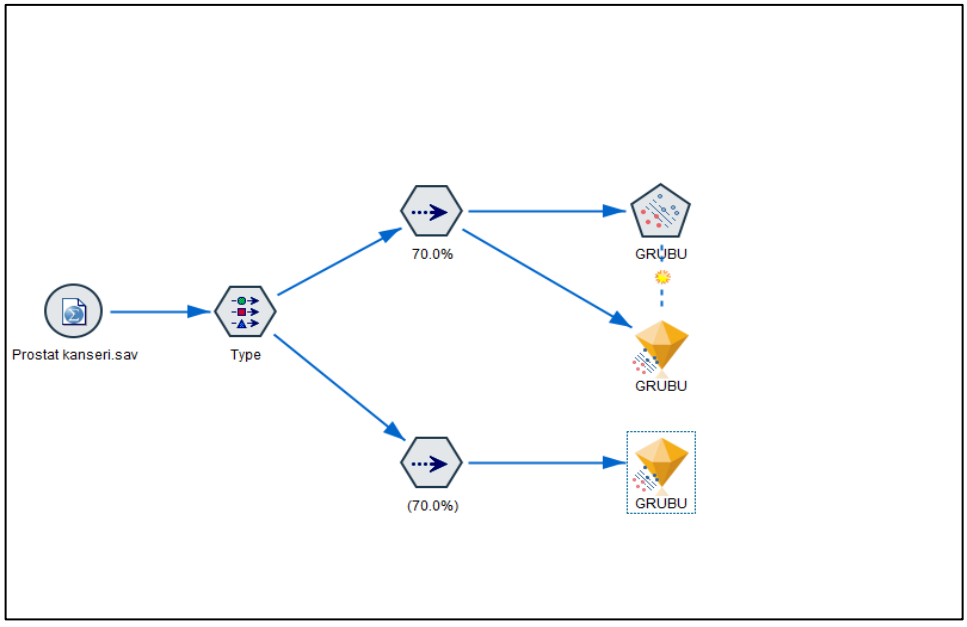

**Şekil 4-2:** Eğitim verisinin oluşturulması.

Eğitim veri seti oluşturulurken %70 ‘e %30 kuralı uygulanarak, tüm veri setinin %70’i eğitim veri setine bölünmüştür. Verinin kalan %30’luk kısmından %70’lik bir veri seti elde edilerek eğitim ve öğrenme verileri hazırlanır.



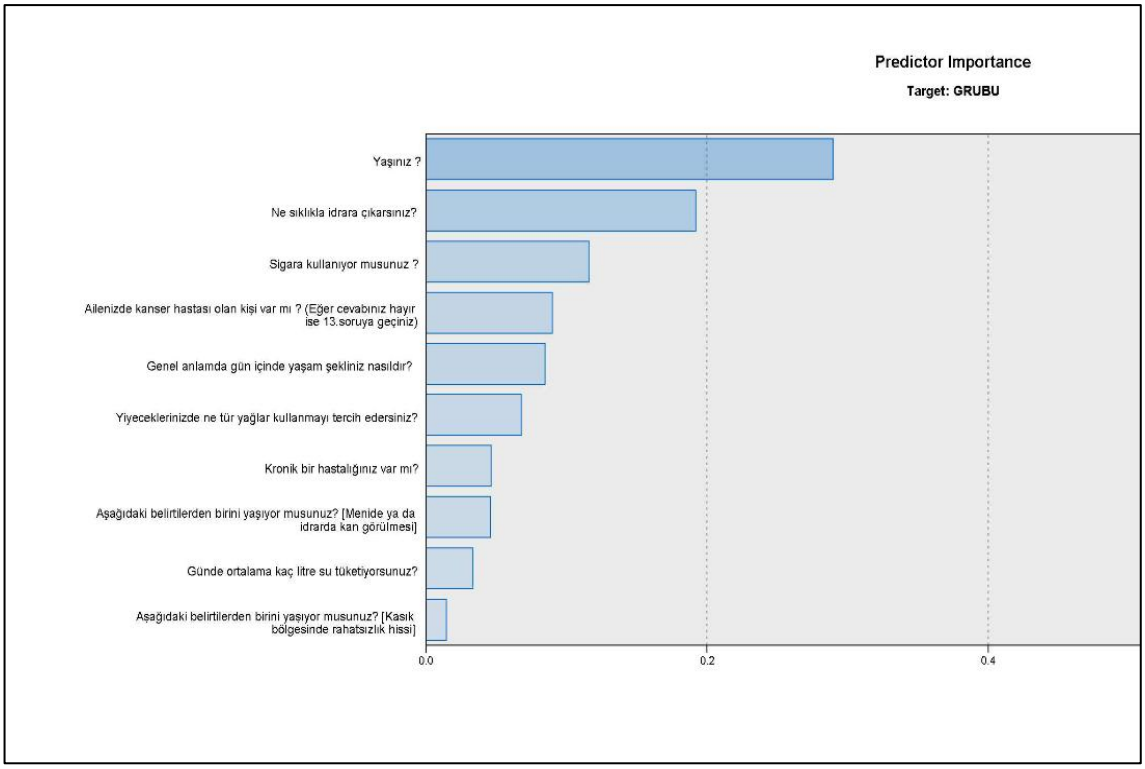

**Şekil 4-3:** Test verisinin oluşturulması

SVM, eğitim veri setinde analiz edilerek model oluşturulmuştur. Oluşturulan model daha sonra deneme veri setine uygulanarak tahmin değerleri elde edilmiştir.




**Şekil 4-4:** SVM eğitim ve deneme veri setine uygulanması.

Analiz sonucunda önem sırasına göre; yaş, idrara çıkma sıklığı, sigara kullanımı, ailede kanser hastalığı görülmesi, gün içindeki yaşam şekli, yiyeceklerde kullanılan yağ, kronik hastalık varlığı, menide veya idrarda kan görülmesi, günlük su tüketimi ve kasık bölgesindeki rahatsızlık hissi değişkenleri ile model oluşturulmuştur.




**Şekil 4-5:** SVM analizi ile oluşturulan modelin değişkenleri.

Prostat kanseri hastalarını sınıflamak için kurulan SVM modeline; yaş değişkeni 0,29, idrara çıkma sıklığı 0,19, sigara kullanımı 0,12, ailede kanser hastalığı görülme 0,09, gün içindeki yaşam şekli 0,08, yiyeceklerde kullanılan yağ 0,7, kronik hastalık varlığı 0,05, mesane veya idrarda kan görülme 0,05, günlük su tüketimi ve kasık bölgesindeki rahatsızlık hissi 0,01 oranda katkı sağlamıştır.

**Tablo 4-7:** SVM modelinin sınıflandırma tablosu

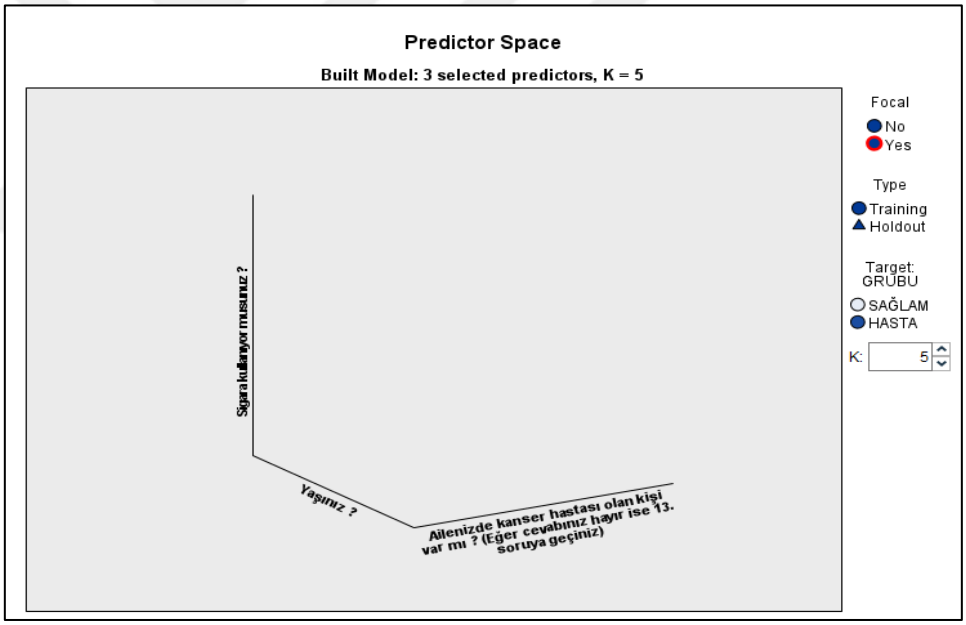
| **Gözlenen** | **Tahmin edilen** | | **Doğru Sınıflama Yüzdesi** |
|---|---|---|---|
| | **Prostat kanseri olmayan** | **Prostat kanseri olan** | |
| Prostat kanseri olan | 5 | 171 | 97,1 |
| Prostat kanseri olmayan | 184 | 2 | 98,92<br>98 |

SVM modelinin eğitim de prostat kanseri olanları doğru tahmin etme oranı %97,1, prostat kanseri olmayanları doğru tahmin etme oranı %98,92, testin toplam doğru tahmin oranı %98‘dir.

**Tablo 4-8:** SVM öğrenmiş modelin sınıflandırma tablosu

| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 4 | 68 | 94,4 |
| Prostat kanseri olmayan | 57 | 10 | 85 |
| | | | 89,92 |

SVM ile eğitilen modelin sonra prostat kanseri olanları doğru tahmin etme oranı %94,4, prostat kanseri olmayanları doğru tahmin etme oranı %85, testin toplam doğru tahmin oranı %89,92 ‘dir.




**Şekil 4-6:** KNN analizi ile oluşturulan modelin değişkenleri

KNN ile oluşturulan model de sigara kullanımı, yaş ve ailede kanser hastası olma değişkenleri anlamlı bulunarak model oluşturulmuştur. KNN modelinin yanlış sınıflandırma yüzdesi %10,8 olarak elde edilmiştir.

**Tablo 4-9:** KNN modelinin sınıflandırma tablosu

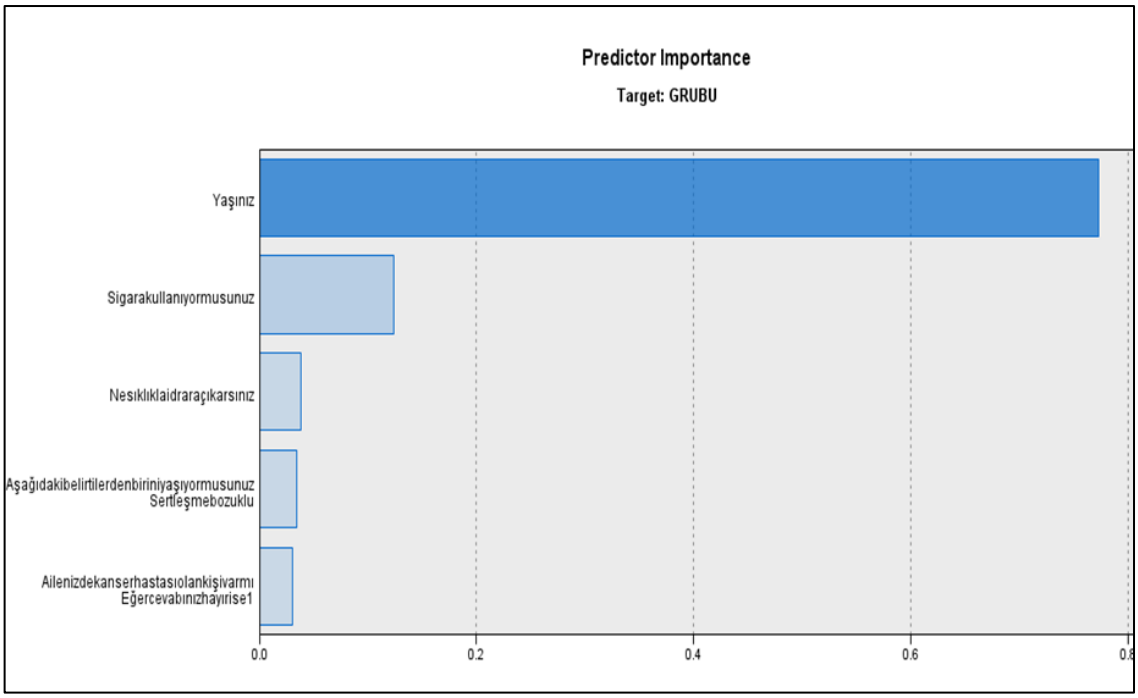
| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 16 | 160 | 90,9 |
| Prostat kanseri olmayan | 182 | 4 | 97,84<br>94,47 |

KNN ile eğitilen modelin prostat kanseri olanları doğru tahmin etme oranı %90,9, prostat kanseri olmayanları doğru tahmin etme oranı %97,84, testin toplam doğru tahmin oranı % 94,47‘dir.

**Tablo 4-10:** KNN ile eğitilmiş modelin sınıflandırma tablosu

| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 8 | 64 | 88,8 |
| Prostat kanseri olmayan | 59 | 8 | 88,0<br>88,48 |

KNN ile oluşturulan modelinin prostat kanseri olanları doğru tahmin etme oranı %88,8, prostat kanseri olmayanları doğru tahmin etme oranı %88, testin toplam doğru tahmin oranı % 88,48 ‘dir.




**Şekil 4-7:** CHAİD analizi ile oluşturulan modelin değişkenleri

CHAİD analizi sonucunda önem sırasına göre model; yaş, sigara kullanımı, idrara çıkma sıklığı, sertleşme bozukluğu yaşama ve ailede kanser hastası varlığı değişkenleri ile model oluşturulmuştur.

Prostat kanseri hastalarını sınıflamak için kurulan CHAİD modeline; yaş değişkeni 0,77, sigara kullanımı 0,12, idrara çıkma sıklığı değişkeni 0,04, sertleşme bozukluğu yaşama 0,03, ve ailede kanser hastalığının varlığı 0,03 oranda katkı sağlamıştır.

**Tablo 4-11:** CHAİD sınıflandırma tablosu

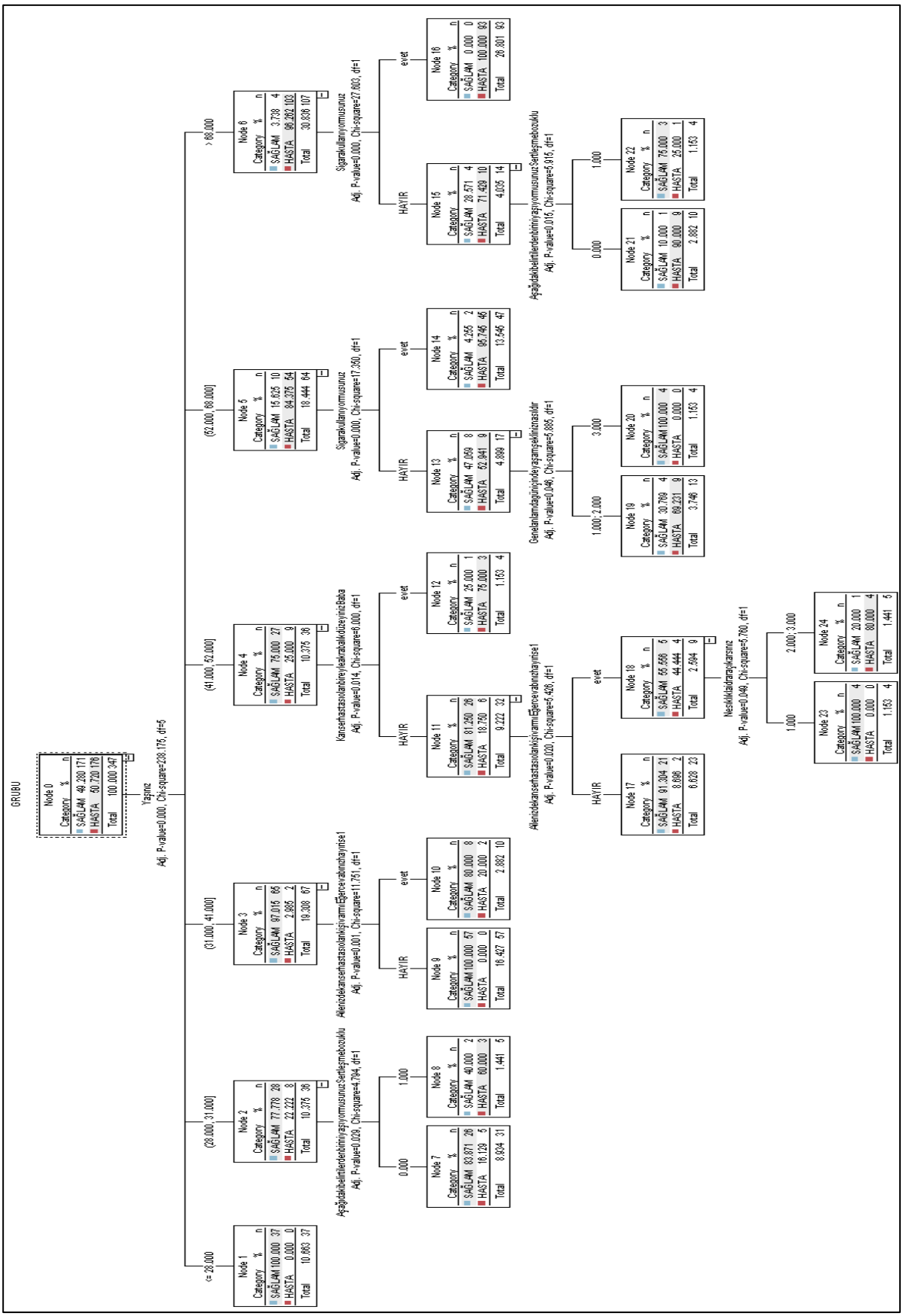
| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 12 | 164 | 93,18 |
| Prostat kanseri olmayan | 171 | 15 | 91,93 |
| | | | 92,54 |

CHAİD analizi ile oluşturulan modelinin prostat kanseri olanları doğru tahmin etme oranı %93,18, prostat kanseri olmayanları doğru tahmin etme oranı %91,93, testin toplam doğru tahmin oranı %92,54‘dır.

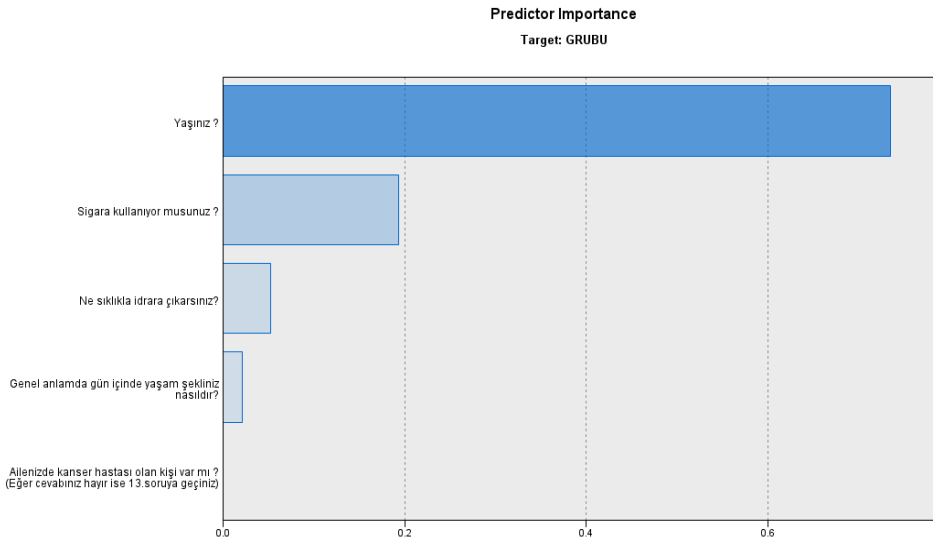
**Tablo 4-12:** CHAİD ile eğitilmiş modelin sınıflandırma tablosu

| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 7 | 67 | 90,54 |
| Prostat kanseri olmayan | 60 | 5 | 92,3 |
| | | | 91,36 |

CHAİD analizi ile oluşturulan modelinin prostat kanseri olanları doğru tahmin etme oranı %90,54, prostat kanseri olmayanları doğru tahmin etme oranı %92,3, testin toplam doğru tahmin oranı %91,36‘dır.




**Şekil 4-8:** CHAİD analizi karar ağacı.




**Şekil 4-9:** C5.0 analizi oluşturulan modelin değişkenleri

C5.0 analizi ile oluşturulan model de önem sırasına göre; yaş, sigara, idrar sıklığı, gün içindeki yaşam şekli ve ailede kanser hastası olması değişkenleri anlamlı bulunmuştur. Prostat kanseri hastalarını sınıflamak için kurulan C5.0 modeline; yaş değişkeni 0,73, sigara kullanımı 0,19, idrara çıkma sıklığı değişkeni 0,05, gün içindeki yaşam şekli 0,02 ve ailede kanser hastalığının varlığı 9,88E-23 oranda katkı sağlamıştır.

**Tablo 4-13:** C5.0 sınıflandırma tablosu

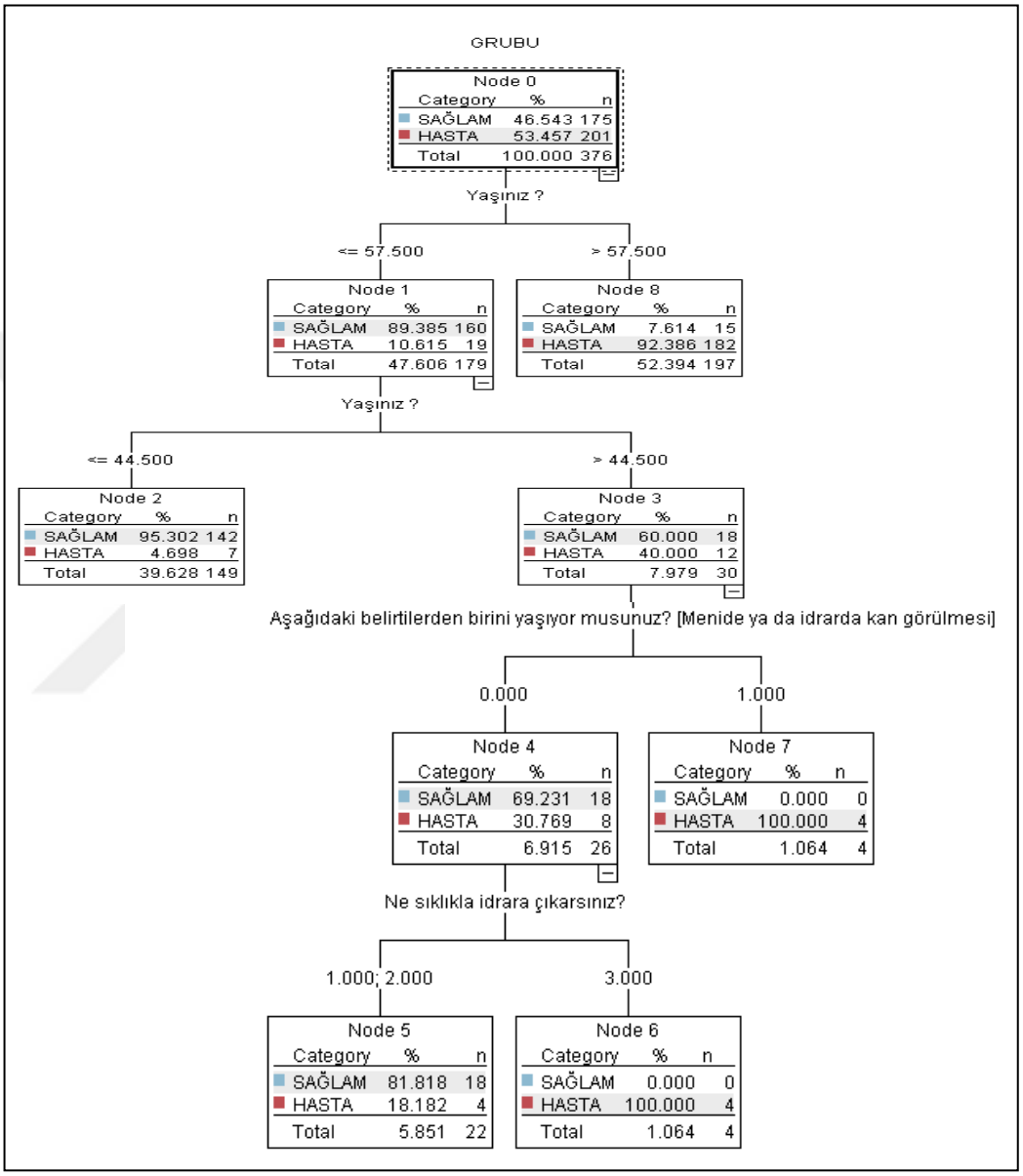
| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 17 | 159 | 90,34 |
| Prostat kanseri olmayan | 180 | 6 | 96,77 |
| | | | 93,64 |

C5.0 analizi ile oluşturulan modelinin prostat kanseri olanları doğru tahmin etme oranı %90,34, prostat kanseri olmayanları doğru tahmin etme oranı %96,77, testin toplam doğru tahmin oranı %93,64‘dir.

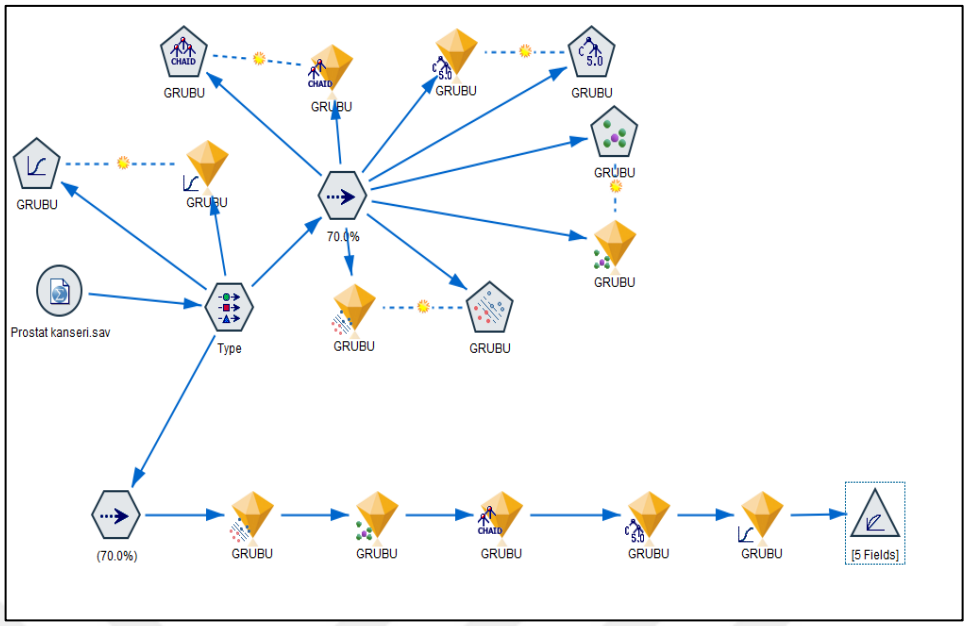
**Tablo 4-14:** C5.0 ile eğitilen modelin sınıflandırma tablosu

| Gözlenen | Tahmin edilen | | Doğru Sınıflama Yüzdesi |
|---|---|---|---|
| | Prostat kanseri olmayan | Prostat kanseri olan | |
| Prostat kanseri olan | 9 | 63 | 87,5 |
| Prostat kanseri olmayan | 60 | 7 | 89,5 |
| | | | 88,48 |

C5.0 analizi ile oluşturulan modelinin prostat kanseri olanları doğru tahmin etme oranı %89,36, prostat kanseri olmayanları doğru tahmin etme oranı %87,17, testin toplam doğru tahmin oranı %88‘dir.

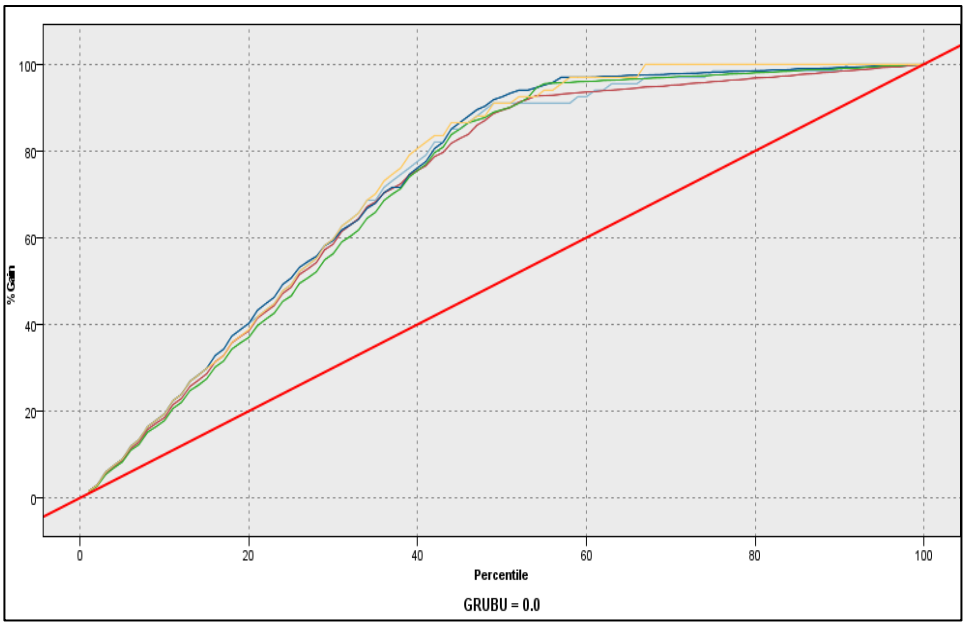



**Şekil 4-10:** C5.0 analizi karar ağacı**.**




**Şekil 4-11:** Lojistik regresyon ve oluşturulan modellerin karşılaştırılması**.**

Lojistik regresyon analizi değişkeni ve prostat kanseri verileri ile eğitilen SVM, KNN, CHAİD ve C5.0 algoritmalar ile oluşturulan modellerin gain değerleri karşılaştırılmıştır.




**Şekil 4-12:** LR, SVM, KNN, CHAİD, C5.0 analizlerinin gainlerinin karşılaştırılması.

Bilgi kazancı karşılaştırılması sonucunda LR > CHAİD > SVM > KNN > C5.0 sıralaması olduğu görülmüştür.

**Tablo 4-15:** Modellerin risk faktörleri

| Modeller | Risk Faktörleri |
|---|---|
| **LR** | Menide ve idrarda kan görülmesi, sigara içme, idrara çıkma sıklığı, ailede kanser varlığı, yaş, gün içindeki yaşam şekli |
| **SVM** | yaş, idrara çıkma sıklığı, sigara kullanımı, ailede kanser hastalığı görülme, gün içindeki yaşam şekli, yiyeceklerde kullanılan yağ, kronik hastalık varlığı, menide veya idrarda kan görülmesi, günlük su tüketimi ve kasık bölgesindeki rahatsızlık hissi |
| **KNN** | yaş, sigara kullanımı, ailede kanser hastalığı varlığı |
| **CHAİD** | yaş, sigara kullanımı, idrara çıkma sıklığı, sertleşme bozukluğu yaşama ve ailede kanser hastası varlığı |
| **C5.0** | yaş, sigara, idrar sıklığı, gün içindeki yaşam şekli ve ailede kanser hastası olması |

**Tablo 4-16:** Modellerin sınıflandırma yüzdeleri.

| Model | Sınıflama | Eğitim verisi | Deneme verisi |
|---|---|---|---|
| **SVM** | Bağımsız değişken sayısı | 10 | 10 |
| | Prostat kanseri olanlar | 97,1 | 94,4 |
| | Prostat kanseri olmayanlar | 98,92 | 85 |
| | **Doğru sınıflandırma yüzdesi** | 98 | **89,92** |
| **KNN** | Bağımsız değişken sayısı | 3 | 3 |
| | Prostat kanseri olanlar | 90,9 | 88,8 |
| | Prostat kanseri olmayanlar | 97,84 | 88 |
| | **Doğru sınıflandırma yüzdesi** | 94,47 | **88,48** |
| **CHAİD** | Bağımsız değişken sayısı | 5 | 5 |
| | Prostat kanseri olanlar | 93,18 | 90,54 |
| | Prostat kanseri olmayanlar | 91,93 | 92,3 |
| | **Doğru sınıflandırma yüzdesi** | 92,54 | **91,36** |
| **C5.0** | Bağımsız değişken sayısı | 5 | 5 |
| | Prostat kanseri olanlar | 90,34 | 87,5 |
| | Prostat kanseri olmayanlar | 96,77 | 89,5 |
| | **Doğru sınıflandırma yüzdesi** | 93,64 | **88,48** |

Tablo 4.15 de verilen modellerin sınıflandırma yüzdelerinden Lojistik regresyonun %92,2 toplam sınıflandırma yüzdesi olduğu da görülmektedir. Lojistik regreson analizinden sonra %91,36 sınıflandırma yüzdesi ile CHAİD analizi en iyi sınıflama yüzdesine sahiptir. SVM algoritması %89,92 doğru tahminleme yüzdesine sahiptir. En az sınıflandırma yüzdesine KNN ve C5.0 algoritmaları % 88,48 doğru tahminleme yüzdesine sahiptir.

# 5. TARTIŞMA

Kılıçarslan ve arkadaşlarının (2019) yaptıkları çalışmada prostat kanserini teşhisi üzerinde durmuşlardır. Hastalığı etkileyen genler teşhis edilmeye çalışılmış, veri madenciliği teknikleri kullanılarak çalışma gerçekleştirilmiştir. Destek makine vektörü ve KNN algoritmaları sonucu sınıflandırmada yüksek başarıya ulaşmışlardır. Atasoy ve Demiröz'ün (2021) gerçekleştirdiği çalışma prostat kanserinde tümör oluşumunda genomik faktörleri incelemişlerdir. Prostat kanserini tanılamak adına makine öğrenmesi yöntemlerini değerlendirip yüksek başarı elde etmişlerdir.

Hastalığın gen yönünün incelenmesi bakımından bu ve benzeri çalışmalara sıklıkla karşılaşılmaktadır. Ancak Literatürde çalışmamız paralelinde, direkt çevresel risk faktörünü ele alan çalışma bulunamadığından doğrudan karşılaştırma yapılamamaktadır. Literatürde makine öğrenmesi modellerinden elde edilen başarı oranları çalışmamızda da elde edilmiştir. Fakat LR analizinin makine öğrenmesi yöntemlerinden daha başarılı olduğu tespit edilmiştir.

Prostat kanserinin risk faktörlerini belirlemeye yönelik çalışmalar literatürde bulunmasına karşın direk risk faktörlerine ve belirlediğimiz risk faktörlerinin incelenmesi üzerine makine öğrenmesi çalışmasına rastlanılmamıştır. Çalışma alan yazına bu açıdan katkı sağlayacaktır. Lojistik regresyon analizi ile makine öğrenmesi yöntemlerinin karşılaştırıldığı çalışmalara rastlanılmaktadır, fakat prostat kanseri üzerine bu bağlamda literatürde boşluklar bulunmaktadır. Prostat kanseri özelinde modellerin denenmesi açısından literatüre katkı bu bağlamda da katkı sağlanacaktır.

Büyük veri ve makine öğrenme yöntemleri gizemli görünse de çoğu klinisyen tarafından bilinen geleneksel istatistiksel yöntemler temelli olduğu anlaşılmaktadır (Gulshan ve ark., 2016). Makine öğrenme yöntemlerinin amacı genel anlamda öğrenme ve öğrendiğini uygulamaktır. Öğrenme işlemlerini ise istatistiksel analizler ile gerçekleştirir. Bilgisayar bilimleri makine öğrenme kavramını geliştirirken istatistik bilimini kullanmaktadır. Son zamanlarda gündem haline gelen makine öğrenmesi ve yapay zekanın altında on yıllardır kullanılan bu geleneksel istatistiki yöntemlerin olduğu unutulmamalıdır. Makine öğrenmesi yöntemlerine yönelik alan yazın incelendiğinde kullanılan algoritmaların istatistikte kullanılan

sınıflama ve tahminleme analizleri olduğu fark edilecektir. Birbiri içine geçmiş yada istatistik biliminin popüler olan bu yöntemin içinde kaybolması veya unutulması çalışmaların performans ve güvenilirliğini düşürecektir. Zira istatistikte kullanılan metot ve yöntemlerin çeşitli varsayımları, kısıtları ve özellikleri bulunmaktadır ve bunların ihlali yanlış sonuçların elde edilmesine neden olacaktır. Günümüzde yürütülen çalışmaların çok branşlı olduğu bilim dallarının birbiriyle etkileşim içinde olduğu gerçeğinden hareketle; araştırmacıların makine öğrenmesi yöntemleri çalışmalarında bilgisayar ve yazılım alanlarından aldıkları destekleri istatistikten de almaları gerekmektedir.

# 6. SONUÇ VE ÖNERİLER

Destek makine vektörü yada makine destek vektörü olarak adlandırılan SVM ile oluşturulan modelde; yaş, idrara çıkma sıklığı, sigara kullanımı, ailede kanser hastalığı görülme, gün içindeki yaşam şekli, yiyeceklerde kullanılan yağ, kronik hastalık varlığı, menide veya idrarda kan görülmesi, günlük su tüketimi ve kasık bölgesindeki rahatsızlık hissi değişkenleri risk faktörü olarak saptanmıştır. Modeldeki risk faktörü sayısı diğer algoritmaların tespit ettiği risk faktörü sayısından fazladır.

SVM modeli eğitim veri setinde; prostat kanseri olmayan (sağlıklı) kişileri %98,92 oranla, prostat kanseri olan (hasta) kişileri %97,1 oranla ve genel tahmini %98 ile gerçekleştirmiştir. Bu değerlerle verinin SVM modelinde başarılı olduğu görülmüştür. Fakat eğitilen model deneme veri setinde prostat kanseri olmayan (sağlıklı) kişileri %85 oranla, prostat kanseri olan (hasta) kişileri %94,4 oranla ve genel tahmini %89,92 ile gerçekleştirmiştir. Eğitilen modelin genel tahmin oranın düştüğü tespit edilmiştir. Özellikle prostat kanseri hastası olmayan bireylerin tahmin oranı eğitimde %98,92 iken eğitilen modelin deneme setinde %85' e düştüğü görülmüştür. SVM ile eğitilen modelin prostat kanseri olmayan (sağlıklı) bireyleri tahmin yeteneğinin diğer algoritmalara göre daha düşük olduğu sonucuna varılmıştır.

K-En Yakın Komşu algoritması (KNN) ile elde edilen modelde yaş, sigara kullanımı ve ailede kanser hastalığı varlığı değişkenleri risk faktörü olarak tespit edilmiştir. Model eğitim veri setinde prostat kanseri olmayan (sağlıklı) kişileri %97,84 oranla, prostat kanseri olan (hasta) kişileri %90,9 oranla ve genel tahmini %94,47 ile gerçekleştirmiştir. Eğitilen model deneme veri setinde prostat kanseri olmayan (sağlıklı) kişileri %89,5 oranla, prostat kanseri olan (hasta) kişileri %87,5 oranla ve genel tahmini %88,48 ile gerçekleştirmiştir. Algoritmanın diğer algoritmalara göre en düşük doğru sınıflandırma yüzdesi olduğu görülmüştür.

Otomatik Ki-Kare Etkileşim Belirleme (CHAİD) analizi ile yaş, sigara kullanımı, idrara çıkma sıklığı, sertleşme bozukluğu yaşama ve ailede kanser hastası varlığı risk faktörü olarak tespit edilmiştir. Model eğitim veri setinde prostat kanseri olmayan (sağlıklı) kişileri

%91,93 oranla, prostat kanseri olan (hasta) kişileri %93,18 oranla ve doğru tahmini %92,54 ile gerçekleştirmiştir. Eğitilen model deneme veri setinde prostat kanseri olmayan (sağlıklı) kişileri %92,3 oranla, prostat kanseri olan (hasta) kişileri %90,54 oranla ve genel tahmini %91,36 ile gerçekleştirmiştir. Veri yapısının CHAİD analizine uygunluğu nedeniyle öğrenen CHAİD modelinin doğru sınıflandırma yüzdesi LR den sonra en yüksek yöntem olarak tespit edilmiştir.

C5.0 algoritmasında oluşturulan modelde yaş, sigara, idrar sıklığı, gün içindeki yaşam şekli ve ailede kanser hastası olması risk faktörüdür. C5.0 eğitiminde sağlıklı bireyleri %96,77, hasta bireyleri %90,34 ve sınıflandırma oranını %93,64 ile gerçekleştirmiştir. Eğitilen model denendiğinde sağlıklı bireyleri %87,5 oranla, hasta bireyleri %89,5 oranla ve sınıflandırma oranını %88,48 ile gerçekleştirmiştir. Algoritmanın diğer algoritmalara göre KNN ile birlikte en düşük doğru sınıflandırma yüzdesi olduğu görülmüştür.

Toplamda 20 bağımsız değişkenin bulunduğu ilk modelden elde edilen bilgilerle kurulan son model de 6 değişken ile başarılı bir model oluşturulmuştur. Lojistik regresyon analizinde oluşturulan modelde risk faktörü olarak; Yaş, sigara içme, ailede kanser varlığı, idrara çıkma sıklığı, gün içindeki yaşam şekli ve menide ve idrarda kan görülmesi bulunmuştur.

Lojistik regresyon analizi modeli prostat kanseri olmayan (sağlıklı) kişileri %92,3 oranla tahmin edebilme yeteneğine sahiptir. Prostat kanseri olan (hasta) kişileri %92,1 oranla tahmin edebilme yeteneğine sahiptir. Analizler sonucunda Lojistik regresyon analizinin prostat kanseri hasta olanlar ile olmayanları sınıflandırma yeteneği %92,2 bulunmuştur. Algoritmalar içinde en yüksek doğru tahmin yüzdesi bu değerle lojistik regresyona aittir.

Mühendislik, finans, tıp gibi çeşitli alanlarda kullanılan temeli 19 yüzyıla dayanan lojistik regresyon analizi prostat kanseri verilerin üzerinde %92,2 yüzdeyle çok yüksek bir sınıflandırma yeteneğine sahiptir. Şekil 4.11 de modellerin bilgi kazançları incelenmiştir, inceleme sonucunda LR'nin daha fazla bilgi kazancı sağladığı görülmüştür. Çalışmada yapılan tüm kıyaslamalar sonucunda; SVM, KNN, CHAİD, C5.0 analizleri içinde prostat kanseri hastalığının teşhisinde ve tahminin de lojistik regresyonun en iyi sınıflandırıcı analiz olduğu görülmüştür.

Eğitilen modellerin deneme verilerinde yapılan tahminlerde, başarı oranın düştüğü görülmüştür fakat bu düşüşün SVM modelindeki kadar düşük olmadığı diğer algoritmaların

sonucuna bakıldığında da görülmektedir. Gerçekleştirilecek yeni çalışmalarda bu durumun incelenmesi SVM algoritmalarının uygulanabilirliğine katkı sağlayacaktır.

Prostat kanseri verileri üzerinde gerçekleştirilen çalışmada ağırlıklı olarak nitel değişkenler ile çalışılmıştır. Hastalık risk faktörlerini belirlemek üzere yapılacak yeni çalışmalarda ölçümler ile elde edilmiş sürekli değişkenlerin ağırlıklı olduğu veriler ile çalışmak faydalı olacaktır. Çalışma 253'ü prostat kanseri olmayan (sağlam), 248'i prostat kanseri olan (hasta) toplam 501 kişi ile gerçekleştirilmiştir. Çalışma öncesinde örneklem büyüklüğü hesaplanarak yeterli sayı elde edilmiştir. Ancak dünyada ve ülkemizde erkekler arasında en sık görülen kanser türlerinden biri olan bu hastalığın incelenmesinde yapılacak yeni çalışmalarda, daha büyük örneklemlerle ve gerçek veriler (simüle olmayan) ile çalışılması önerilmektedir. Çalışmada farklı algoritmalar ile oluşturulan modellerden, elde edilen risk faktörlerinin yeni çalışmalarda incelenerek uzman görüşleriyle doğruluğunun incelenmesi önerilmektedir.

# KAYNAKLAR


Afrin, F. , Nahar, I. , 2015. *Incremental learning based intelligent job search system*, Doktora Tezi- BRAC Üniversitesi.

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. John Wiley and Sons, New York.

Aktaş Cengiz, (2009), "Lojistik Regresyon Analizi: Öğrencilerin Sigara İçme Alışkanlığı Üzerine Bir Uygulama", *Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 26, 107-121.

Albayrak A.S. (2006), *Uygulamalı Çok Değişkenli İstatistik Teknikleri*, Asil Yayın Dağıtım, Ankara.

Allison, Paul D. (2010) *Measures of fit for logistic regression*. Statistical Horizons LLC and the University of Pennsylvania.

Alpaydın E., (2010) *Introduction to Machine Learning*, 2nd ed., The MIT Press, London, 2010

Alpaydin, E. (2020). Introduction to Machine Learning (Fourth bas.). MIT. ss. 19, 1-3

Alpar, R. (2017). Uygulamalı çok değişkenli istatistiksel yöntemler.

Arabacı Ö. (2002), *Lojistik Regresyon Analizi ve Bir Uygulama Denemesi*, Uludağ Üniversitesi Sosyal Bilimler Enstitüsü, (Yayınlanmamış Yüksek Lisans Tezi), Bursa

Arul K. Jothi, Subburam S, Umadevi V, Hemavathy K., (2021) *Heart disease prediction system using machine learning*, Mater. Today Proc., s.1–3.

Ayaz, M. (2021). *Makine öğrenmesi algoritmaları ile covid-19 hastalarının belirlenmesi* Yüksek Lisans Tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü.

Aydın Atasoy, N. Demiröz, A. (2021). *Makine Öğrenmesi Algoritmaları Kullanılarak Prostat Kanseri Tümör Oluşumunun İncelenmesi*. *Avrupa Bilim ve Teknoloji Dergisi*, 87-92.

Ayodele T. (2010*). Available at InTech open*: http://www.intechopen.com/books/new-advances-inmachine-learning/types-of-machine-learning-algorithms

Berenson, M.L. ve Levine, D. M. (1996). *Basic Business Statistics: Concepts and Applications*, Sixth Edition, Prentice-Hall International, 837-838.

Brant R. (1990) Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics. 49(4): 1171-1178.

Burges C. J., A tutorial on support vector machines for pattern recognition, *Data mining and knowledge discovery*, 1998, 2(2), 121-167.

Camastra, F. ve Vinciarelli, A., (2008). Machine Learning for Audio, Image and Video Analysis: Theory and Applications, *Springer*, ISBN: 1-84800-007-3.

Caruana, R., Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168), ACM.

Christensen, R., (1997), *Log-Linear Models*, Springer Science & Business Media.

Cordeiro, G.M., (2004)., On Pearson's residuals in generalized linear models, *Statistics and Probability Letters*, 66, 213-219.

Cragg, J.G., Uhler, R.S. (1970). The demand for automobiles. *The Canadian Journal of Economics*, 3: 386-406.

Cruz, J. A., Wishart, D. S., (2006). Applications of machine learning in cancer prediction and prognosis, *Cancer informatics*, 2,11-16.

Çakır, M., (2005). *Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makina Öğrenmesi Teknikleri: Ampirik Uygulamalar ve Karşılaştırmalı Analiz*, Uzmanlık Yeterlilik Tezi, Türkiye Cumhuriyet Merkez Bankası İstatistik Genel Müdürlüğü.

Çölkesen, İ. (2009). *Uzaktan algılamada ileri sınıflandırma tekniklerinin karşılaştırılması ve analizi.* Gebze Yüksek Teknoloji Entitüsü, Jeodezi ve Fotogrametri Mühendisliği, Y. Lisans Tezi.

Diri B. (2020) *Makine Öğrenmesine Giriş Yıldız Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü,* Ders Notu https://avesis.yildiz.edu.tr/diri/dokumanlar

Dolgun, Saraçbaşı O. (2015) *Lojistik Regresyon Çözümlemesi.* Ankara: Hacettepe Üniversitesi Yayınları.

Elhan, A. H. (1997). *Lojistik regresyon analizinin incelenmesi ve tıpta bir uygulaması.* Yüksek Lisans Tezi, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü.

Elmas Ç., (2003). *Yapay Sinir Ağları (Kuram, Mimari, Eğitim, Uygulama)*, Seçkin Yayıncılık: Ankara.

Eray, O. (2008). *Destek vektör makineleri ile ses tanıma uygulaması.* Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü, Denizli.

Fagerland MW, Hosmer DW. (2016) Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 86(17): 3398-3418.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems.* USA: O'Reilly Media, Inc.

Goodfellow, I., Yoshua, B., ve Courville, A. (2016). *Deep Learning.* MIT Press. 10.8.2022 tarihinde

Gulshan, V, (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316.22: 2402-2410.

Gürsakal, N., (2021). *Yapay Zeka Fareli Köyün Kavalcısı,* (1. Baskı, Bursa), Dora Basım-Yayın Dağıtım Ltd.Şti.

Hagan R., Gillan C. J., and Mallett F., (2021) Comparison of machine learning methods for the classification of cardiovascular disease, *Informatics Med.* 24:100606.

Hair, J. F., Black, W. C., Babin, B., Anderson, R. E., Tatham, R. L. (2006). *Multivariate data analysis* (6th ed). Upper Saddle River, NJ: Prentice-Hall.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), 18- 28

Herbrich, R. (2002). *Learning Kernel Classifiers.* The MIT Press

Hoare, R., (2004), *Using CHAID for Classification Problems*. New Zealand Association Conference,Wellington,s.1

Hosmer, D. W., Lemeshow S. (1980). Goodness-of-fit tests for the multiple logistic regression model. Communications in Statistics. *Theory and Methods* 9: 1043–1069.

Hosmer, D. W., Lemeshow S.. (2000). *Applied Logistic Regression*. 2nd ed. New York: Wiley.

Hosmer, D. W., Lemeshow, S., (1989) *Applied Logistic Regression*. John Wiley & Sons, New York, 5-50 (1989).

Hosmer, DW, Lemeshov S. (1980) Goodness of fit Tests fort he multiple Logistic Regression Model, *Communications in Statistics*, 9,1043-1069.

http://www.ttbilgin.com/2018-2019-guz/veri-bilimine-giris/hafta6/siniflandirma2.pdf

Işığıçok E. (2004), *Toplam Kalite Yönetimi Bakış Açısıyla İstatistiksel Kalite Kontrol,* Ezgi Kitabevi Yayınları, Bursa

İçöz O., Kozak M. (2002). *Turizm Ekonomisi,* Turhan Kitabevi, Ankara. Kültür ve Turizm Bakanlığı İstatistikleri.

İşeri İ., (2014) *Mamogram Görüntülerinden Makine Öğrenmesi Yöntemleriyle Meme Kanseri Teşhisi,* Yüksek Lisans Tezi. Sakarya Üniversitesi Fen Bilimleri Enstitüsü,

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer p.18

Karagöz, Y. (2019). SPSS AMOS META Uygulamalı Nitel-Nicel Karma Bilimsel Araştırma Yöntemleri ve Yayın Etiği, 2. Basım, Ankara: Nobel Yayıncılık.

Kartal, E. (2015). Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama. Yüksek Lisans Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü / Enformatik Ana Bilim Dalı

Kılıçarslan, S., Adem, K., Cömert, O. (2019). Parçacık Sürü Optimizasyonu Kullanılarak Boyutu Azaltılmış Mikrodizi Verileri Üzerinde Makine Öğrenmesi Yöntemleri ile Prostat Kanseri Teşhisi. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi,* 7(1),769-777.

Larose, D.T., (2005). *Discovering knowledge in data an introduction to data mining.* NY: John Wiley & Sons.

Liu, X. (2015). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling.* Sage publications.

Maddala, G. S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge

Malthus, T.R. (1789). *An Essay on the Principle of Population*. https://www.econlib.org/library/Malthus/malPop.html (16.04.2019)

McFadden, (1973) *Quantitative Methods for Analyzing Travel Behaviour of Individuals: Some Recent Developments*, Hensher, D.A. & Stopher, P.R.: (eds): Behavioural travel modelling. London: Croom Helm, pp. 279-318

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* (Edited by P. Zarembka), Academic Press, New York. P.105-142.

Meyer, D., Wien, F.T. (2001*). Support vector machines. Porting R to Darwin/X11 and Mac OS X*, 1, 23

Minitab,https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistical-modeling/regression/how-to/fit-binary-logistic-model/interpret-the-results/all-statistics-and-graphs/coefficients-and-regression-equation/

Nacar, E.N., Erdebilli, B. (2021). *Makine Öğrenmesi Algoritmaları İle Satış Tahmini.* Endüstri Mühendisliği, 32 (2) , 307-320.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78: 691-692.

Nguyen, A., (2020). *Comparative study of C5.0 and CART algorithms.* Retrieved from:http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/NguyenPresentation%20Data%20mining.pdf.

Oğuztürk, G. (2018). *Diyabet hastalığının makine öğrenmesi algoritmaları ile en iyi doğru tahmininin elde edilmesi.* Yüksekte lisans Tezi, Kırıkkale Üniversitesi, Fen Bilimleri Enstitüsü, Kırıkkale.

Okal, G. (2019). *Meme kanseri riskinin termal görüntüleme ve makine öğrenmesi ile saptanması*. Yüksek Lisans Tezi, Pamukkale Üniversitesi Fen Bilimleri Enstitüsü.

Örenç, S. (2019). *Kronik obstrüktif akciğer hastalığı teşhisi için makine öğrenmesi temelli yeni bir yöntem geliştirilmesi* Yüksekte lisans Tezi, Sakarya Üniversitesi, Fen Bilimleri Enstitüsü, .

Özdamar, K., (2004). *Paket programlarla İstatistiksel veri analizi-1*. Kaan Kitapevi, Eskişehir.

Özekes S., (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Dergisi*, 3, 65-82

Özkan Y., (2008). *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık Eğitim, İstanbul, ISBN 978-975-6797-82-2

Öztemel E. (2012). *Yapay sinir ağları*. Papatya Yayıncılık, s.48.

Pearl, R. ve Reed, LJ (1920) 1790'dan beri Birleşik Devletler Nüfusunun Büyüme Hızı ve Matematiksel Temsili Üzerine. *Amerika Birleşik Devletleri Ulusal Bilimler Akademisi Bildirileri,* 6, 275-288.

Pehlivan, G. (2006). *Chaid analizi ve bir uygulama.* Yüksek Lisans Tezi. Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü İstanbul

Roberts G, Rao NK, Kumar S. (1987). Logistic regression analysis of sample survey data. *Biometrika,* 74: 1-12.

Sebe, N., Cohen, I., Garg, A., Huang, TS (2005). Bilgisayarla görmede makine öğrenimi. *Springer Bilim ve İş Medyası*.

Selçuk, Z. (1999). *Gelişim ve Öğrenme: Eğitim Psikolojisi.* Ankara: Nobel Yayın Dağıtım.

Sharma V., Yadav S., and Gupta M., (2020). Heart Disease Prediction using Machine Learning Techniques," Proc.-IEEE 2020 2nd *Int. Conf. Adv. Comput. Commun. Control Networking,* 1, 6: 177–181.

Silahtaroğlu G., (2008). *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*. Papatya Yayıncılık Eğitim, İstanbul, ISBN 978-975-6797-81-5.

Tanık, N. (2021). *EMG verilerine dayalı olarak makine öğrenmesi yöntemleri ile polinöropati teşhisi.* Yüksek Lisans Tezi, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü.

Tatlıdil H. (2002), *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* Ziraat Matbaacılık A.Ş. Ankara

Tatlıdil, H. (1996). *Uygulamalı Çok Değişkenli İstatistiksel Analiz,* Cem Ofset Ltd. Şti., Ankara.

Türk Dil Kurumu, (2022), Öğrenmek, http://www.tdk.gov.tr/, [Ziyaret Tarihi: 5.8.2022].

Ulupınar, S. D. (2007). *2001 Kriz Dönemi, Öncesi ve Sonrasında Türk Ticari Bankalarının Karlılıklarının Lojistik Regresyon Analizi ile incelenmesi*, Yüksek Lisans Tezi, Marmara Üniversitesi, İstatistik Bilim Dalı İstanbul. S.39.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, 54(3), 426-482.

Wright, R.E. (1994) Reading and Understanding Multivariate Statistics, *American Psychological Association*, p.220.

Yahyaoui, A. (2017). *Göğüs hastalıklarının teşhis edilmesinde makine öğrenmesi algoritmalarının kullanılması* Doktora Tezi, Sakarya Üniversitesi Fen bilimleri Enstitüsü).

Yıldız, A. (2019). *Makine öğrenmesi yöntemleri ile tiroit hastalığının teşhisi* (Yüksek Lisans Tezi, Sakarya Üniversitesi, Uygulamalı Bilimler Anabilim Dalı

Yılmaz, B. (2022). *Bilgisayarlı tomografi görüntüleri üzerinden makine öğrenme algoritmaları kullanılarak 1, 11. Ve 12. torakal vertebralardan cinsiyet tahmini*. Yüksek Lisans Tezi, Karabük Üniversitesi, Lisansüstü Eğitim Enstitüsü.

# EKLER

## Ek-1

**PROSTAT KANSERİ BİLGİ TOPLAMA FORMU**

**1.** Yaşınız …….

**2.** Kilonuz ……

**3.** Medeni durum

Bekar ☐
Evli ☐
Diğer ☐

**4.** Eğitim düzeyiniz

Okur-yazar ☐
İlkokul ☐
Ortaokul ☐
Lise ☐
Üniversite ☐

**5.** Mesleğiniz

İşçi ☐
Serbest meslek ☐
Öğrenci ☐
Akademik personel ☐
Devlet memuru ☐
Çalışmıyor ☐
Emekli ☐

**Ek-1 (devamı)**

Sağlık personeli ☐

Öğretmen ☐

**6.** Doğduğunuz şehir……..

**7.** Yaşadığınız şehir……

**8.** Sigara kullanıyor musunuz

Evet ☐

Hayır ☐

Eskiden kullanıyordum ☐ içilen yıl sayısı ve adedi ……yıl……adet

**9.** Alkol alıyor musunuz

Evet ☐

Hayır ☐

Eskiden kullanıyordum ☐ kullanın yıl sayısı…….

**10.** Şimdiye kadar hiç PSA testi yapıldı mı

Evet ☐ 1'in üstü ☐ 1'in altı ☐ Hayır ☐

**11.** Ailenizde kanser hastası olan kişi var mı (eğer cevabınız hayır ise 13. Soruya geçiniz)

Evet ☐ Hayır ☐

**12.** Kanser hastası olan kişi ile akrabalık düzeyiniz

Anne ☐ Teyze ☐

Baba ☐ Hala ☐

Kardeş ☐ Dayı ☐

Dede ☐ Amca ☐

Nine ☐

(Kanser türünü bu alana yazınız)……..

**Ek-1 (devamı)**

**13.** Beslenme şekliniz ( bu soruda birden fazla seçenek işaretleyebilirsiniz)

Kırmızı et ağırlıklı ☐

Beyaz et ağırlıklı ☐

Sebze ağırlıklı ☐

Bakliyat ve tahıl ağırlıklı ☐

**14.** Günde ortalama kaç litre su tüketiyorsunuz

Ortalama 1litre ☐ ortalama 2.5 litre ☐ 2.5 ve üzeri ☐

**15.** Hangi pozisyonda idrar yaparsınız

Genelde ayakta ☐ Genelde oturarak ☐ Hep oturarak ☐ Hep ayakta ☐

**16.** Kronik bir hastalığınız var mı

Evet (var) ☐ (hastalığınızı buraya yazınız)…….

Hayır (yok) ☐

**17.** Ne sıklıkla idrara çıkarsınız

Günde 5 kez ve aşağısı ☐ Günde 5 – 10 kez arası ☐ 10 kez ve üzeri ☐

**18.** Genel olarak gün içindeki yaşam şekliniz nasıldır

Hareketsiz bir yaşam şeklim vardır ☐

Spor yaparım ☐

Spor yapmam ama gün içinde hareketliyimdir ☐

**19.** Yiyeceklerinizde hangi yağları kullanmayı tercih edersiniz

Hayvansal yağları ☐

Bitkisel yağları ☐

**20.** Aşağıdaki belirtilerden birini yaşıyor musunuz (birden fazla cevap işaretlenebilir)

| | | | |
|---|---|---|---|
| İdrar yapma güçlüğü | ☐ | Kasık bölgesinde rahatsızlık hissi | ☐ |
| İdrar akışında kuvvet azalması | ☐ | Kemik ağrıları | ☐ |
| Menide ya da idrarda kan görülmesi | ☐ | Sertleşme bozukluğu | ☐ |
| Boşalma esnasında ağrı | ☐ | Hayır hiçbirini yaşamıyorum | ☐ |

# İNTİHAL RAPORU İLK SAYFASI

LOJİSTİK REGRESYON ANALİZİ ve MAKİNE ÖĞRENMESİ ALGORİTMALARI ile PROSTAT KANSERİ RİSK FAKTÖRLERİNİN BELİRLENMESİ ÜZERİNE BİR ÇALIŞMA

ORİJİNALLİK RAPORU

| %12 | %11 | %3 | %3 |
|---|---|---|---|
| BENZERLİK ENDEKSİ | İNTERNET KAYNAKLARI | YAYINLAR | ÖĞRENCİ ÖDEVLERİ |

BİRİNCİL KAYNAKLAR

| | | |
|---|---|---|
| 1 | acikbilim.yok.gov.tr<br>İnternet Kaynağı | %3 |
| 2 | dergipark.org.tr<br>İnternet Kaynağı | %1 |
| 3 | openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | %1 |
| 4 | robotikcerrahi2017.blogspot.com<br>İnternet Kaynağı | %1 |
| 5 | www.researchgate.net<br>İnternet Kaynağı | %1 |
| 6 | acikarsiv.ankara.edu.tr<br>İnternet Kaynağı | <%1 |
| 7 | Submitted to Istanbul University<br>Öğrenci Ödevi | <%1 |
| 8 | acikarsiv.aydin.edu.tr<br>İnternet Kaynağı | <%1 |

sabita.medipol.edu.tr