T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YAKLAŞIMI İLE TİROİT KANSERİNİN TANILANMASI

DOKTORA TEZİ

Mehmet Emin ASAN

Endüstri Mühendisliği Anabilim Dalı

ŞUBAT 2024

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YAKLAŞIMI İLE TİROİT KANSERİNİN TANILANMASI

DOKTORA TEZİ

Mehmet Emin ASAN

Endüstri Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Harun TAŞKIN

ŞUBAT 2024

Mehmet Emin ASAN tarafından hazırlanan "VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YAKLAŞIMI İLE TİROİT KANSERİNİN TANILANMASI" adlı tez çalışması 14.02.2024 tarihinde aşağıdaki jüri tarafından oy birliği ile Sakarya Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı'nda Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Jüri Başkanı :

Jüri Üyesi :

Jüri Üyesi :

Jüri Üyesi :

Jüri Üyesi :

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Sakarya Üniversitesi Fen Bilimleri Enstitüsü Lisansüstü Eğitim-Öğretim Yönetmeliğine ve Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesine uygun olarak hazırlamış olduğum "Veri Madenciliği ve Makine Öğrenmesi Yaklaşımı ile Tiroit Kanserinin Tanılanması" başlıklı tezin bana ait, özgün bir çalışma olduğunu; çalışmamın tüm aşamalarında yukarıda belirtilen yönetmelik ve yönergeye uygun davrandığımı, tezin içerdiği yenilik ve sonuçları başka bir yerden almadığımı, tezde kullandığım eserleri usulüne göre kaynak olarak gösterdiğimi, bu tezi başka bir bilim kuruluna akademik amaç ve unvan almak amacıyla vermediğimi ve 20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince Sakarya Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Enstitü tarafından belirlenmiş ölçütlere uygun rapor alındığını, etik kurul onay belgesi aldığımı, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun ortaya çıkması halinde doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi beyan ederim.

(14/02/2024)

Mehmet Emin ASAN

Eşime ve çocuklarıma

TEŞEKKÜR

Doktora eğitimim süresince her konuda desteğini ve yardımlarını benden esirgemeyen değerli danışman hocam Prof. Dr. Harun TAŞKIN'a teşekkürlerimi sunarım.

Tez izleme jürimde yer alan ve beni yönlendiren değerli hocalarım Prof. Dr. Cemalettin KUBAT ve Prof. Dr. Cemil ÖZ'e; akademik hayatıma katkı sunan ve derslerinden çok istifade ettiğim değerli hocam Sayın Prof. Dr. Özer UYGUN'a çok teşekkür ederim.

Araştırma konumun belirlenmesinde ve tezim için gerekli verilerin temin edilip veri seti oluşturulmasında büyük emekleri olan hocalarım Prof. Dr. Murat ALEMDAR'a, Doç. Dr. Recayi ÇAPOĞLU'na ve Uzm. Dr. Tarık HARMANTEPE'ye katkılarından dolayı çok teşekkür ederim.

Eğitim hayatım süresince ve sonrasında maddi ve manevi desteklerini benden esirgemeyen saygıdeğer hocam ve ağabeyim Sayın Doç. Dr. Rüstem KELEŞ'e; eğitim hayatıma her daim katkılar sunan ve öğretileri hâlâ önümü aydınlatan ilkokul öğretmenim Sayın Hasan ÇELİKOĞLU'na; pozitif, enerjik ve samimi duruşuyla daima manevi ilham kaynağım olan dostum, ağabeyim ve hocam Sayın Bünyamin KAYIKÇI'ya teşekkürlerimi sunarım.

Bugünlere ulaşmamda emeklerini hiçbir zaman ödeyemeyeceğim aileme, daima yanımda olan değerli eşim Fatma ÇOLAK'a, ablalarım Serpil BAĞ ve Figen ÇOLAK'a, ağabeylerim Nevzat ASAN ve Necip ASAN'a, varlığı ile bana enerji veren oğlum Vefa'ya ve hayatıma şifa olan kızım Şifa'ya sonsuz teşekkürlerimi sunarım.

Mehmet Emin ASAN

İÇİNDEKİLER

Sayfa

KISALTMALAR

CNN : Evrişimsel sinir ağları
DVM : Destek vektör makineleri
GA : Genetik algoritmalar
İBYSA : İleri beslemeli yapay sinir ağları
İİA : İnce iğne aspirasyonu
KNN : K-en yakın komşu
LR : Lojistik regresyon
MÖ : Makine öğrenmesi
NB : Naif bayes
RO : Rastgele orman
YSA : Yapay sinir ağları
YZ : Yapay zekâ
WBCD : Wisconsin meme kanseri veri seti

SİMGELER

a : Ağırlık

b : Sapma

da : a'ya göre kayıp fonksiyonunun kısmi türevi

db : b'ya göre kayıp fonksiyonunun kısmi türevi

d(a,b) : Öklid uzaklığı

e : Sapma

J(w,b) : Destek vektör makinelerinde kullanılan maliyet fonksiyonu

L : Öğrenme oranı

P(x) : Lojistik regresyon sınıflandırma

R(i) : Korelasyon değeri

W : Ağırlık

X : Girdi bağımsız değişkeni[nümerik]

y_i : Destek vektör makineleri algoritmasında sınıflandırma sonucu

Y : Sınıf [ikili bağımlı değişken]

Y = wX – b : Doğrusal hiperdüzlem denklemi

(X1 – X2) : Destek vektör makinelerinde marjin

TABLO LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

VERİ MADENCİLİĞİ VE MAKİNE ÖĞRENMESİ YAKLAŞIMI İLE TİROİT KANSERİNİN TANILANMASI

ÖZET

Sağlık sistemi, hedeflenen kitlenin sağlıkları ile ilgili tüm hizmetlerini görmek üzere ilgili kurumların, kaynakların ve insanların bir arada örgütlendiği bir sistem olarak tanımlanabilir. DSÖ'ne göre Sağlık sistemi, sağlığı korumak, hastalıkları tedavi etmek ve bireylerde sağlıklı gelişimi sağlamak üzere kişileri, eylemleri ve organizasyonları kapsamaktadır.

Her yıl artarak devam eden hastalık ve ölüm oranları zaman, para ve insan gücü maliyetleri açılarından büyük giderlere muhatap olmaktadır. Hızlı artan bu maliyetler daha fazla insan gücü istihdamı ile çözülememektedir. Teknolojinin gelişimi ve birçok alanda olduğu gibi tıp alanında da yararlı neticeler vermesi sonucu, yukarıda belirtilen problemlerin çözümü ve maliyetlerin olabildiğince aza indirilerek verimliliğin arttırılması amacıyla, uzmanlar tarafından yürütülen karar süreçleri zeki karar destek sistemleri ile desteklenme potansiyeline sahiptir.

Veri bilimi (Data science) ve eldeki hasta verilerinin birer anlamlı bilgiye dönüştürülmeleri için kullanılacak teknikler elde etmek istenilen sonuçların anahtarı niteliğindedir. Bu anahtar, maliyetleri azaltacak ve doktorların üzerinden fazlalık işleri alarak hastalara daha fazla bakım ve daha verimli tedavi olanakları sağlayacak şekilde sağlık sistemlerini başarıya ulaştıracaklardır.

Veri bilimi bu başarısını, çoklu veri kullanımı ile tanı ve tedavide çoklu seçeneklere sahip olmasına borçludur ve hatalı sonuçlar vermesi çok küçük düzeylerde kalmaktadır. Buna mukabil birçok doktorun son tıbbi gelişmeler hakkında yayınlanan örneğin binlerce makaleyi (verileri kontrol altında tutmak açısından) okuyabilmesi ve elde ettiği tüm verilerden analizler çıkartarak tanı ve tedavi yapabilmesi mümkün görünmemektedir.

Uzman kontrolünde yazılım ve donanım olarak kullanılan teknolojik yapılar, günümüzde yapay zekâ adıyla anılan ve kendi kendine öğrenerek karar çıktıları üretebilen makine öğrenmesi türünden algoritmalar ile desteklenebilmektedirler.

Bu araştırmada önerilen yöntem, uygun veri toplama (bizim durumumuzda, çevrimiçi eldeki veri setleri), veri seçimi, veri ön işleme işlemlerinin gerçekleştirilmeleri ile birlikte destek vektör makineleri ve lojistik regresyon gibi makine öğrenmesi yöntemlerinin karar destek sistemleri olarak modellenmelerinde veri madenciliği ve makine öğrenmesi yaklaşımları olmuştur.

Elde edilen bilgiler, tablolar ve diyagramlarda gösterilmek üzere sonuçların görselleştirilmesi ve değerlendirilmesi ile başlayan farklı veri madenciliği süreçlerini de içermektedir. Tanı ve tedavi için eldeki kullanılabilir veriler, destek vektör makineleri, lojistik regresyon teknikleri ile diğer makine öğrenmesi algoritmaları ile uzmanlık seviyesinde karar verme süreçlerine yardımcı olabilir bilgilere dönüştürülmüştür.

Öznitelik seçme teknikleri amaca yönelik araştırılan tüm modelleri ve bu modellerin sonuçlarını etkilemektedir. Bu sebeple mümkün olan tüm öznitelik seçme algoritmaları denenerek en uygun doğru sınıflandırma elde edilmiştir. Farklı öznitelik seçme algoritmalarından elde edilen en uygun özniteliklerin, sonuçları doğru vermede (Accuracy) çok önemli olduğu ortaya konmuştur. Ayrıca model değerlendirmesinde *Spesifiklik* ölçüsü gibi farklı ölçülerin kullanılabileceği ve bu ölçüler üzerinden doğru sonuçlar ve analizler yapılabileceği gösterilmiştir.

Destek vektör makineleri sınıflandırıcısı kullanılarak oluşturulan bir model üzerinden bir web uygulaması çalıştırılmıştır. Girdi olarak elde var olan tiroit kanser hasta verilerini alan web formu, modelin verdiği karar üzerinden hastanın kanserli olup olmadığı sonucunu web ara yüzü üzerinden ekrana yansıtmaktadır.

DIAGNOSIS OF THYROID CANCER USING DATA MINING AND MACHINE LEARNING APPROACH

SUMMARY

The health system can be defined as a system in which relevant institutions, resources and people are organized together to provide all health-related services to the targeted population. According to WHO, the health system includes people, actions and organizations to protect health, treat diseases, and ensure healthy development in individuals.

Today's Health services within the health system stand out with their "Medical Practice" aspect rather than "Medical Science". Today, the diagnosis is provided partly through the information provided by the patients, the patient's medical history and/or the symptoms of the disease/diseases they have (even if the patients cannot adequately convey the symptoms they have in an appropriate language). Diseases can be diagnosed based on information/experiences obtained from previous procedures on different patients or information obtained from old lectures in medical schools that doctors can partially remember. With exceptions, it is often possible to receive three different diagnoses, and treatment recommendations from three different doctors for a health problem. In this case, the clearest result is the delayed treatment, and the monetary and temporal costs for the patient.

Disease and death rates, which continue to increase every year, result in great expenses in terms of time, money and manpower costs. These rapidly increasing costs cannot be solved by employing more manpower. As a result of the development of technology and its yielding beneficial results in the field of medicine as in many fields, decision processes carried out by experts have the potential to be supported by intelligent decision support systems in order to solve the above-mentioned problems and increase efficiency by minimizing costs as much as possible.

In order for the field of Medicine in general, and health services in particular, to provide higher quality and more efficient services, they need to be supported through various systems and the costs of this level of services must be reduced. The health system, which is becoming more specialized day by day but cannot adequately meet our expectations of receiving quality health care, causes our health workers to be under more labor force. This situation necessitates the need to support the healthcare system, especially our specialist doctors and healthcare professionals. Once we obtain collective data about patients and start operating these data within the framework of Data Science by establishing appropriate systems and ensuring data compatibility between them, the knowledge-based expert systems to be used will make inferences on this data, allowing doctors to be faster, more effective and more efficient in their work.

Data Science and the techniques used to transform existing patient data into meaningful information are the key to the results we want to achieve. This key will

make healthcare systems successful by reducing costs and taking redundant work off doctors, providing patients with more care and more efficient treatment opportunities.

Data science owes its success to the use of multiple data and multiple options in diagnosis and treatment, and its erroneous results are minimal. On the other hand, it does not seem possible for many doctors to read, for example, thousands of articles published about the latest medical developments (to keep the data under control) and to make diagnosis and treatment by analyzing all the data they obtain.

According to statistical data, Cancer has been one of the leading causes of death worldwide, with an estimated 9.6 million deaths in 2018. Although Thyroid Cancer, which is not among the most common types of cancer in terms of the majority of cases, is not statistically at the top in terms of risk of death and disease, the difficulty of diagnosis and the inability to perform a biopsy compared to other types of cancer have brought different methods for diagnoses to the agenda. In this type of cancer, where a definitive diagnosis requires surgery, misdiagnosis results in unnecessary surgery. Sometimes, when a patient is diagnosed as having cancer and is biopsied after the surgery, it is concluded that he or she is cancer-free, while in some cases the opposite results occur. On this occasion, the need to use data mining and machine learning algorithms for thyroid cancer diagnosis has emerged.

Technological structures used as software and hardware under expert control can be supported by algorithms of the Machine Learning type, now known as Artificial Intelligence, which can produce decision outputs by self-learning. Based on this, after obtaining the necessary ethics committee permissions from the General Surgery Department of Sakarya University Research Hospital, we tried to create a model that can make diagnoses in various Machine Learning algorithms by using the thyroid cancer patient data we obtained.

The method proposed in this research has been Data mining and Machine learning approaches in modeling Machine learning methods such as Support vector machines and Logistic regression as Decision support systems, along with appropriate data collection (in our case, offline available data sets), data selection, data pre-processing operations.

The information obtained also includes different data mining processes, starting with visualizing and evaluating the results to be displayed in tables and diagrams. Available data for diagnosis and treatment have been transformed into information that can help decision-making processes at the expert level with Support vector machines, Logistic regression techniques and other machine learning algorithms.

Feature selection techniques affect all models researched for the purpose and the results of these models. For this reason, all possible feature selection algorithms were tried and the most appropriate correct classification was obtained. It has been demonstrated that the most appropriate features obtained from different feature selection algorithms are very important in providing accurate results (Accuracy). In addition, it has been shown that different measures such as the Specificity measure can be used in model evaluation and accurate results, and analyses can be made through these measures.

We ran a web application on a model we created using the support vector machines classifier. The web form, which takes the Thyroid cancer patient data as input, reflects the result of whether the patient has cancer or not based on the decision made by the model, on the screen via the web interface.

This model can help doctors as a decision support system to diagnose thyroid cancer. The model classifier we created in this section can also be used in other binary disease diagnoses.

1. GİRİŞ

1.1. Genel Bakış ve Problem Tanımı

Günümüz sağlık hizmetleri asıl itibari ile *Tıp Bilimi* olmaktan çok *Tıp Uygulaması* tarafı ile öne çıkmaktadır. Günümüzde tanı, kısmen hastaların verdiği bilgiler, hastanın medikal geçmişi ve/veya sahip oldukları hastalık/hastalıkların belirtileri üzerinden (her ne kadar hastalar, kendilerinde var olan belirtileri uygun bir dille yeterince aktaramıyor olsalar bile) sağlanmaktadır. Hastalıklar, daha önceki farklı hastalar üzerinden yapılan işlemlerde elde edilen bilgiler/tecrübeler veya doktorların kısmen hatırlayabildikleri, tıp okullarında derslerden elde edilen bilgiler üzerinden teşhis edilebilmektedir. İstisnalar hariç, bir sağlık problemi için çoğu zaman üç farklı doktordan üç farklı tanı ve tedavi önerisi almanız mümkündür. Bu durumda ortaya çıkan en net sonuçsa hasta açısından geciken tedavi ile birlikte parasal ve zamansal maliyetler olmaktadır.

Bilişsel yöntemler ve beraberinde tecrübesizlik, erken sonuç alma isteği (ilk bulgularla hemen harekete geçerek alternatif incelemeler yapmaksızın tanı sonucuna varma isteği) genel itibari ile yapılan tanı işlemlerinin pek çoğunu kapsamaktadır. Bu türden yanlış tanılar, yeniden tanı ve tedavi yapmaktan kaynaklanacak harcamaların yanı sıra, tazminat türü diğer harcamaları da beraberinde getirerek maliyetleri arttıracaktır. En uygun tanının konulması ve bu tanıya en uygun tedavinin uygulanması, verilerin interaktif olarak analiz edilmesi ve uyum sağlanması gerekmektedir.

Doktorlar, tanı ve tedavi söz konusu olduğunda taramalar, testler, geleneksel yöntemlerle yapılan muayeneler, reçete yazma, davranışlar ve belirtiler üzerinden nitelendirmelerde bulunmak gibi birçok işleme ihtiyaç duyarlar. Öte yandan bilgi parçacıklarının bir birleri ile uyum içinde olması ve tüm verilerin kullanılarak sonuca ulaşılmaya çalışılması teknolojiyi gerekli kılmaktadır. Bahsi geçen teknolojik yöntemler, tüm maliyetler ve elde edilen sonuçlar açısından geleneksel tanı ve tedavi yöntemlerinden daha iyi sonuçlar verdiği son zamanlarda yapılan araştırmalardan anlaşılmaktadır [1]. Doktorlardan beklenen çalışma, hasta ile ilgili tüm verilerin toplanması ve bu verilerin güncel tıbbi bulgulara ve hastaların geçmişine göre

değerlendirilmesi ve irdelenmesidir. Tanı ve tedavi ile ilgili tüm bu çalışmaları bilgisayarların ve bilgi sistemlerinin, donanım, yazılım, veri ve iletişimi sağlayan çoklu etmen yöntemlerini kullanarak en az hata oranı ile daha iyi yapabildiği ortaya konmuştur [2].

Yeni tıbbi gelişmelere zaman ayırmak açısından ortalama bir doktorun elinde bulundurduğu ve birikimlerine temel oluşturan bilgi ve bulgular genel itibari ile tıp fakültesinde elde ettiği bilgi ve tecrübeler üzerinden olmaktadır. Uzmanlık açısından bakıldığında bile, uzmanların kendilerini geliştirecek zamanı bulabilmesi halinde bile, insanoğlunun çok fazla irili ufaklı hastalıkları ile muhatap olması ve bunları hatırlayarak çözümler üretebilmesi mümkün görünmemektedir.

Genelde tıp alanı, özelde sağlık hizmetlerinin daha kaliteli ve verimli hizmet verebilmesi ve maliyetlerin düşürülmesi için çeşitli sistemler vasıtası ile desteklenmeleri gerekmektedir. Gün geçtikçe ihtisaslaşan ancak kaliteli sağlık hizmeti alma benzeri beklentilere yeterince cevap veremeyen sağlık sistemi, sağlık çalışanlarının daha fazla iş gücü altında kalmalarına sebep olmaktadır. Bu durum sağlık sisteminin, özellikle uzman doktorların ve sağlık çalışanlarının desteklenmeleri gerekliliğini zorunlu hale getirmektedir. Hastalar hakkında toplu veriler elde edildiğinde ve bu veriler uygun sistemler kurularak ve aralarında veri uyumu sağlanarak işlenmeye başlanıldığında, kullanılacak bilgi temelli uzman sistemler bu veriler üzerinden çıkarımlar yapabilecekler, doktorlar işlerinde daha hızlı, etkili ve verimli olabileceklerdir.

Bilgisayarlar, yazılım ve donanım olarak gerekli sistemin kurulması halinde parlak bir tıp uzmanından daha iyi olarak eldeki verileri, karmaşık olsalar dahi organize etme ve uygun verileri kullanarak işe yarar bilgiyi ortaya çıkarabilme kabiliyetine sahiptirler. Ayrıca hastalık belirtileri, hastalık geçmişi, hasta davranışları ve çevresel faktörler gibi birçok husus arasında bütünleşme ve denge kurma özelliklerine de sahiptirler [3]. Tüm bu avantajların elde ediliyor olması, teknolojik sistem kullanımının (Çoklu etmen sistemleri, uzman sistemler, veri madenciliği ve yapay zekâ teknikleri) ne kadar önem arz ettiğini ortaya koymaktadır.

Yukarıda adı geçen başta yapay zekâ teknikleri olmak üzere tüm diğer teknik ve metotlar, insani olandan kaynaklı eksikliklerin telafisinde, sıradan bir uzman

doktordan beklenen işleri yerine getirebilmekte ve hatta pratisyen hekimlerin uzmanlık açıklarını kapatabilecek oranda sonuçlar elde edebilmektedir [4].

Yapay zekâ temelli karar destek sistemlerinin oluşturulması ve mevcut sağlık sistemine bütünleştirilmesi bazı tereddütleri beraberinde getiriyor olsa da maliyet, zaman, tanının daha yüksek oranlarda doğru yapılması, ölüm oranlarının azaltılması ve benzeri birçok nedenden gerekli görülmektedir [5].

1.2. Araştırma Motivasyonu

Sağlık sistemleri yönetişim, liderlik, finansman, hizmet sunumu, insan gücü, ilaç, tıbbi teknoloji ve enformasyon türünden temel bileşenlerden oluşmaktadır. Bileşenlerin her biri için asıl hedef kaliteyi, etkinliği ve verimliliği arttırmaktır. Bileşenlerin her birinde hedeflenen sonuçların elde edilmesi, bütüncül olarak sağlık siteminde verimliliği, kaliteyi ve etkinliği arttıracaktır.

Teşhis süreci, eğer hastalık sık görülen ve bilinen bir duruma tekabül etmiyorsa, bilinenin aksine zor bir süreçtir. Türk TORAKS derneğine göre hasta, üzerinde 100 binden fazla hastalıklardan birini yâda daha fazlasını taşıma olasılığı ile muayene odasına girer [6]. ve doktor, hastayı muayene ederek ve çeşitli tetkiklerle bu olasılığı olabildiğince en aza indirmeye çalışır. Elemeler neticesinde elde kalan bariz bir teşhis söz konusu ise, doktor bu hastalığa yönelik tedavilere başlar. Belirtiler açısından yakın olasılıklarla birden fazla hastalık mevcut ise, daha fazla tetkik ve muayene ile daha fazla bilgi elde edilmeye çalışılır. Tetkikler tomografi, MR, PET, ultrasonografi yanı sıra, hastalar için daha sıkıntılı süreçleri barındıran endoskopi, bronokskopi, anjiyografi gibi tetkik yöntemler olabilir. Kesin teşhis için hastadan biyopsi yaptırması da istenebilir. Tüm bu süreçlerin defalarca her bir hastaya ayrı ayrı uygulanması hem uzman doktor hem de hastalar açısından eziyete dönüşebilmektedir. Uzmanların birey olarak, yorgunluk, unutkanlık ve bıkkınlık gibi insani zafiyetleri ve emeklilikle birlikte bilgi birikimlerini de beraberinde emekli ettikleri göz önünde bulundurulduğunda, sistemleştirilmiş ve veriye dayalı bir karar mekanizmasının varlığı önem kazanmaktadır.

İstatistiksel verilere göre kanser, 2018'de tahmini 9.6 milyon kişi ile dünya çapında ölümün önde gelen nedenlerinden biri olmuştur. Vakaların çokluğu açısından en yaygın kanser türleri Akciğer (2.09 milyon vaka), Meme (2.09 milyon vaka), Kolorektal (1.80 milyon vaka), Prostat (1.28 milyon dava), Deri (melanom dışı) (1.04

milyon vaka), Mide (1.03 milyon vaka) iken, ölümlere en çok neden olan kanser türleri ise Akciğer (1.76 milyon ölüm), Kolorektal (862 000 ölüm), Mide (783 000 ölüm), Karaciğer (782 000 ölüm), Göğüs (627 000 ölüm) kanser hastalıkları olmuştur [7]. Tiroit kanseri ise istatistiksel olarak ölüm ve hastalığa yakalanma riski olarak ilk sıralarda yer almasa da teşhis edilmesinin zorluğu ve diğer kanser türlerine göre biopsi yapılamaması tanı için farklı yöntemleri gerekli kılmaktadır. Kesin bir tanının ameliyat gerektirdiği bu kanser türünde, yanlış teşhis sonucu gereksiz bir ameliyat ile sonuçlanır. Bazen kanserli olduğu sonucuna varılan hastaya, ameliyattan sonra biyopsisi yapıldığında kansersiz olduğu sonucuna varılırken, bazı durumlarda da tam tersi sonuçlar ortaya çıkmaktadır.

Dünyada birçok ülkede olduğu gibi, maliyetlerin düşürülmesi temelinde ele alınan konulardan biri yapay zekâ destekli karar destek sistemleridir ve Türkiye'de de son zamanlarda üzerinde çalışılan önemli konulardan biri olmuştur. Ulusal bağlamda e-sağlık uygulamaları olarak başlayan karar destek distemleri bir adım daha ileri götürülerek yapay zekâ sistemleri ile bütünleşmeye başlamıştır. Şimdilik e-sağlık sistemleri veri paylaşımı ve harcanan zaman bağlamında verimlilik sağlanıyorsa da maliyet, kalite, hız, tanı ve tedavilerde kesinlik, doğru uzmanlığa yönlendirme, ölüm oranlarının düşürülmesi gibi konularda, yapay zekâ, makine öğrenmesi teknikleri ve veri madenciliği algoritmaları en uygun sonuçların elde edilebilmesinde umut verici gelişmeler sunmaktadır.

1.3. Araştırmanın Amaçları

Araştırmalarda elde edilecek bulguların yapay zekâ teknikleri ve veri analizi algoritmaları bileşimi içinde analiz edilmesi ile elde edilecek sonuçların, uzmanların tanı ve tedavi kararlarında olumlu yönde etkili olacağı öngörülmektedir. Birçok araştırma ve tezler bu öngörümüzü haklı çıkaracak yönde gelişmeler ortaya koymaktadır. WATSON, MYCN gibi projeler başarı ile uygulanan ve önemli sonuçların alındığı önemli yapay zekâ temelli projelerden bazılarıdır. Uygun yapay zekâ tekniklerinin keşfedilmesi ve beraberinde veri analiz algoritmalarının, veri madenciliği ve Derin öğrenme metotlarının kullanımı ile oluşturulan karar destek sistemleri e-sağlık sistemlerine büyük katkı sağlayacaklardır.

Sağlık hizmetlerinde karar verici uzmanların, tanı ve tedavi için karar vermede ellerindeki en önemli faktör kaliteli verilerdir. Verilerin varlığı tek başına maliyet ve

hız açılarından olumlu sonuçlar doğurmaz. Tanı ve tedavi kaliteli olabilir ancak verimli olmayabilir. Bunun nedeni elde bulunan verilerin en kısa zamanda ve en uygun veri ilişkileri ile bir sonuç ortaya koyamamasıdır. Veri tabanlarından veya veri madenciliğinden bilgi elde etmek, büyük veri tabanlarından uygun veri ilişkileri ve bilgiye dönüşen veri kalıpları çıkarmak demektir. Veri miktarı nedeniyle ve faydalı sonuçlar elde etme arzusu ile sistematik bir yöntem uygulanmalıdır. Kaliteli verilerin, kirli (açık ve net olmayan) verilerden daha fazla doğru sonuçlar vereceği bir gerçekliktir.

Makine öğrenmesi algoritmaları ve veri madenciliği teknikleri kullanılarak veri analizlerinin yapılması ve uygun sistemin modellenmesi amaçlanmaktadır. Bu vesileyle;

- Sağlık alanında, ihtiyaca binaen ortaya çıkan karar destek sistemleri içerisinde makine öğrenme algoritmaları ve veri madenciliği teknikleri ve bunların bütünleşmesi ile elde edilen sistemlerin hastalıkların tanı ve tedavisinde kullanımı bu tezin ana kapsamını ortaya koymaktadır. Karar verilen yapay zekâ alanının ve bu alan üzerinden oluşturulan tekniklerin (algoritmaların) kullanılması ile hayata geçirilen karar destek sistemleri, kanserli hastalıkları erken tanı ve tedavilerine hızlı ve etkili çözümler sunma görevini üstlenecektir. Bu tezin boyutları maliyet, kalite, çözüm için uygun metot, risk faktörlerini azaltma, hastalık sıklığının(frekans) en aza indirilmesi, yanlış tanının önlenmesi, hastalıkların öngörülmesi gibi başlıklar altında sıralanmaktadır.

- Sağlık teknolojileri hastalar, doktorlar ve sağlık kuruluşları açısından radikal şekilde değişimler göstermektedir. Geleceğin tıbbı daha yenilikçi, daha yaratıcı, daha hasta odaklı, daha dijitalleşmiş ve daha sürdürülebilir hale gelmektedir. Tüm bu sonuçları elde etmeye çalışan gelişmiş ülkelerden geri kalmamak ve uygulanabilir bir ön çalışma meydana getirebilmek adına geliştirilmiş ve uygulanabilir bir karar destek sistemi modelinin ortaya konulması.

- Hastalar için makine öğrenmesi modelleri tarafından tanı konulmak amacıyla kullanılacak verilerin, bir veri tabanında (Öğrenen makine gibi yapay zekâ teknikleri için bilgi tabanı olarak adlandırılır) oluşturulması.

- Verilerin makine öğrenmesi teknikleri tarafından kullanılması için karar destek modelinin yazılım ve donanım kullanılarak oluşturulması.

- Veri madenciliği süreçleri ile elde edilen veri guruplarının (kaliteli bilgi), tespit edilmesi, en uygun algoritma ile ortaya çıkarılması ve destek vektör makineleri ve lojistik regresyon gibi makine öğrenmesi teknikleri ile modellenmesi.
- Sağlık Hizmetlerinin gelişimine ve büyümesine katkı sağlaması.
- Sektör profesyonelleri için ve hizmet veren kurum ve kuruluşlarda daha iyiye ve kaliteliye ulaşma arzusu oluşturmak adına, yenilikçi teknolojik yaklaşımların ve işbirliklerinin desteklenmesi amaçlanmaktadır.

Ayrıca;

- Mevcut makine öğrenmesi algoritmalarının tiroit kanseri teşhisinde doğru sınıflandırmaya daha fazla katkıda bulunup bulunmayacağı,
- Farklı metotlarla seçilen farklı özniteliklerin kullanılması ile ortaya çıkan modelin tiroit kanseri ön tanısı doğruluk oranında artış yapıp yapmayacağı,
- Öğrenme odaklı sınıflandırma ve veri madenciliği tekniklerini kullanarak hastanın tiroit kanseri olup olmadığını ayırt etmek amacıyla veri setini en iyi temsil eden özniteliklerin olup olmadığı, varsa hangileri olacağının tespit edilmesi ve amaca en uygun ve en iyi modelin bulunup bulunamayacağı sorularına cevaplar aranacaktır.

1.4. Araştırmanın Katkısı

Araştırma, uygun koşullar altında ulusal ve uluslararası alanda sağlık sistemlerine de yardımcı olacak şekilde erken tanı ve tedavi aşamalarında karar vericilerin doğru karar vermelerine yardımcı olmayı hedeflemektedir.

Bu tezin amacı ve dolayısıyla amaca binaen katkıları, veri madenciliği ve makine öğrenmesi algoritmalarını kullanarak büyük ölçekli veriden anlamlı bilgiler çıkarmak üzere analizler yapmak ve amaca yönelik modeller oluşturmak olacaktır. Amaca yönelik ilgili işlemler, eldeki veri setinde kayıp verilerin tamamlanması, uygun ve kullanılabilir veri tiplerine dönüştürülmesi, tüm veri setini temsil edebilecek özniteliklerin seçimi için en uygun algoritmanın bulunması ve en iyi sonucu verebilecek sınıflandırma modelinin elde edilmesi ve mümkünse makine öğrenmesi temelli yeni hibrit modellerin tavsiye edilmesi başlıklarını içermektedir.

Makine öğrenmesi teknikleri kullanılarak yapılan uygulama ile tiroit kanseri hasta verileri üzerinden verilen ameliyat kararlarının kesine yakın olmasını sağlayacak katkılar sunulabilecektir.

2. KAYNAK ARAŞTIRMASI

2.1. Sağlık Hizmetleri ve Süreçleri

Sağlık hizmetleri insanlar için hastalıkların tanısı ve tedavisi yolu ile sağlığın korunması yada daha iyilenmesi/iyileştirilmesi için verilen hizmetlerin tümü için kullanılan bir kavramdır. Hastalık ve sağlık kavramlarının kapsadığı alanlarda süreci başlatır ve tamamlar. Sağlık hizmetleri verilen hizmeti kapsadığı gibi bu hizmetleri veren başta doktor ve hemşire olmak üzere diğer sağlıkla ilgili tüm meslekleri ve sektörleri kapsar. Ancak burada sağlık hizmetleri daha çok klinik düzeyde hastanelerde başta doktorlar, hemşireler, teknisyenler ve tüm diğer hastane görevlileri tarafından verilen hizmetlerin bütünü olarak ele alınacaktır.

Sağlık hizmetleri süreç açısından dört basamaklı sağlık hizmetleri yanı sıra halk sağlığı için yapılan çalışmaları da kapsamaktadır. Çeşitli sağlık süreçleri farklı kültürel, sosyo-ekonomik, örgütsel ve disipliner bakış açısına göre farklılıklar göstermesi yanı sıra ilk basamak sağlık hizmetlerinin yürütülmesi sağlık-bakım sürecinin ilk unsurunu oluşturur. Süreç ilk basamaktan sonra devem ederek 4. Basamağa kadar ilerlemektedir. Birinci basamak konsültasyon türünden en temel sağlık hizmetlerini içerirken, ikinci basamak kısa olsada ciddi bir hatalığın tanı ve tedavi sürecini içermektedir. Süreçteki üçüncü basmak daha çok yatan hastalar ile daha ileri tetkiklere ihtiyaç duyan yada daha modern techizatlara ihtiyaç nedeni ile sevk edilmek durumunda kalının hastalar için verilen hizmetleri içermektedir. Dördüncü bakım yada hizmet süreci daha yüksek seviyede uzmanlaşmış ve genelde kolaylıkla ulaşılamayan ileri tıp düzeylerinde ulaşılan hizmetleri temsil etmektedir. Bu dört basamaklı hizmetlerin dışında toplum sağlığı adıyla ortaya çıkan ve çoğunlukla gıda güvenliği, doğum kontrol önlemleri, bulaşıcı hastalıkların önlenmesi, aşılama türünden halk sağlığını ilgilendiren hizmetler de mevcuttur.

2.1.1. Sağlık ve hastalık kavramları

Bir birine yakın ancak farklı yaklaşımlarla anlamlandırılan ve tarif edilen sağlık kavramı Dünya sağlık örgütüne göre; yalnızca hastalığa sahip olmama yada sakat

kalmamış olma halini tarif etmeyip fiziki, sosyal ve zihinsel açıdan iyilik üzere olma durumunu ifade etmektedir [8].

Sağlık kavramı tarihi süreç içerisinde ihtiyaca binaen çeşitli ve ilgili disiplinler içinde farklı anlamlarla ifade edilse de her zaman hastalık kavramına karşıt bir pozisyonda yer edinmiştir. Sağlığın konumu, anlamı ve içeriği hastalık kavramının anlamı ve içeriğine bağlı olarak değişimlere uğramıştır. Hastalık'ın net bir hal üzere olmaktan çıkıp süreç durumuna dönüşmesi sözkonusu iken, sağlık ve kapsadığı alan da aynı şekilde net ve belirgin bir durumdan çıkarak bir sürece dönüşmüştür.

Dünya sağlık örgütüne göre, sağlıklı olmanın yada sağlığın ana belirleyicileri mevcuttur. Bunlar; sosyal, fiziksel ve ekonomik çevreler ile her bireyin bireysel özellikleri ve yaklaşımlarıdır [9].

Sosyal, fiziksel ve ekonomik açıdan iyi hal üzere olmak sağlıklı olmaya tekabül ederken, bu belirleyicilerin anormal ve keyif giderici olmaları hastalıklı olmayı beraberinde getirmektedir. Klinik düzeyde ve dar anlamda ele alındığında ise Hastalık, maddi ve manevi aykırılıkların iyi olma halini ortadan kaldıran durumlarını ifade etmektedir. Bu durumlar farklı adlar altında hastalık kelimesi ile bir arada kullanılarak sağlıklı olmayışın kök nedeni belirtilir. Kanser hastalığı, psikolojik yada ruhsal hastalıklar, fiziksel hastalıklar türünden tanımlamalar sağlıklı olmama hallerinin nedenlerini birlikte ortaya koymaktadır. Bu nedenler hastalıkların tanı süreçlerinde daha fazla detaylandırılır ve çeşitli tetkiklerle ispatlanmaya çalışılır.

2.1.2. Hastalık tanısı ve tedavi süreçleri

Tanı, kişide var olan hastalık belirtilerine ve bulgularına dayanarak kişinin hastalık tanımlamasının yada tıbbi durum analizinin yapılmasını ifade etmektedir. Tanı için hastaneye başvuran kişiden alınan ön bilgiler, tahlil ve tetkikler ile fiziki muayeneden elde edilen bilgiler gereklidir. Bilinmelidir ki tanı için gerekli bilginin çeşitliliği ve yoğunluğu bilginin alınacağı sağlık hizmeti basamağı ile doğrudan ilgilidir. Örneğin birinci basamak sağlık hizmetleri kategorisine giren hastalıklar için yapılacak tanılar ile ikinci basamak sağlık hizmetleri bağlamında konulmaya çalışılan hastalık tanıları için gerekli bilgi çeşitliliği ve yoğunluğu farklı olacaktır.

Tanı için temelde ihtiyaç duyulan bilgiler, fiziki muayene, tahliller, aile hikayesi, hastalardan alınan bilgilerden oluşurken elde edilen tüm bu bilgilerin hastalık çeşitlerine göre yöntemleri değişebilmektedir. Enfeksiyonel bir hastalıkta kan tahlilleri

ön plana çıkarken, fiziki bir rahatsızlıkta gözle muayene, film çekimi gibi yöntemler ön plana çıkabilmektedir. Benzer şekilde kanser ön şüphesiyle taraması yapılan hastalar için kan tahlilleri ile filmlerin sonuçları yanı sıra belirgin kanser hastalarında; örneğin tiroit kanser şüphesi olan hastalar için nodül araştırmasında kullanılan ince iğne anspirasyon yöntemi ile elde edilen sonuçlar da kullanılır.

Farklı hastalıklar için konulan tanılar neticesinde varılan kararlara göre tedavi yöntem ve süreçleri de farklılıklar göstermektedir. Çeşitli hastalıkların tedavilerinde kullanılan farklı yol ve yöntemler mevcuttur. Modern tıpta tedavilere yardımcı çeşitli araç ve gereçler yanı sıra ilaç tedavileri önemli bir yer tutmaktadır. Genellikle tanının net ve doğru konulduğu durumlarda (geç kalınmamış tanılar için geçerli olmak kaydıyla) tedavi süreçleri doğru ilerlemekte ve sonuç iyileşmeyle sonuçlanmaktadır.

Tanının doğru ve zamanında konulması tedavinin en önemli basamağını oluşturmaktadır. Tanı için halihazırda var olan yöntemler yanı sıra teknolojinin artan oranda kullanılması olumlu yönde katkılar sağlamaktadır. Özellikle verilerin tam ve güvenilir olduğu durumlarda, veriler arası bağlantıyı oratya koyan çeşitli yazılımsal karar destek sistemleri ortaya konulmuştur. Hastalık tanısında, karar destek sistemin yapay zekâ ile ilişkilendirilerek ve veri madenciliği yöntemleri kullanılarak sonuçlara varılması son yıllarda başvurulan önemli yöntemlerden biri olmuştur.

2.1.3. Kanser hastalıkları

Kanser vücuttaki hücre DNA'sının normal işleyişinin dışına çıkarak anormalleşmesi sonucu hızlıca ve kontrol dışı çoğalmaya başlaması durumudur [10].

Dış çevre veya kendi içinden gelen kanserojen madde etkisiyle yaş, cinsiyet ve beslenme gibi yan faktörlerin de katılımıyla oluşan kötü huylu tümörler vücudun farklı yerlerinde oluşması durumuna göre farklı isimler alırlar. Modern tıp alanında bu türden bir hastalığın kanser olarak adlandırılabilmesi için şu beş şartı yerine getirmiş olması gerekmektedir.

- İlgili hücrelerin sınırsız çoğalma eğilimi göstermeleri
- Kanserli bölge hücrelerinin büyümesine mukavemet gösteren faktörlere olumlu cevap vermemesi
- Normal hücreler gibi ömrünü tamamlamaması – ölümsüz olması
- Damarlardan tomurcuklanma yöntemi ile yeni damar oluşumunu tetiklemesi

- Dokulara nüfuz ederek çoğalması [11].

Mikroskopta incelendiğinde normal hücrelere benzemediği görülen kanserli hücreler bağışıklık sistemi tarafından yok edilerek vücut dışına atılamazlarsa çoğalarak vücudu ele geçirmeye çalışırlar. Çoğalmaları açısından çok yavaş yada hareketsiz olarak bulunan ve iyi huylu tümörler olarak adlandırılan kanserli hücreler zaman içinde kötü huylu kanserli hücrelere dönüşerek yayılım gösterebilirler. Bu tür tümörlerin zaman içinde gözlem altında tutulmaları uygun olacaktır [12].

Dünyada kanser türleri içerisinde akciğer kanseri kanserli hastalıklar içinde %20 oranı ile başı çekmektedir. Bunun yanı sıra %9 ile karaciğer, %9 ile kolon, %9 mide, %7 meme ve % 47 ile diğer kanserler mevcuttur [13].

Farklı kanser türlerinde farklı tanı koyma ve tedavi etme yaklaşımları mevcuttur. Farklı olsa da genel itibari ile tarama testleri ile elde edilen belirtiler üzerinden yada mümkünse parça alımının biyopside değerlendirilmesi ile ortaya çıkan sonuçlar ile tanı konulur. Tanı kesinleşirse kanser hastalığının türüne göre ameliyat, ilaçlı tedavi, ışın tedavisi veya kemoterapi gibi tedavi yöntemlerine başvurulur [14].

Kanserli hastalıklar arasında diğer kanser türlerine oranla daha az rastlanılan ancak endişe kaynağı olmaya devam eden kanser türlerinden biri de tiroit kanseridir.

Tiroit kanseri, tiroit bezinin dokularından meydana gelen bir kanser türüdür. Anormal hücrelerin tiroit bezi dokularında hızlıca çoğalması neticesinde vücudun diğer bölgelerine de dağılma eğilimi gösterebilirler. Erken yaşta radyasyona maruz kalma, tiroid bezinde büyüme eğilimi, aile hastalık geçmişi ve aşırı kilo bu hastalık için risk faktörleri arasında yer alır. Tiroit kanserinin dört türüne rastlanır; bunlar papiller tiroit kanseri, foliküler tiroit kanseri, medüller tiroit kanseri ve anaplastik tiroit kanseridir. Tanı çoğunlukla ultrason sonuçları ve ince iğne aspirasyonu (İİA) yöntemleri ile konulmaktadır. Tedavi için cerrahi yöntemler, radyasyon uygulaması, kemoterapi, hormon takviyesi uygulanmaktadır [15].

Tiroit kanseri tiroit hastalıklarından farklıdır. Kliniğe başvuran ve daha sonra kanser tanısı konulan hastaların şikayetleri; boyun bölgesinde şişlik, nefes darlığı, ses kısıklığı, yutma güçlüğü, boyun ve kulak bölgesinde ağrılar ve yorgunluk iken, kanser dışındaki tiroit hastalarının şikayetleri; sinirlilik hali, uyku problemleri yaşama, kilo kaybetme, güçsüzlük, titreme, adet düzensizliği ve ısıya karşı hassasiyettir.

Tanı sırasında tiroit kanseri için tiroit hastalıklarında kullanılan kan tahlilleri gibi sonuçlar üzerinden gidilmez. Ön fiziki muayeneden sonra İİA yöntemi ile tiroit bezinden hücre rezeke edilerek elde dilen hücreler laboratuvar ortamında incelenir. Elde edilen bulgular pozitif ise cerrahi müdahale gerçekleştirilir. Bazı durumlarda bu yöntemle rezeke edilen hücreler doğru sonucu veremeyebilmektedir. Bu durumda eldeki daha fazla veri üzerinden farklı yorumlamalarla tanı konulmaya çalışılmaktadır. Tiroit kanser tanısı hastanın kanserli olduğu yönünde ise, hasta ameliyat edilerek tiroit bezi alınmaktadır.

Sakarya Üniversitesi Genel Cerrahi bölümünde görev yapan operatör doktorlardan alınan bilgilere göre; ameliyat sonrası genelde hastaların şikayetleri ortadan kalkar. Ameliyat edilen hastalar ameliyat sonrası altı ayda bir olmak üzere kontrollere çağrılmaktadırlar. Tiroit bezleri alındığı için hastalara hormon ilaçları başlanmaktadır. Altı aylık periyotlarda ilaçların dozu kontrol edilerek ilaç doz düzenlemesi yapılmaktadır. Tiroit bezi dışında vücutta hormon üreten başka bir organ olmadığından bu bezin yokluğu halinde gerekli hormon replasmanı yapılmaktadır. Hormon eksikliklerini gidermek amacıyla kullanılan ilaçların yokluğu halinde hipotiroidiye bağlı halsizlik, saç dökülmesi, üşüme ve kabırlık gibi belirti ve bulgular ortaya çıkmaktadır.

Vücuttaki ana işlevi, vücuda alınan gıdaların enerjiye dönüştürülmesi sırasında metabolizmanın hızını dengelemek olan tiroit bezi T3 ve T4 gibi hormonlarla da metabolizmanın dengesini sağlar. Tiroit bezinin varlığı özellikle vücudun hormonal dengesi için çok önemlidir ve ameliyatla alınması halinde vücut için önemli bir kayıp anlamına gelmektedir. Ameliyat sonrası ihtiyaç duyulan hormonal denge ilaçlarla sağlanıyor olsa da ilaçların ömür boyu reçetelendirilmesi ve vücuda verdiği yan etkileri göz önüne alındığında ameliyat kararlarının isabetli verilmesi ve yanlış ameliyatların yapılmaması çok önem arzetmektedir. Bu önemli nedenlerle, yanlış ameliyatların önüne geçmek amacıyla, makine öğrenmesi yöntemleri türünden mümkün olan tüm araçların kullanılması elzem görünmektedir.

2.2. Endüstri Mühendisliği

Endüstri mühendisliği kullanıcı, hammadde ve araçlardan oluşan sistemlerin kurulması ve bu sistemlerin işlem sürekliliğinin çeşitli yöntemlerle sağlanması ile ilgilenen müdendislik dalıdır. Endüstri mühendisliği diğer tüm mühendislik dallarının

bir çoğunun derslerini de vermekle birlikte işletme, yönetim ve ekonomi gibi dallardan da eğitimler içerir. Diğer tüm alanlar için yöneticilik vasıflarına sahip olacak şekilde bu mühendisliğe özel dersler ve eğitimler de verilir. Farklı alanlardaki çalışmalarına matematik, bilgisayar bilimleri, doğa bilimleri ve sosyal bilimleri de dahil ederek çeşitli analizler ve tasarımlar ortaya koyar. Tüm bu süreçlerde sonuç odaklı hizmetlerde verimle birlikte kaliteyi arttırmayı ve dolayısı ile maliyetleri en aza indirmeyi hedefler. Her türlü uygulamada verim ve kaliteyi hedefe koyarken sadece malzeme ve araçları değil aynı zamanda insan faktörünü de azami derecede dikkate alır ve böylelikle sosyal bilimlerle de etkileşim içinde olur.

Yöneylem araştırması, üretim sistemleri, modelleme ve optimizayon gibi önemli dersler ihtiva eden endüstri mühendisliği aynı zamanda yönetim bilişim sistemleri, bilgisayar programlama derslerini de içermektedir. Bu türden derslerle birlikte endüstri mühendisliği bilgisayar mühendisliği, bilişim bilimleri, ve bilgi teknolojilerinin alt bileşenlerinden olan karar destek sistemleri, yapay zekâ ve makine öğrenmesi gibi farklı model ve yöntemlerle bir arada çalışmalar yapma olanağı bulmaktadır.

Sağlık bilimlerinde giderek artan şekilde bilişimin, bilgisayar bilimlerinin ve yöntemlerinin kullanımı endüstri mühendisliğinin bahsi geçen bu disiplinler üzerinden sağlık bilimleri ile ilgili çalışmalar yapmasına yol açmıştır. Endüstri mühendisliğini genel manada problemi tanımladıktan sonra çözüm türetme ve karar verme süreçlerini takip eden bir mühendislik dalı olarak ele alındığında, sağlık sitemlerinde süreçlerin analizi, tasarımı, hastalık tanı ve tedavilerinde karar destek sistemlerinin oluşturulması beklenen sonuçlar arasında olmaktadır. Veri toplama ve analiz yöntemleri kullanarak sonuç çıkarma işlemlerini en iyi şekilde ortaya koyan endüstri mühendisliği, sıkı bir ilişki içinde bulunduğu Yönetim bilişim sistemleri üzerinden yapay zekâ ve alt bileşeni olan makine öğrenmesi yöntemlerini kullanarak hastalık tanı ve tedavilerinde karar destek sistemlerinin sağlık hizmetlerinde yerini almasını sağlamaktadır.

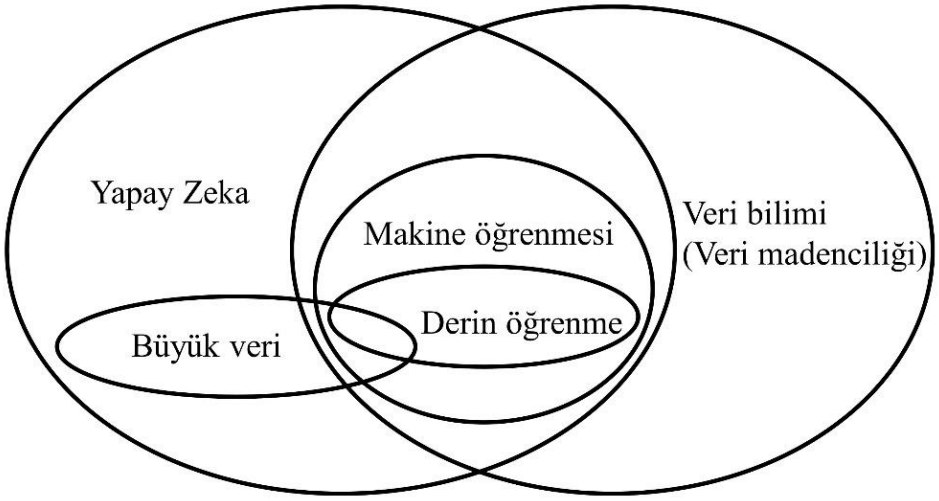
2.3. Hastalık Tanısında Yapay Zeka Kullanımı ve Diğer Disiplinlerle İlişkisi

Yapay zekâ, bilgisayar bilimlerinin gelişmesi ile ortaya çıkan ve zeki makine olarak ta adlandırılan, insanlarda var olan doğal zekâya nispetle aynı sonuçları vermek üzere tasarlanmış algoritmik bilgisayar sistemleri olarak tarif edilirler.

Yapay zekâ terimi genellikle öğrenme ve problem çözme gibi insanların insan zihniyle ilişkilendirdiği bilişsel işlevleri taklit eden makineleri (veya bilgisayarları) tanımlamak için kullanılır. Yapay zekâ kavramı zeki olarak tarif edilebilecek her bir görevle ilgili ve ilişkilidir [16].

Modern yapay zekâ (YZ) tekniklerinin uygulandığı alanlar yaygındır [17], ve burada listelenemeyecek kadar fazladır. YZ, bilgisayar bilimi alanı içine giren ve diğer disiplinleri de ilgilendiren zor problemleri çözmek için birçok araç ve yöntem geliştirmiştir.

Son zamanlarda sağlık alanında, tıp uygulamalarında karar destek sistemleri olarak kullanılan yazılımlar ve bu yazılımlar üzerinden uygulamaya konulan teknik ve algoritmalar yapay zekâ uygulamaları olarak sistemleştirilmiştir. Yapay zekâ uygulamaları birçok alanda kullanılmakla birlikte, sağlık alanında da artan oranda araştırmalara konu olmakta ve sağlıkta tanı-tedavi için destek sistemleri olarak yerini almaktadır. Yapay zekâ daha genel bir alanı ifade den kavram olması nedeniyle, tüm işlemlerini kendi alt bileşenleri üzerinden ve/veya diğer disiplinlerle birlikte yürütür. Yapay zekâ alanının diğer alt bileşenler ve disiplinlerle ilişkisi Şekil 2.1 ve Şekil 2.2'de gösterilmiştir.



Şekil 2.1. Yapay zekâ ve diğer disiplinlerle ilişkisi.

Şekil 2.2. Yapay zekânın diğer bilimlerle ilişkisi.

Sağlık sistemlerinde yapay zekâ kullanımı sadece hastalık tanısı koyma ile sınırlı olmamakla birlikte sağlıkta yapay zekâ kavramı, daha çok ilaç ve hastalık ile ilgili çok fazla ve compleks olan verilerin yeni yöntemlerle açıklanması, analiz edilmesi ve varsa eksiklerin tamamlanması amacıyla, yazılımların ve algortimaların kullanımıyla ilişkilidir. Yapay zekâ yaklaşımı ile oluşturulan tanı ve tedavi aplikasyonlarında temel amaç hastalık tanısı çıktıları ile girdi verileri arasındaki net ilişkiyi ortaya koymaktır.

Hastalık tanı ve tedavilerinde yapay zekâ kullanımı söz konusu olduğunda kastedilen çoğunlukla makine öğrenmesi algoritmalarıdır. Bu yaklaşımla makine öğrenmesi algoritmaları üzerinden yapılan hastalık tanı ve tedavi çalışmaları aynı zamanda yapay zekâ çalışmaları olarak ta ele alınabilir.

Yapay zekâ teknolojisini sağlık hizmetinde kullanılan geleneksel teknolojilerden ayıran durum, veri edinme, veriyi işleme ve son kullanıcılara iyi tanımlanmış bilgiyi çıktı olarak verme yeteneğidir. YZ bunu makine öğrenme algoritmaları ve Derin öğrenme metotları yoluyla yapar. Bu algoritmalar işleme konmuş kalıpları tanıyabilir ve bu sonuçlar üzerinden kendi mantık örgülerini oluşturabilir. Hata payını azaltmak için, YZ algoritmaları tekraren test edilmeleri gerekir.

YZ algoritmaları insanlardan iki şekilde farklı davranır: (1) algoritmalar değişmezler. Bir hedef belirlediğinizde, algoritma kendini bu amaca yönelik otomatik olarak ayarlayamaz ve sadece açıkça söyleneni ve verilen/öğretilen yolu anlayabilirler. (2)

Bazı derin öğrenme algoritmaları, giren veri ve çıkan veri üzerinden tanımlanırlar; Algoritmalar, verilen veriler arasında tanımlanan ilişkiler üzerinden son derece kesine yakın olacak şekilde sonuçlar çıkartarak tahminde bulunabilirler, ancak bu sonuçların nedenleri üzerinde tahminler yürütemezler [18].

Sağlıkla ilgili YZ uygulamalarının temel amacı, önleme veya tedavi teknikleri ile hastalık sonuçları arasındaki ilişkileri analiz etmektir. YZ programları, teşhis süreçleri, tedavi protokolü geliştirme, ilaç geliştirme, kişiselleştirilmiş ilaç ve hasta izleme veya bakım gibi uygulamalara yönelik olarak geliştirilmiş ve uygulanmıştır [19]. Mayo Kliniği, Memorial Sloan Kettering Kanser Merkezi [20, 21], ve İngiliz Ulusal Sağlık Servisi [22], gibi departmanlar için yapay zekâ algoritmaları geliştirilmiştir. IBM [23], ve Google [24], gibi büyük teknoloji şirketleri de sağlık hizmetleri için YZ algoritmaları geliştirmişlerdir. Buna ek olarak, hastaneler maliyet azaltan, tasarruf ettiren, hasta memnuniyetini arttıran ve personel ve işgücü ihtiyaçlarını karşılayan operasyonel girişimleri desteklemek üzere YZ yazılımları kullanma arayışlarına girmişlerdir [25]. Şirketler, sağlık yöneticileri için yapay zekâ destekli modeller kullandırarak, hasta başvurularını ve hasta kalış sürelerini azaltmak, personel sayılarını optimize ederek iş operasyonlarını iyileştirmelerine yardımcı olmak üzere öngörü sunan ve analiz yapabilen çözümler geliştirmişlerdir [26].

Yapay zekâ algoritmaları kullanılarak, koroner arter hastalığı doğru bir şekilde teşhis edilmiş ve risk sınıflandırmasında umut verici sonuçlar elde edilerek triyaj aracı olarak potansiyel yararları gösterilmiştir [27, 28]. Bu sonuçlar umut verici olsada az sayıda çalışma, makine öğrenmesi modellerinin doğruluğunu ve kalitesini doktorların tanı koyma yeteneğiyle doğrudan karşılaştırmışlardır [29]. Gelişen bir diğer araştırma alanı da yapay zekânın kalp atımlarını/seslerini sınıflandırmada ve kalp kapakçığı hastalıklarının tanısında kullanılmasıdır [30].

Dermatoloji alanı görüntüleme verilerinin çokça kullanıldığı bir uzmanlık alanıdır. Görüntü verilerinin derin öğrenme yöntemleri ile çok iyi işlenebiliyor olması ve tanı için çıkarımlarda bulunulması bu hastalığın tanısında yapay zekânın kullanımını çekici kılmıştır. Dermatoloji hastalık tanısında derin öğrenme kullanılarak elde edilen sonuçlar [31]. gelecek vadetmektedir.

Yapay zekâ kullanılarak Gastroloji alanında da çeşitli klinik sonuçlar elde edilmiştir. Mide kanserinin erken safhasında yapay zekânın tanı için kullanılmasına yönelik ilk

denemeler, çıktı verileri üzerinden yapılan doğru sınıflandırmada hassasiyet (Sensitivity) ölçümleri açısından uzman endoskopistlerin sonuçları ile çok yakın olduklarını göstermiştir [32].

Ayrıca yapay zekâ teknikleri kullanılarak enfeksiyon hastalıkları için üretilen ilaçlarda uygun doz ayarlamaları yapılmıştır [33].

Yapay zekâ ayrıca kas-iskelet sistemi içinde doktorların gözden kaçırdığı ve Afrika kökenlileri daha fazla etkileyen diz ağrısının nedenlerinin belirlenmesinde [34], nöroloji alanında alzaymer'a ön tanı konulmasında [35], onkoloji alanında kanserli hastalığın tanısında [36], göz hastalıkları tanısında [37], patolojide [38], psikiyatride [39], bilgisayarlı tomografi (BT) ve manyetik rezonans (MR) görüntüleme yoluyla elde edilen veriler üzerinden hastalıkları tespit etmek üzere radyoloji alanında [40], kullanılmıştır.

Jiang ve arkadaşları [41], farklı hastalıklar için kullanılan çeşitli YZ tekniklerinin olduğunu gösterdiler. Bu tekniklerin bazıları destek vektör makineleri, yapay sinir ağları, karar ağaçları başta olmak üzere çeşitli hastalıkların tanı ve tedavilerinde destek sağlamak üzere kullanılan tekniklerdir.

Araştırmacılar tarafından otomatik teşhis sistemleri geliştirmek için en sık kullanılan ve bir kısmı makine öğrenmesi alt dalına dahil olan YZ tekniklerinin bir kısmı aşağıdaki gibi sıralanabilir:

- Yapay sinir ağları [42]
- Bulanık mantık [43]
- Destek vektör makineleri [44]
- Genetik algoritmalar [45]
- Derin öğrenme olarak sıralanabilir.

2.4. Makine Öğrenmesi

Makine öğrenmesi, sensörlar ya da veri tabanları aracılığı ile elde edilen veriler üzerinden öğrenmeyi mümkün kılan, algoritmaların oluşumunu ve kullanılmasını konu edinen, yapay zekânın bir alt dalıdır.

Makine öğrenmesi, deneylerle elde edilebilen sonuçların, matematiksel, istatistiksel ve yapay zekâ teknikleri ve veri madenciliği yöntemleri kullanılarak, birbiri ile ilişkili veriler üzerinden otomatikleştirilmesi işlemidir. Bu işlem ile çeşitli algoritmalar kullanılarak tekraren kullanılmak üzere örüntüler elde edilebilir.

Makine öğrenmesi çeşitli alanlarda farklı çözümler elde etmek üzere farklı algoritmalar geliştirebilirler. Bu farklılık, makine öğrenmesi yöntemlerinin olasılık, istatistik, matematik, yapay zekâ, veri madenciliği, örüntü tanıma, optimizasyon gibi diğer bilim dallarıyla da ilişki içinde olmasını gerekli kılmaktadır [46].

Makine öğrenmesi yöntemlerinde amaç ve odaklanılan çerçeve, makinelere (bilgisayarlara), insanlar için karmaşık görünen örüntüleri algılama ve veri kullanarak tutarlı ve işe yarar kararlar verebilme yetisi kazandırmaktır. Makinenin verilerden öğrenmesi olarak özetlenebilen bu yöntem, öğrenme işlemlerini gerçekleştirirken, yaklaşım olarak çeşitli algoritmalar kullanmaktadır. Çözmek istenilen probleme ve kullanılacak verinin türü ve özelliğine göre değişebilen algoritmaların dahil olduğu ve makine öğrenme tipleri (ML type) olarak ele alınan önemli başlıklar şu şekilde sıralanabilir:

- Denetimli öğrenme
- Denetimsiz öğrenme
- Yarı-denetimli öğrenme
- Takviyeli öğrenme
- Kendi kendine öğrenme

Makine öğrenmesi verileri kullanarak veri seti içinden çeşitli örüntüler elde edebilirler. Bu örüntüler üzerinden önceki veriler kullanılarak yeni çıkarımlar yapabilirler ve yeni kararlar ortaya koyabilirler. Veri odaklı çalışan makine öğrenmesi, verilerden daha sonra kullanmak üzere örüntüler elde etmesi ve bu bilgileri daha sonra tekrar kullanma kabiliyeti dışında veri madenciliği yöntem ve yaklaşımlarına da sahiptir.

Makine öğrenmesi otomatik dil tercümesi, hastalık tanısı koyma, hisse senedi fiyat tahmini yapma, çevrimiçi dolandırılıcılık tespiti, sanal kişisel asistanlık, kötü amaçlı elektronik posta filtrelemesi, kendini süren arabalar, trafik yoğunluğu tahminleri, ses ve konuşma algılama, görüntü tanıma gibi bir çok alan ve projede kullanılmaktadır.

Makine öğrenmesi denetimsiz öğrenme tekniği altında kümeleme algoritmasını kullanırken, denetimli öğrenme tekniği altında ise sınıflandırma ve regresyon algoritmalarını kullanır.

Sınıflandırma tekniği altında destek vektör makineleri, karar ağaçları, k-en yakın komşu, naïve bayes, lojistik regresyon, yapay sinir ağları ve derin öğrenme algoritmaları yaygın olarak kullanılmaktadır.

2.4.1. K-en yakın komşular algoritması

K-en yakın komşular (KNN) algoritması hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılabilen, uygulaması kolay, basit bir denetimli makine öğrenme algoritmasıdır.

KNN algoritması, benzer şeylerin birbirine yakın yerlerde bulunacağını varsayar. Diğer bir deyişle, benzer şeyler birbirine yakındırlar.

K- En yakın komşu veri sınıflandırma algoritması, tüm mevcut durumları muhafaza eden ve yeni durumları benzerlik ölçülerine (Bu benzerlikler veriler üzerine uygulanan mesafe fonksiyonları ile elde edilir.) göre sınıflandıran basit bir algoritmadır. Benzerlik ölçülerinde kullanılan fonksionlar, kullanılan verinin tipine göre belirlenirler. Veriler sürekli değişken veri tipinde ise, k-en yakın komşu algoritmalarında mesafeler üzerinden veriler arası benzerlikler euclidean, manhattan, minkowski mesafe fonksiyonları ile ölçülür. Kategorik değişken veri tipi söz konusu olduğunda hamming mesafe fonksiyonu kullanılmalıdır. Eldeki veriler üzerinden doğru olan k'yi seçmek için, KNN algoritması farklı k değerleriyle birkaç kez çalıştırılmalıdır ve algoritmanın sahip olduğu veriler üzerinden doğru tahminler yapma durumu devam ediyorken her çalıştırılmada giderek azalan hata sayısında en az hatayı veren k seçilir. Diğer bir değişle, sınıflandırma veya regresyon söz konusu olduğunda, veriler için doğru k'yi seçmek hedefi ile bir çok k denenmeli ve en iyi k seçilmelidir.

2.4.2. Yapay sinir ağı

Yapay sinir ağları (YSA), insan beyninin sınıflandırma şeklini taklit etmeye çalışan makine öğrenme modellerindendir. Nöral ağ birçok farklı nöron katmanından oluşur, her katman önceki katmanlardan girdi alır ve çıktıları başka katmanlara iletir. Her katman çıktısının bir sonraki katman için girdi haline gelmesi, maliyet işlevine veya optimize edici duruma bağlı olarak, söz konusu bağlantıya verilen ağırlıkla doğrudan ilişkilidir. Sinir ağı, önceden belirlenmiş sayıda yineleme için tekrarlar yapar. Her

tekrarın ardından modelin nerede geliştirilebileceğini görmek için maliyet fonksiyonu analiz edilir. En iyi sonucu veren işlev için, maliyet en aza indirilene kadar, maliyet işlevi tarafından sağlanan bilgilere dayanarak ağın ağırlıkları değiştirilir.

Bu bağlamda, bir sinir ağı sınıflandırma problemlerinin çözülmesine yardımcı olabilecek çeşitli makine öğrenme algoritmalarından birini teşkil etmektedir. Başka algoritmaların yapamayacağı şekilde tahmin işlemlerini dinamik olarak oluşturması ve insan düşüncesini taklit etme yeteneğine sahip olması en önemli özelliklerindendir.

2.4.3. Karar ağacı

Karar ağacı algoritması istatistik, veri madenciliği ve makine öğrenmesinde kullanılan öngörülü modelleme yaklaşımlarından biridir. Bir ögeyle ilgili (dallarda temsil edilen) gözlemlerden, ögenin hedef değeri (yapraklarda temsil edilen) ile ilgili sonuçlara gitmek için bir karar ağacı (tahmini model olarak) kullanır. Hedef değişkenin ayrı bir değer kümesi alabileceği ağaç modellerine sınıflandırma ağaçları denir; bu ağaç yapılarında yapraklar sınıf etiketlerini ve dallar bu sınıf etiketlerine yol açan özelliklerin birleşimlerini temsil eder. Hedef değişkenin sürekli değerler alabileceği karar ağaçlarına regresyon ağaçları denir. Karar ağaçları, anlaşılırlıkları ve basitlikleri göz önüne alındığında en popüler makine öğrenme algoritmaları arasındadır [47].

2.4.4. Naive bayes sınıflandırıcı

Bayes algoritmasına dayanan olasılık tabanlı bir sınıflandırıcıdır. Bağımlı olasılık kavramına göre, bir veri noktasının özelliklerinin (giriş değişkenlerinin) her birinin hedef sınıfların her birinde tek tek bulunma olasılığını hesaplar. Daha sonra olasılıkların maksimum olduğu kategoriyi seçer. En basit bayes ağ modelleri arasındadır [48].

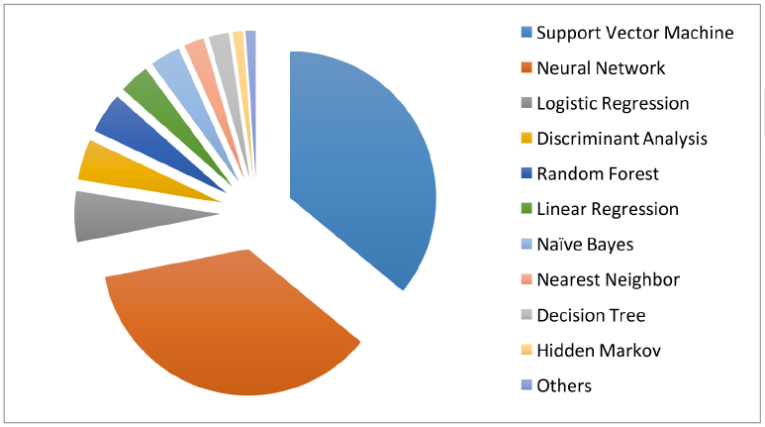
Naive bayes sınıflandırıcıları, yüksek düzeyde ölçeklenebilir niteliktedirler ve bir öğrenme probleminde değişken sayısı (özellikler / öngörücüler) içinde doğrusal bir dizi parametre ihtiyacı hisseder. Bilinmeyen veri kümelerinin sınıfını tahmin etme konusunda bayes olasılık teoremi üzerinde çalışır.

2.5. Hastalık Tanı ve Tedavilerinde Makine Öğrenmesi Kullanımı

Kimi uzmanlara göre uygulamalar henüz sonuç elde etme açısından erken aşamalarda olsa da makine öğrenmesi, hastalık tanısı sonuçlarında giderek artan doğru isabetler

elde etme, maliyetten tasarruf etme ve bunlardan daha önemlisi hayat kurtarabilme potansiyeline sahiptir.

Makine öğrenme algoritmaları hastalıkları daha erken safhalarında tespit edebilirler. Örneğin, makine öğrenmesi modelleri, tümörleri tespit etmek, ölçmek ve analiz etmek için tıbbi görüntülerden elde edilen verileri kullanarak kanser teşhisine yardımcı olabilirler. Veri ve görüntü analizini uzman doktorlardan daha hızlı gerçekleştirmek için işlem hızı avantajını uygulayan makine öğrenmesi, taramaları daha kısa sürede tamamlayabilir. Bu durum sevk bekleme sürelerini azaltabilir ve personel yetersizliği veya diğer zorluklarla karşı karşıya kalan muayenehanelerin yükünü azaltabilir. Yıllar içerisinde makine öğrenmesi algoritmaları kullanılarak yapılan çalışmalar üzerinden elde edilen istatistiksel bilgilere göre hastalık tanı ve tahminlerinde kararlara destek için kullanılan algoritmaların kullanım ve tercih sıklıklarına göre sıralamaları Şekil 2.3'te verilmiştir.



Şekil 2.3. Medikal literatürde kullanılan algoritmalar (PubMed).

2.5.1. Makine öğrenmesi kullanılarak yapılmış hastalık tanısı çalışmaları

Yapay zekânın bir alt dalı olan makine öğrenmesi (MÖ) finansal, çevresel, pazarlama, güvenlik, üretim, endüstriyel ve tıbbi uygulamalar başta olmak üzere bir çok alanda karmaşık problemleri çeşitli algoritmalarla çözme yeteneğine sahiptir. MÖ bir çok hastalığın tanısında ve hastalık sonucu tahminlerinde çokça başvurulan bir karar destek sistemi olmuştur.

Tıbbi görüntüleme analizi yapma ve kanser tanısı koyma gibi işlemlerde çözümler üretmiştir [49]. Bilgisayar destekli tanı yada karar destek sistemleri adlarıyla anılan uygulamalarda çokça yararlanılan makina öğrenmesi, çeşitli algoritmalar kullanarak

hastalık tanısı koyma aşamasında, doğruluğun ve güvenilirliğin arttırılmasında doktorlara yardımcı olmuşlardır [50].

Wei ve ark. [51], dijital memogramlar ile kümelenmiş mikrokalsifikasyonların sınıflandırılmasında makine öğrenmesi uygulamalarını araştırmışlardır. Çalışmada destek vektör makineleri (DVM) ve ileri beslemeli yapay sinir ağları (İBYSA) teknikleri kullanılmıştır. Görüntü özelliklerine dayanarak bir mikrokalsifikasyon kümesinin iyi yada kötü huylu olup olmadığını araştırmışlardır. 697 klinik memogramdan oluşan veriler kullanılarak elde edilen doğruluk oranları sırasıyla DVM için %85,45 ve İBYSA için %80,07 olarak bulunmuştur.

Ing ve arkadaşları 530 hastaya ait 10 adet girdi verisi üzerinden yaptıkları çalışmada atardamar biyopsi sonuçlarının tahminini yapmak üzere destek vektör makineleri algoritması kullanarak %82,7 ve lojistik regresyon algoritması kullanarak % 82,5 doğruluk sonuçları elde etmişlerdir [52].

Ambrish ve arkadaşları Kaliforniya Üniversitersi UCI veri tabanından aldıkları Kalp-damar hastalarına ait verileri kullanarak tanı tahmini yapmak üzere lojistik regresyon algoritmasını farklı eğitim ve test oranları ile çalıştırarak en yüksek % 87,10'luk doğruluk değeri elde etmişlerdir [53].

Khaleel ve Al-Bakry Pima Indian diyabet hasta verilerini kullanarak yaptıkları tahmin çalışmasında lojistik regresyon, naïve bayes ve k-en yakın komşu algoritmalarını kullandılar ve en iyi sonucu %94 kesinlik oranı ile lojistik regresyon ile elde etmişlerdir [54].

2.5.2. Makine öğrenmesi kullanılarak yapılmış kanser tanısı çalışmaları

Kanser araştırmaları toplumsal etkisi yüksek önemli bir alandır. Makine öğrenmesi, metodlarının ve algoritmalarının kanser araştırmalarında karar destek sistemleri olarak kullanılmaları, tanı ve tedavide sınıflandırmalar yaparak doktorlara yardımcı olmaları açılarından büyük kabiliyetlere sahiptirler [55, 56].

Podolsky ve ark. [57], akciğer kanseri için veri analizi ve kanser tanısı koyma çalışmaları yaptılar. Yapılan bu çalışmada kamuya açık dört farklı veri seti kullanılmış ve k-en yakın komşular, naïve bayes, destek vektör makineleri ve karar ağacı algoritmaları üzerinden sınıflandırmalar yapılmıştır. Destek vektör makineleri doğruluk açısından %85,2 ile en iyi sınıflandırma sonucunu vermiştir.

Akay [58], yaptığı çalışmada, destek vektör makineleri algoritmasının iyi bir analitik doğruluk yeteneğine sahip olduğunu ispatlamaya çalışarak, öznitelik seçme tekniği ile birleştirilmiş destek vektör makinelerine dayalı meme kanseri tahmin analizi yaptı. Wisconsin meme kanseri veri seti (WBCD) dahil olmak üzere farklı ortak veri setleri üzerinde deneyler gerçekleştirdi. Kullandığı algoritmanın etkinliğini değerlendirmek üzere karışıklık matrisi üzerinden özgüllük, duyarlılık, sınıflandırma doğruluk performansı, pozitif ve negatif tahmin değerleri kullandı. Elde ettiği sonuçlara göre en yüksek sınıflandırma doğruluğunu %99,5 ile seçilen beş adet öznitelik bağımsız değişken üzerinden destek vektör makineleri modeli ile ortaya koydu.

Polat ve ark. [59], meme kanseri tanısı için bir LS-SVM sınıflandırıcı algoritması önermişler ve 10 parçalı çapraz doğrulama kullanarak tanı için %98,53 oranında sınıflandırma doğruluğu elde etmişlerdir.

Chen ve ark. [60], 440 akciğer hastasından alınan verileri kullanarak yaptıkları sınıflandırma çalışmasında yapay sinir ağları tekniği ile ve çapraz doğrulama işlemi yapmaksızın tahminde %83,5 doğruluk oranı sonucu elde etmişlerdir.

Chang ve ark. [61], destek vektör makineleri algoritması ve çapraz doğrulama kullanarak 31 adet ağız kanseri hastası veri seti ile yaptıkları çalışmada %75 doğruluk oranına ulaşırken, aynı algoritmayı 69 hasta verisi ile çalıştıran Rosado ve ark. [62], %98 doğruluk oranına ulaşmışlardır.

Olatunji ve ark. [63], tiroit kanserini çok erken aşamalarda (semptomatik öncesi aşama) tespit ederek erken uyarı sistemi görevi görebilecek makine öğrenme tabanlı araçlar geliştiren çalışmalar yaptılar. Ayrıca olabildiğince az girdi değişkenleri kullanarak mümkün olan en yüksek doğruluğu elde etmeyi hedeflediler. Bu çalışmada tiroit kanseri tahmin etmede Suudi Arabistan veri kümesi kullanılmıştır. Bu çalışmada kullanılan teknikler arasında Rastgele orman (RO), yapay sinir ağı (YSA), destek vektör makinesi (DVM) ve naif bayes (NB) yer almaktadır. Elde edilen en yüksek doğruluk oranı RO tekniği ile %90,91 olurken, DVM, YSA ve NB için sonuçlar sırasıyla %84,09, %88,64 ve %81,82 olmuştur. Elde edilen bu sonuçlara mevcut 15 özellikten yalnızca yedisi kullanılarak ulaşılmış ve RO tekniğinin daha iyi olduğu, dolayısıyla bu spesifik problem için önerilen makine öğrenmesi algoritmalardan biri olduğu gösterilmiştir.

Zachariah ve ark. [64], kolon kanseri tahmini yapan bir çalışmada polip patolojisinin sonuç tahmini üzerinde durmuşlardır. Çalışmada kullanılan evrişimsel sinir ağları (CNN) üzerinden 0.90 doğruluk, 0.96 (sensitivity) duyarlılık ve 0.90 özgüllük (specificity) değerleri elde edilmiştir.

Murugan ve ark. [65], cilt görüntülerine dayanarak cilt kanseri hastalığını tahmin eden makine öğrenmesi algoritmaları üzerinde çalıştılar. Veriler için cildin dış görünümleri bir medyan filtresi kullanılarak filtrelenmiş ve ortalama kaydırma segmentasyonu kullanılarak renk seviyesi, görüntü uzunluğu ve çapı türünden özelliklere göre guruplara ayrılmıştır. Parçalara ayrılmış görüntüler, öznitelik çıkarımına girdi olarak beslenmiştir. Ortaya çıkarılan öznitelikler, destek vektör makinesi, olasılıksal sinir ağları, rastgele orman ve DVM+RO birleşiminden oluşan sınıflandırma teknikleri kullanılarak sınıflandırılmıştır. Burada birleştirilmiş DVM+RO sınıflandırıcısı diğer sınıflandırıcılara göre daha iyi sonuçlar vermiştir. RO, DVM, DVM+RO için sırasıyla %76,36 , %87,8 ve %89,31 doğruluk değerlerine ulaşılmıştır.

Minnoora ve Baths [66], 30 adet öznitelik girdi verisi ihtiva eden Wisconsin Meme Kanseri veritabanı veri seti kullanarak kanser tahmini yapabilen en iyi makine öğrenmesi algorimasını araştırmışlardır. Sonuçları itibari ile 212 adedi kötü huylu olan toplamda 569 adet girdi verisinin bulunduğu veri seti kullanılarak öznitelik seçme yöntemleri yardımıyla, sırasıyla 16 ve 8 özniteliğe sahip veri setleri oluşturuldu ve eğitildi. Rastgele Orman modeli, farklı hiperparametrelerle denenerek sırasıyla %100 ve %99,30 doğrulukla test edildi. Bu sonuç ayrıca diğer dört makine öğrenmesi sınıflandırma algoritmaları ile de karşılaştırıldı. destek vektör makinesi, karar ağacı, çok katmanlı algılayıcı ve k-en yakın komşular algoritmalarının kullanıldığı çalışmada, rastgele orman algoritmasının meme kanseri tanısı için daha iyi bir yöntem olduğu gösterildi.

Aljuaida ve ark. [67], derin sinir ağlarının (ResNet 18, ShuffleNet ve Inception-V3Net) bir kombinasyonunu kullanan ve BreakHis üzerinde öğrenmeyi herkese açık olarak aktaran, meme kanseri sınıflandırması (hem ikili hem de çoklu sınıf) için yeni bir bilgisayar destekli tanı yöntemi sunmuşlardır. Önerilen yöntemlerle iyi huylu veya kötü huylu kanser vakalarının ikili sınıflandırmasında, ResNet, InceptionV3Net ve ShuffleNet için, sırasıyla %99,7, %97,66 ve %96,94'lük en iyi ortalama doğrulukları elde etmişlerdir.

2.6. Veri Madenciliği

Veri madenciliği, hedeflenen modellemeleri ortaya çıkarma ve büyük veri kümeleri arasında ilişkiler kurarak veri analizi yoluyla problemlere çözüm bulmak üzere yapılan sıralama sürecine verilen addır. Veri madenciliği araçları, işletmelerin gelecekteki eğilimlerinin tahmin edilmesine katkı sağladığı gibi tüm özel ya da tüzel kişilerin karar verme süreçlerine de katkı sunar.

Veri madenciliği, genel olarak, el ile analizi imkânsız olacak olan mevcut veri yığınlarından ortaya konulan hedeflerle ilgili gizlenmiş bilgilerin çıkarılması sürecini ifade eder. Bu tanımlama son zamanlarda çok sıklıkla kullanılan *veri tabanlarında bilgi keşfi* ifadesine karşılık gelmektedir. Veri tabanlarında bilgi keşfi, veri madenciliğinin yanı sıra, veri hazırlama, veri seçimi, veri temizleme ve veri ön işleme, ilgili ve işe yarar ilişkiler için gömülü veri sunma, elde edilen modelleri ortaya koyma ve görselleştirme gibi bir dizi faaliyet içerir [68].

Veri madenciliği, yararlı verileri bulmak ve bu veriler üzerinden karar vericiye karar verme aşamasında destek vermek amacıyla büyük miktarlarda veri arasından aramalar yapmak üzere kullanılan mantıksal işlemler olarak tarif edilebilir. Bu tekniğin amacı, önceden bilinmeyen kalıpları bulmak ve bu kalıpları istenilen sonuçlarla optimum düzeyde eşleştirmektir. Bu modeller bulunduktan sonra, benzer kararlar hızlı ve yerinde verilerek karar süreçlerinin daha az maliyetle daha verimli ve daha fazla kalite ile sonuçlanması sağlanabilmektedir.

Veri madenciliği, verilerin üretildiği hemen hemen her sektörde geniş bir uygulama alanına sahiptir; bu nedenle veri madenciliği, veri tabanı ve bilgi sistemlerindeki en önemli öncülerden ve disiplinler arası gelişmelerden biri olarak kabul edilmektedir.

2.6.1. Veri madenciliği ve makine öğrenmesi ilişkisi ve sınıflandırma

Makine öğrenmesi, algoritmalar kullanılarak eğitilen veri üzerinden model oluşturur ve daha sonra girilen diğer benzer veriler bu modele göre işlem görerek problemin çözümüne yönelik zeki sonuçlar ortaya koyar.

Daha ileri düzey ve karmaşık görevler için, insanın gerekli algoritmaları elle oluşturması zor olabilir. Uygulamalarda programcıların gereken her adımı belirtmeleri yerine makinenin (bilgisayarın) kendi algoritmasını geliştirmesine yardımcı olması daha etkili olabilir.

Tüm bunlar, veri tabanlarından ve veri setlerinden gerekli bilgiyi çeşitli teknikler vasıtası ile ortaya çıkaran veri madenciliği ile ilintili ve içi içedir. Veri setleri ya da veri tabanları öznitelik, değişken ya da girdiler olarak ifade edilen veri örnekleri ile oluşturulmuştur.

Her bir veri örneği girdi olarak kabul edildiğinde bir çıktı mevcutsa, makine öğrenmesi işlevi denetimli öğrenme adını alır. Veri setinin bağımsız değişkenlerine ait bağımlı bir değişken (sonuç-sınıf) mevcut değilse, bu tür makine öğrenmesi algoritmalarına denetimsiz öğrenme adı verilmektedir. Veri madenciliğinde sınıflandırma ve regresyon algoritmaları bu bağlamda denetimli öğrenmeye dahil iken, kümeleme algoritmaları çoğunlukla denetimsiz öğrenmeye dahil edilmektedir.

Veri madenciliği ve makine öğrenmesi birlikte bilgisayar bilimleri altında yer almaktadırlar. Her ikisi de eldeki veri setlerinden karar destek sistemleri (KDS) için en uygun sonuçları verecek örnek modeller ortaya çıkarmak için kullanılırlar. Aynı amaçlarla veriler arasından çözüm sunan örnek modeller ortaya çıkarsalar da her iki yaklaşım arasında belirgin farklılıklar mevcuttur.

Veri madenciliğinde kullanılan en yaygın teknik sınıflandırma tekniğidir [69]. Sınıflandırma, büyük kalabalık verilerin önceden tanımlanmış bir sınıf grubuna ayrılması işlemidir. Dolandırıcılık tespiti, hastaları sağlık merkezlerine ya da ilgili uzmanlık alanına göre sınıflandırma, kredi riski uygulamaları ve hastalık tanı ve tedavilerinde hasta verilerinin kullanımı sınıflandırmanın uygulandığı konulardan birkaçıdır. Sınıflandırma süreci genellikle denetimli öğrenme ve sınıflandırma adımlarından oluşur ve çoğunlukla öngörülü modellemeler için kullanılır. Sınıflandırmada kullanılan yaygın algoritmik modellerden bazıları karar ağaçları, yapay sinir ağları, bayesian sınıflandırması, destek vektör makineleri ve ilişkilendirmeye dayalı sınıflandırmalardır [70].

İstatistik biliminde sınıflandırma, kategorik olarak üyeliği bilinen gözlemleri (veya örnekleri) içeren bir eğitim veri setine dayanarak, yeni bir gözlemin grup kategorilerinden hangisine ait olduğunu tanımlama işlemidir. Hastanın gözlenen özelliklerine (cinsiyet, yaş, kan değerleri, birtakım belirtilerin varlığı veya yokluğu, vb.) dayalı olarak belirli bir hastalığa tanı konulması sınıflandırma yaklaşımına örnek verilebilir.

Veri madenciliğinde sınıflandırma, makine öğrenmesi teknikleri ile ele alındığında denetimli öğrenme örneklerine dâhil olmaktadır. Diğer yandan denetimsiz öğrenme örnekleri, kümeleme olarak bilinir ve kendi içindeki veri benzerliği veya veriler arası mesafenin her bir ölçüsüne göre verilerin kategoriler halinde gruplandırılmalarını içerir.

Sınıflandırmayı özellikle somut bir uygulama ile oluşturan bir algoritma, sınıflandırıcı olarak adlandırılır. *Sınıflandırıcı* terimi bazen matematikteki *fonksiyon* terimine karşılık gelir ve girdi olarak kullanılan verileri belli kategorilere dâhil eder.

Farklı disiplinlerde farklı anlamlarda kullanılsa da, makine öğrenmesi alanında her bir gözlem genellikle örnek adıyla ele alınırken, bu örnekler içinde mevcut olan ve her birini açıklayıcı görev üstlenen bu değişkenler öznitelik (kimi yerde özellik) olarak adlandırılırlar. Sonuç tahmini yapılan olası kategoriler ise *sınıf* olarak adlandırılırlar.

2.6.2. Sağlık hizmetlerinde veri madenciliği kullanımı

Veri madenciliği uygulamaları günlük yaşamda giderek daha fazla kullanılmaktadır. Ticaret, sağlık ve eğitim alanlarında karar destek olanaklarından faydalanmak üzere veri madenciliğinin çeşitli yöntemlerinden yararlanılmaktadır.

Sağlık hizmetleri hem gelişmiş hem de gelişmekte olan ülkelerde hızla büyüyen önemli bir alandır. Türkiye için de sağlık hizmetlerinin özellikle son çeyrek yüzyılda hızlı bir şekilde büyüdüğüne tanık olunmaktadır. Sağlık hizmetlerinin içerdiği tüm alanlar için üretilen veriler çok yüksek hacme sahiptir ve çok çeşitlidir ve bu nedenlerle sağlık hizmetleri endüstrisinin çeşitli alanlarında mevcut olan veri madenciliği fırsatları değerlendirmeye alınmakta ve çeşitli proje çağrıları açılmaktadır [71].

Elektronik sağlık veri kapasitesinin sürekli artması nedeniyle, veriler arasında bir çeşit karmaşıklık ve karışıklık mevcuttur. Teknolojik araçların yeterince kullanılmadığı geleneksel yöntemler kullanıldığında, bu verilerden anlamlı bilgiler elde etmek çok zorlaşmaktadır. Ancak, istatistik, matematik, veri madenciliği ve yapay zekâ teknikleri gibi diğer birçok disiplin alanındaki ilerlemeler nedeniyle, veriler arasından anlamlı bağlantı kalıpları bulmak ve çıkarmak mümkün olmaktadır. Veri madenciliği yaklaşımı çok sayı ve çeşitte bulunan sağlık verilerinin işlenmesi ve analiz edilmesinde çok önemli yararlılıklar sağlamaktadır.

İş ve perakende alanındaki başarıların yanı sıra, veri madenciliği tıp ve sağlık alanındaki avantajlarını yansıtmaya devam etmektedir. Algoritmaların adaptasyonu

problemi, sağlık hizmetlerinin veriler düzeyinde karmaşıklığı ve bu alanda, etik kurallar ve çekinceler nedeni ile, bu konularda teknolojik altyapının yavaş ilerliyor olması nedenleriyle sağlık sistemlerinde veri madenciliği çalışmaları hala temel aşamadadır [72].

Araştırmacılar, tıbbi sonuçları tahmin etmek üzere konsültasyon boyunca hasta verilerini ve göreceli de olsa hasta yanıtlarını toplamayı hedeflemektedirler. Ayrıca, belirli bir cerrahi işlemin, tıbbi testlerin ve ilaçların etkinliğini tahmin etmek ya da hastalığa tanı koymak üzere veri madenciliğini kullanmayı amaçlamaktadırlar. Bu durum, klinik karar verme standartlarının yükseltilmesine yardımcı olabilir ve insanların sağlığına kavuşmasına ve bu alanda güvenliğinin sağlanmasına katkıda bulunabilir [73].

Sağlık hizmeti sunulan ortamlar genellikle veri açısından zengin ancak karar destekleyici olması açısından bilgi yoksunu olarak algılanmaktadır. Sağlık sistemlerinde çok fazla veri üretilmekte ancak, veriler arasında kolaylıkla görülemeyen ilişkileri ve eğilimleri keşfetmek üzere gerekli analiz araçları eksik kalmaktadır.

Bu nedenle klinik kararlar, örneğin kanser teşhisi genellikle veri tabanında saklanan bilgi bakımından zengin verilerden ziyade doktorların sezgilerine ve deneyimlerine dayanarak verilir. Bu uygulama, istenmeyen önyargılara, hatalara ve hastalara sunulan hizmet kalitesini etkileyen çok fazla tıbbi maliyetlere yol açmaktadır.

Erken evrede teşhis edilmeleri halinde kanser türlerinin çoğunun tedavi edilebilir olduğu bilinmektedir. Dünya Sağlık Örgütü verilerine göre, tüm kanser türlerinin en az %40'ı tedavi edilebilir ve aynı zamanda diğer kronik hastalıklarda da görülen risk faktörlerinden uzak durulabilir [74].

Kanser, gelişmiş ülkelerdeki insanlarda ölümün ana nedenlerinden biri olmuştur. Kanser ölümlerini azaltmanın en etkili yolu kanser hastalığını daha erken tespit etmektir. Sağlık sistemi, erken tanı yapmak ve cerrahi biyopsiye gerek duymaksızın, iyi huylu tümörleri kötü huylu tümörlerden ayırmak üzere kullanabileceği doğru ve güvenilir tanı prosedürlerine ihtiyaç duymaktadır.

Kanser hastalıklarında erken tanı ve tedavi için öneri amaçlı veri madenciliği tekniğinin bir karar destek sistemi olarak kullanımı ile ilgili başarılı örnekler mevcuttur.

Chang ve ark. [75], yetişkinlerde ve çocuklarda cilt hastalıklarını karakterize etmek üzere entegre bir karar ağacı modeli kullandılar. Araştırmada altı ana cilt hastalığına ilişkin beş deneyin sonuçları analiz edilmektedir. Bu araştırma ile karar ağacı kullanılarak dermatolojide en iyi öngörücü modeli oluşturma amacı güdülmüş ve karar ağacı ile sinir ağları sınıflandırma yöntemleri birleştirilmiştir. Deneysel sonuçlara dayanarak, yapay sinir ağı algoritmalarının cilt hastalıklarının öngörülmesinde doğruluk sonuçları açısından % 92.62 kesinliğe sahip olduğu bulunmuştur. Khan ve ark. [76], meme kanseri hastasının sağ kurtulma ihtimalini tahmin etmek üzere karar ağacı kullandılar.

3. MATERYAL VE YÖNTEM

3.1. Veri Madenciliği Metodolojisi

Bu tezde, veri madenciliği yöntemlerinden olan sınıflandırma yaklaşımı üzerinden denetimli algoritmalar kullanılmıştır. Bu yöntemin tercih edilmesi eldeki tiroit kanserli hasta verilerinin bağımlı değişkeni olan sonuç değişkenlerinin ikili veri tipine sahip olması ve bu veri tipinin sınıflandırma yöntemlerine uygun olması nedeniyledir.

Veri madenciliği algoritmalarının ve tekniklerinin ortaya çıkma nedenlerinin ve kullanım alanlarının farklılaşması ile optimal veri madenciliği tekniklerinin, modellerinin ve algoritmalarının, projenin veya araştırmanın gereklerine göre seçilmesi ve bunun için önemli miktarda veri toplanması gerektiği burada özellikle belirtilmelidir. Bu yaklaşım, beklenen sonuçları elde etmek ve bunları uygun şekilde analiz etmek için gereklidir.

Bu alt bölüm, veri madenciliği algoritmalarını uygulamak üzere çoğu analist/araştırmacı tarafından kullanılan tekniklere odaklanmaktadır. Veri tabanlarından bilgi devşirmek üzere sıklıkla kullanılan bu algoritmalardan ve tekniklerden bazıları sınıflandırma, kümeleme, regresyon, ilişkisel kural tabanlı tahmin yürütmedir.

3.1.1. Veri toplama

Kalitesi düşük veri kümeleri kalitesi düşük karar yapılarına neden olacaktır. Teori dışında gerçek klinik veri setleri tutarsız, net olmayan ve kayıp veriler içermektedir. Veri setlerinin büyük olması, muhatap alınan hasta sayılarının çokluğu, sağlıklı veri akışının olmayışı gibi sebeplerle veriler elde edildiği ilk anda kullanıma hazır değillerdir. Verilerin toplanma işlemleri bu nedenlerle çok önemlidir. Bu tezde kullanılmak üzere elde edilen veriler Sakarya Üniversitesi Araştırma Hastanesi, Genel cerrahi poliklinikleri üzerinden elde edilen, troit kanseri hastalarına ait verilerdir. İlgili kanser hasta verileri excel tablosu kullanılarak kayıt altına alınmıştır. Tablo 3.1'de öznitelik sütunu altında adları verilen veriler klinik hasta verilerinden alınarak doktor asistanları tarafından elle tek tek girilmiştir. Giriş sırasında eksik veriler mevcut olsa da gürültülü veriler elenmiştir.

3.1.2. Veri seçimi

Veriler, üzerinde çalışılan hastalık verilerinden alınarak bir ön hazırlıktan geçirilmelidir. Ham veri olarak tabir edilen bu veriler poliklinik veri tabanlarında düzenli tutulsalarda dikkatli bir ön işlemden geçirilmelidirler. Bu tezde kullanılan tiroit kanser hastalığı verileri, Sakarya Üniversitesi Araştırma hastanesine çeşitli nedenlerle başvuran ve tiroite bağlı kanser tanısı konan ve tümü ameliyat edilen hasta verilerinden alınmıştır. Bu veriler içinden yaş, cinsiyet, tiroksin ile ilgili sorgular ile antitroit ilacı kullanıp kullanmadığı, hamile olup olmadığı, farklı başka tedavilere tabi tutulup tutulmadığı sorularına alınan cevaplar yanısıra, hipotiroit durumu, lityum, guatır, tümör, hipopitüiter, TSH, sT3, sT4, T4U, FTI, TBG gibi ölçüm sonuçları alınmıştır.

Veri seti, başvuran hastalar içerisinden tiroit kanseri teşhisi konmuş ve ameliyat edilmiş hasta verilerinden oluşmaktadır. Veri seti içinde bulunan ve her biri anlamları ve tipleri ile birlikte verilen tiroit kanser hasta verileri Tablo 3.1'de bir arada verilmiştir.

Tablo 3.1. Tiroit kanseri veri setine ait verilerin anlamı ve veri tipleri.

Öznitelik	Anlamı	Orijinal veri tipi ve içerik	Veri tipi (Dönüştürme sonrası)
Cinsiyet	Kadın ya da Erkek	İki terimli (Binominal: Kadın:1 Erkek: 2)	Sayısal-ikili (0,1)
Yaş	Yaş	Sürekli (20-87aralığı)	Sayısal-Sürekli
Usg	Ultrasonografi	Çok terimli (Polinominal, kötü huylu, olası kötü huylu, iyi huylu, olası iyi huylu)	Sayısal-kategorik (0,1,2,3)
Usg Lap	Ultrasonografi ile Lenfadenopati	İki terimli (Binominal, yok, mevcut)	Sayısal-ikili (0,1)
Şüpheli Nodül Çapı	Kanser şüpheli nodülün çapı	Çok terimli (Polinominal ,1 cmden küçük,1-2 cm arası,2-3 cm arası,3-4 cm arası,4 cmden büyük)	Sayısal-kategorik (0,1,2,3,4)
Boy/En Oranı	Kanserden şüphe edilen nodülün boy/en oranı	İki terimli (Binominal, Genişliğinden daha uzun olan, Uzunluğu genişliğinden az)	Sayısal-ikili (0,1)
Nodül Ekosu	Nodülün izoekoik, hipoekoik olma durumu	Çok terimli (Polinominal, İzoekoik, Hipoekoik, Hiperekoik)	Sayısal-kategorik (0,1,2)
Mikro kalsifikasyon	Göğüs dokusunda rastgele oluşan kalsiyum çökeltisi.	İki terimli (Binominal, mevcut, yok)	Sayısal-ikili (0,1)

Tablo 3.1. (**Devamı**) Tiroit kanseri veri setine ait verilerin anlamı ve veri tipleri.

Öznitelik	Anlamı	Orijinal veri tipi ve içerik	Veri tipi (Dönüştürme sonrası)
Cinsiyet	Kadın ya da Erkek	İki terimli (Binominal: Kadın:1 Erkek: 2)	Sayısal-ikili (0,1)
Yaş	Yaş	Sürekli (20-87aralığı)	Sayısal-Sürekli
Usg	Ultrasonografi	Çok terimli (Polinominal, kötü huylu, olası kötü huylu, iyi huylu, olası iyi huylu)	Sayısal-kategorik (0,1,2,3)
Usg Lap	Ultrasonografi ile Lenfadenopati	İki terimli (Binominal, yok, mevcut)	Sayısal-ikili (0,1)
Şüpheli Nodül Çapı	Kanser şüpheli nodülün çapı	Çok terimli (Polinominal ,1 cmden küçük,1-2 cm arası,2-3 cm arası,3-4 cm arası,4 cmden büyük)	Sayısal-kategorik (0,1,2,3,4)
Boy/En Oranı	Kanserden şüphe edilen nodülün boy/en oranı	İki terimli (Binominal, Genişliğinden daha uzun olan, Uzunluğu genişliğinden az)	Sayısal-ikili (0,1)
Nodül Ekosu	Nodülün izoekoik, hipoekoik olma durumu	Çok terimli (Polinominal, İzoekoik, Hipoekoik, Hiperekoik)	Sayısal-kategorik (0,1,2)
Mikro kalsifikasyon	Göğüs dokusunda rastgele oluşan kalsiyum çökeltisi.	İki terimli (Binominal, mevcut, yok)	Sayısal-ikili (0,1)

3.1.3. Veri ön işleme

Veri ön işleme veri madenciliği tekniği olup, ham verinin anlaşılabilir bir formata sokulmasını amaçlamaktadır. Veri ön işleme, ilk elden anlaşılması zor ham verinin anlaşılabilir ve işlenmeye hazır hale getirilmesi süreçlerinin metodudur. Veri kümeleri arasında ki tutarsızlık, veri ön işleme metodunun önündeki en önemli engellerden biridir. Veri temizleme, veri ön işleme ve öznitelik mühendisliği gibi bir çok farklı tanımları da olan verinin hazırlık süreci, tahmin sonuçlarını en iyi verecek modelin bulunması öncesi özellikle makine öğrenmesi alanında verilerin ön işleme tabi tutularak hazır hale getirilmesi, amaçlanan sonuçların kesine yakın olmasını etkileyecek derecede önemlidir.

Veri ön işleme işlemleri elle yapılması zor ve zaman alan bir süreçtir. Zaman içinde yazılım algoritmaları üzerinden çeşitli veri ön işleme teknikleri geliştirilmiştir. Eldeki verilerin modellemeler üzerinden istenilen en uygun sonuca ulaştırılması süreci, kullanılacak algoritmalarda en iyi uyumu sağlayacak duruma getirilmeleri ile hedefine ulaşabilecektir. Sınıflandırma ve regresyon analizleri üzerinden oluşturulan makine

öğrenmesi modellemelerinin en uygunu elde edilmeden önce yapılması gereken çeşitli işlemler mevcuttur. Bu işlemler;

- Verilerin temizlenmesi
- En uygun verilerin seçimi (Öznitelik seçimi)
- Verinin dönüştürülmesi (gerekli ise)
- Öznitelik mühendisliği (Mevcut veriden yeni veriler türetmek)
- Boyutsallık azaltımıdır.

Bu işlemlerin tamamı uygulama safhasında kendilerine özel algoritmalara tabidirler. Makine öğrenmesi algoritmaları uygulamadan önce verinin özelliklerine ve yapısına has olacak şekilde seçilecek çeşitli tekniklerle/algoritmalarla hazır hale getirilmelidirler.

MÖ algoritmaları üzerinden ham veriler ile sağlıklı değerlendirmeler yapılamaz; bunun yerine, seçilen makine öğrenmesi algoritmalarının gereksinimlerini karşılamak için veriler dönüştürülmelidirler.

Ham verilerin kullanımı üzerinden makine öğrenmesi algoritmalarının performansları sağlıklı değerlendirilemez. Uygun ve sağlıklı bir değerlendirme için seçimi yapılan algoritmanın ön koşullarına ve kurallarına göre verilerin hazır hale getirilmeleri gerekmektedir.

3.1.3.1. Veri seti elde etme

Veri seti hazırlamadaki ilk basamak çeşitli ve farklı kaynaklardan veri elde etmektir. Çeşitli veri tabanlarından elde edilen ilgili veriler bir veri seti oluşturmak üzere bir araya getirilmelidirler. Bu işlem yapılırken veri tipleri ve formatları göz önüne alınması yanı sıra tabloların birleştirilmesinde normal formlara dönüştürülmesine de dikkat edilmelidir. Bu tezde kullanılan veri seti Sakarya Üniversitesi, Tıp Fakültesi altında faaliyet gösteren Araştırma hastanesi Genel Cerrahi bölümünden elde edilmiştir. Veriler bir veri tabanından tablolar halinde çekilmemiş olup, her bir hasta için veriler tek tek elle bir Excel tablosuna giriş yapılarak oluşturulmuştur. Veri seti oluşturma aşamasında veri tipleri ve formlarının aynı olmasına dikkat edilmiştir.

3.1.3.2. Veri temizleme

Kayıp verilerin tamamlanması, kirli verilerin anlamlandırılması, aykırı ve anlamsız verilerin ayıklanması ve veri tutarsızlıklarının giderilmesi işlemlerini içerirler. Bu tezde kullanılan tiroit kanseri verileri, önceden öznitelik adları belirlenmiş sütunlara tek tek ve uygun veri tipleri ve veri formatları gözetilerek el ile işlenmiştir. Veri setinde kirli veri bulunmamaktadır. Üç farklı sütunda eksikliği görülen 5 adet veri içinde, çoklu yerine koyma (Multiple Imputation) gözetilerek string veri tipi için *Medyan,* nümerik olanlar içinse *Mean* kullanılarak tamamlanmıştır.

3.1.3.3. Veri entegrasyonu

Veri entegrasyonu farklı fakat ilgili veri tabanlarının ya da veri tablolarının birleştirilmesi işlemlerini içerir. Eldeki veri seti tek excel tablosundan oluşmaktadır. Veri entegrasyonu ve/veya veri birleştirmeleri söz konusu olan durumlar oluşmamıştır. Eldeki excel veri tablosunda farklı sütunlar olan MCV ve RBC değerleri kullanılarak MCV/RBC oranı elde edilmiş ve sonuçlar Mentezer Index adıyla yeni bir sütun olarak tabloda yer almıştır.

3.1.3.4. Veri dönüştürme

Veri dönüştürme işlemleri normalizasyon ve veriyi bir araya toplama işlemlerini içerir. Eldeki tiroit kanseri veri setinde var olan veriler binominal, string ve nümerik veri tiplerinden oluşmaktadır. Bu verilerin tümünü belli modellere, özellikle makine öğrenmesi modellerine uygulamak için bir biri ile uyumlu hale getirmek gerekmektedir. Uyum için gerekli işlemlerden biri veri tipi uyumu iken bir diğeri de verilerin kendi aralarında sayısal olarak büyük aralıklara sahip olmamalarıdır. Örneğin yaş verileri 26 ile 72 aralığında bulunan büyük sayılar iken kan tahlilleri 1,2–2,5 aralığında küçük sayılardan oluşmakta ve ayrıca evet-hayır karşılığı olarak verilen 0 veya 1 gibi binominal değerlerinin yanında büyük kalmaktadır. Bu ve benzeri sayısal değerler arasında çok büyük farkların olması sonuç almada hata paylarını yüzdelik olarak arttırmaktadır. Bu nedenle normalizasyon teknikleri kullanılarak veriler arası değer yakınsamaları belli bir seviyeye (genelde 0 ile 1 aralığına) çekilmelidir.

3.1.3.5. Veri azaltma

Veri azaltma, fazla miktardaki verinin kullanılan teknik üzerinden elde edilen sonucu etkilemeyecek şekilde azaltılması ve/veya ayrıklaştırılması işlemlerini içerir. Veri özniteliklerinin fazlalığı sınıflandırmada iyi sonuçlar alınabileceği anlamına

gelmemektedir. Hatta ilgisiz sütunların varlığı, daha az öznitelik ile elde edilebilecek sonuçları çok fazla oranda olumsuz etkileyebilir. Ayrıca veri setindeki öznitelik sütunlarının fazlalığı kurulmaya çalışılan modelin sonuç çıkarmasını işlem zamanı olarak olumsuz etkiler. Eldeki tiroit kanseri veri setinde bulunan iki adet sütun (öznitelik) çok fazla verinin aynı olması nedeni ile (%97 aynı veri) setten tamamı ile çıkarılmıştır. *Takma ad, hasta numarası* gibi veri sütunları ihtiyaç duyulmaması nedeni ile setten çıkarılmıştır.

3.2. Öznitelik Seçme

Çok büyük miktarda veriyi işlemek üzere elde bulunduran karar destek sistemleri (KDS), etkili ve verimli bir sonuç elde etmek için veri ön işlemeyi çok dikkatli, etkili ve verimli bir şekilde yapmalıdır. Verilerin analiz edilmesi ve birbirleri ile olan bağlantılarının özellik seçimi ile dikkatlice ortaya konması istenilen sonuçların elde edilmesinde hayati bir öneme sahiptir.

Öznitelik seçiminin asıl amacı, belirlenen özellikler üzerinden veriler arası ilişkiyi en büyük tutacak şekilde en az sayıda özellik seçimi yapmaktır. Mümkün olan en az sayıda özniteliği en iyi sonucu verecek şekilde ayarlamaya çalışmak, öznitelik seçimi basamağının önceliğidir. Öznitelik seçimi gereksinimi;

- Problem boyutunu azaltma gerekliliği.
- Bilgisayar depolama alanı gereksinimini azaltma.
- Hesaplama süresini kısaltma.
- Tahmin kalitesini artırmak hedefiyle öznitelik sayısında azalmaya gitme.
- Sınıflandırıcının alakasız özellikler seçmesini baştan engellemek ve gürültülü olarak adlandırılan kirli ve tutarsız verilerin kaldırılması ihtiyaçlarından doğmaktadır.

Veri madenciliği ve makine öğrenmesi teknikleri ile işleme konulacak verilerin uygun değer düzeylerinde olmaları ve birbirleri ile olan bağlantılarının istenilen sonuca götürecek düzeyde olması karar süreçlerinin en önemli adımlarından biridir. Kullanılacak veri madenciliği teknikleri ve yapay zekâ algoritmaları ne kadar iyi seçilirse seçilsin, kullanılan verilerin veri öznitelikleri istenilen sonuca yönelik olmadıkça istenilen iyi sonuçlar elde edilemez. Hastalık tanısı için eldeki bulgular,

tahliller, hastalık belirtileri, hasta şikâyetleri, hastanın geçmiş hikâyesi ve şimdiki yaşam koşulları birer veri özelliği olarak ele alınabilir.

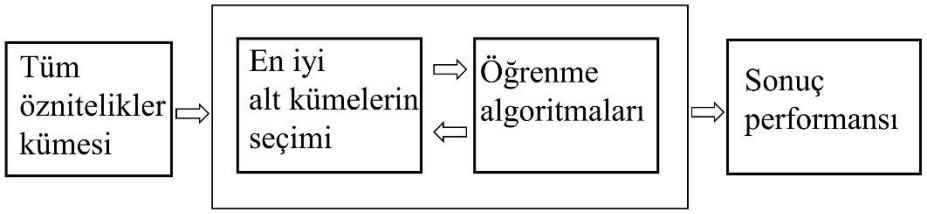
Her ne kadar hastalık tanıları için çeşitli teknolojik yöntemlerle elde edilen bulgular, hasta şikâyetleri, geçmiş hasta hikâyeleri birer veri olma özelliği taşısa da istenilen tanı ve tedavi yöntemleri için verilerin analiz edilerek işe yarayabilecek verilerin elde edilmesi gerekmektedir. Verilerin mümkün olduğunca işe yarar ve kendi aralarında uygun ilişkiler kurabilen veriler olması tercih edilmelidir. Bu sebeple veri analizlerinin iyi yapılması ve kurulması düşünülen KDS için kullanılması planlanan verilerin iyi seçilmesi gerekmektedir. Bu seçim veri madenciliği söz konusu olduğunda *öznitelik seçimi* kavramı olarak ortaya çıkmaktadır.

Veriler çok sayıda özniteliğe sahiptir, ancak tüm bu öznitelikler kendi aralarında yeterli ilişkiye sahip olmayabilir. Sınıflandırma ya da kümelemede daha iyi sonuçlar elde edebilmek için öznitelik seçimi yapma işlemleri, ilgisiz ve diğer tüm özniteliklerle uyum sağlamayan -varsa- bazı öznitelikleri, eldeki verilerden fazlaca bilgi kaybı olmadan ortadan kaldırmak için yapılır.

Temel olarak üç çeşit öznitelik seçme yaklaşımı vardır: Sarmalayıcı (Wrapper), filtreleme (Filter) ve gömülü (Embedded) yöntemler. Gömülü yöntemde, öznitelik seçimi aynı zamanda öğrenme algoritmasının bir parçasıdır.

3.2.1. Sarmalayıcı öznitelik seçme tekniği

Sarmalayıcı yaklaşım veya Sarmalayıcı yöntem için işlem basamakları Şekil 3.1'de gösterilmiştir. Bu yöntemde, özniteliğin durumu kullanılan sınıflandırıcıya bağlıdır, yani verilen özniteliğin ilgili model için gerekli en iyilerden olup olmadığını belirlemek için sınıflandırıcının sonucunu kullanır. Bu yöntem, filtre yönteminin dezavantajını ortadan kaldırır, yani sınıflandırıcı ile etkileşimi içerir ve ayrıca öznitelikler arası bağımlılıkları göz önüne alır. Bu yöntem bağımlılıkları göz önüne alması nedeni ile filtre yönteminden daha yavaştır ve bu durum zaman kullanımı açısından dezavantaj sayılabilir. Bu yöntemde özniteliğin kalitesi, kullanılan sınıflandırıcının performansı ile doğrudan ölçülür.



Şekil 3.1. Sarmalayıcı öznitelik seçme akış şeması.

Sarmalayıcı yöntemleri, basitçe alakalı özniteliklerden ziyade optimum öznitelikleri tanımlamakta daha iyidir. Bunu, öğrenme algoritmasının sezgisel özelliklerini ve eğitim setini kullanarak yaparlar. İleriye doğru eleme ve geriye doğru eleme, önemsiz öznitelikleri alt kümeden kaldırmak sarmal metodu tarafından kullanılan yöntemlerdir. Sarmalayıcı metod genel olarak sıralı seçim algoritmaları ve sezgisel arama algoritmaları olarak sınıflandırılır.

3.2.1.1. Sıralı seçim algoritmaları

Sıralı öznitelik seçimi [77,78,79], algoritması (Sequential Selection Algorithms) içi boş bir setle başlar ve ilk adımda objektif fonksiyon kullanıldığında en yüksek değeri veren bir öznitelik ekler. İlk adımdan sonra, kalan öznitelikler mevcut altkümeye ayrı ayrı eklenir ve yeni altküme değerlendirmeye alınır. Maksimum sınıflandırma doğruluğu sağlayan öznitelikler tek tek kalıcı olarak alt kümeye dâhil edilir. Gerekli sayıda öznitelik elde edilinceye kadar bu işlemler tekrarlanır.

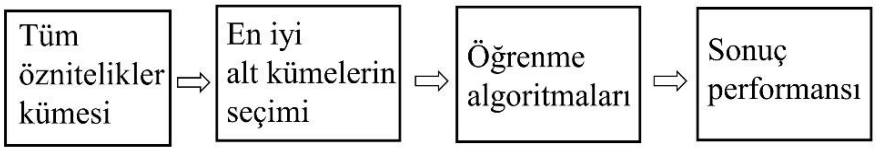
3.2.1.2. Sezgisel arama algoritmaları

Sezgisel arama algoritmaları (Heuristic Search Algorithms) genetik algoritmalar (GA), karınca kolonisi optimizasyonu, parçalı sürü optimizasyonu gibi algoritmaları içerir. GA, optimizasyon ve arama problemlerine doğru veya yaklaşık bir çözüm bulmak için hesaplamada kullanılan bir arama tekniğidir. Genetik algoritmalar, en uygun olanın var olması ilkesine dayanır. Karınca koloni optimizasyonu, gerçekte karıncaların yiyecek kaynaklarını ararken bulabilecekleri en kısa yollara dayanmaktadır. Parçalı sürü optimizasyonu yaklaşımı çaprazlama ve mutasyon operatörleri kullanmaz, dolayısıyla GA üzerinde etkilidir ancak birkaç matematiksel operatör gerektirir. Bu tür matematiksel işlemler, kullanıcı tarafından belirlenmiş çeşitli parametreler gerektirir ve bu parametrelerle ilgilenmek, optimal değerler üzerine karar vermek kullanıcılar için zor olabilir. Her ne kadar karınca kolonisi optimizasyonu ve parçalı sürü

optimizasyonu algoritmaları GA ile neredeyse aynı şekilde çalışsa da GA üstel arama alanlarındaki sadeliği ve güçlü arama özelliğinden dolayı büyük ilgi görmüştür.

3.2.2. Filtreleme öznitelik seçimi teknikleri

Filtre yaklaşımı veya filtre yöntemi işlem adımları Şekil 3.2'de gösterilmiştir. Bu yöntem, sınıflandırıcılara bağlı olmaksızın öznitelik seçimi yapar. Basit ve sınıflandırıcılardan bağımsız olması bu yöntemin avantajı olmasıyla birlikte öznitelik seçiminin yalnızca bir kez yapılması öngörülmektedir. Sınıflandırıcı ile etkileşime girmemesi, özniteliklerin birbirine bağımlılıklarını göz ardı etmesi bu yöntemin dezavantajı olarak kabul edilir ve son tahlilde her öznitelik ayrı olarak kabul görür.



Şekil 3.2. Filtreleme öznitelik seçme akış şeması.

Öznitelik sıralaması filtreleme yönteminde ana yaklaşım olarak kullanılır. Değişkenlere uygun bir sıralama kıstası kullanılarak bir puan verilir ve bazı eşik değerlerin altında bir puana sahip olan değişkenler ortadan kaldırılır. Bu yöntemler zamansal maliyetler açısından daha ucuzdur, öznitelikler arası uyumdan kaçınır ve öznitelikler arasındaki bağımlılıkları görmezden gelir. Bu nedenle, seçilen alt küme istenilene en uygun olmayabilir ve yedek bir alt küme elde edilebilir. Filtreler öznitelik seçim metodunda temel olarak ve genelde kullanılan algoritmalar aşağıdaki gibidir.

3.2.2.1. Ki-kare testi

Ki-kare (Chi-square test) filtre yöntemi testi, iki olay arasındaki bağımsızlığı kontrol eder. Genel ifade iki değişken arsındaki ilişkiyi test eder. İki olay X, Y, P (XY) = P (X) P (Y) veya eşdeğer P (X / Y) = P (X) ve P (Y / X) = P (Y) durumunda bağımsız olarak tanımlanır. Daha özel olarak, öznitelik seçiminde, belirli bir terimin oluşumunun ve belirli bir sınıf oluşumunun bağımsız olup olmadığını test etmek için kullanılır.

3.2.2.2. Öklid uzaklığı

Öklid uzaklığı öznitelik seçim tekniğinde, öznitelikler arasındaki ilgileşim öklid mesafesi cinsinden hesaplanır. İki nokta arası uzaklığı ifade eder. Örnek özelliği "a" "n" içeriyor diyorsa, bu "n" öznitelik sayısı, aşağıdaki denklemi kullanarak

aralarındaki mesafeyi hesaplar ve diğer "n-1" öznitelikleriyle karşılaştırılır (denklem 3.1). Öznitelikler arasındaki mesafe kalır ve yeni özniteliklerin eklenmesinden sonra bile etkilenmezler.

$$d(a,b) = \{\sum_i (a_i - b_i)^2\}^{1/2} \tag{3.1}$$

3.2.2.3. Korelasyon kıstasları

Veriler arası istatistiksel ilişkiyi temsil eder. Varlıklar yada veri noktaları arası ilişki pozitif, negatif veya sıfır olarak ifade edilir. Pearson korelasyon katsayısı en basit kriterlerdir ve aşağıdaki denklem ile tanımlanır: x_i i. değişken olduğunda, Y çıkış sınıfı varyansı ve kovaryans belirtir (denklem 3.2). Dezavantajı korelasyon sıralamasının sadece değişkenler ve hedefler arasındaki doğrusal bağımlılıkları tespit edebilmesidir.

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}} \tag{3.2}$$

3.2.2.4. Bilgi kazanımı

Bilgi kazanımı (BK) bize, verilen bir özelliğin ne kadar önemli olduğunu söyler. BK öznitelik seçim yöntemi, en yüksek bilgi kazancı puanlarına sahip terimleri seçer. Bilgi kazanımı, eğer bir öznitelik içinde veya ilgili sınıf dağılımında aranan bilgi mevcutsa, sınıf tahminiyle ilgili parça halindeki bilgi miktarını/miktarlarını ölçer. Bu durumda öznitelik bağımlı sonuç değişkeninden ne kadar bağımsız ise BK o denli düşük çıkacaktır. BK'nın olabildiğince artış göstermesi ilgili sınıfa yakınlığını ve dolayısıyla ilgili sınıfa aidiyetini ortaya çıkaracaktır.

3.2.2.5. Müşterek bilgi

Müşterek bilgi (MB) algoritmasındaki bilgi-teorik sıralama kriterleri iki değişken arasındaki bağımlılık ölçüsünü kullanır. Eğer bağımlı değişken (sınıf) ile bağımsız bir değişken arasında yüksek müşterek bir bilgi söz konusu ise bağımlı sonuç değişkeninin ilgili bağımsız değişkenle yüksek düzeyde müşterek bir bağımlılığı olduğu söylenebilir.

3.2.2.6. Korelasyona dayalı öznitelik seçimi

Korelasyona dayalı öznitelik seçimi algoritması, sınıf etiketini tahmin etmek için bireysel özniteliklerin faydalarını ölçen bir yöntem kullanır ve öznitelikleri

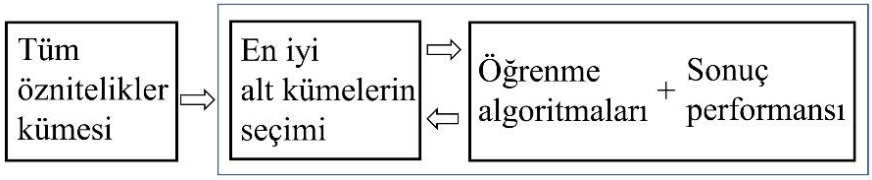
aralarındaki korelasyon düzeyini kullanarak seçer. Yüksek derecede korelasyonlu olan ve alakasız öznitelikleri seçmekten kaçınır. Bu algoritma parametresi, sınıfın zayıf tahminini sağlayan alakasız, gereksiz öznitelikleri filtrelemek için kullanılır.

3.2.2.7. Hızlı korelasyon tabanlı öznitelik seçimi

Hızlı korelasyon tabanlı filtre, bir dizi öznitelikle başlayan, özniteliklerin bağımlılıklarını hesaplamak için simetrik belirsizlik kullanan ve sıralı arama stratejisiyle geriye doğru bir seçim tekniği kullanarak en iyi alt kümeyi bulan çok değişkenli bir öznitelik seçim yöntemidir.

3.2.3. Gömülü öznitellik seçimi teknikleri

Gömülü yaklaşım veya gömülü yöntem işlem adımları Şekil 3.3'te gösterilmiştir. Sınıflandırıcı yapıya yerleştirilmiş olan uygun değer öznitelikler alt kümesini arar. Sınıflandırıcının doğruluğu sadece sınıflandırma algoritmasına değil, aynı zamanda kullanılan öznitelik seçim yöntemine de bağlıdır. Alakasız ve uygunsuz özniteliklerin seçilmesi sınıflandırıcıyı yanlış ve istenmeyen sonuçlara götürebilir.



Şekil 3.3. Gömülü öznitelik seçme akış şeması.

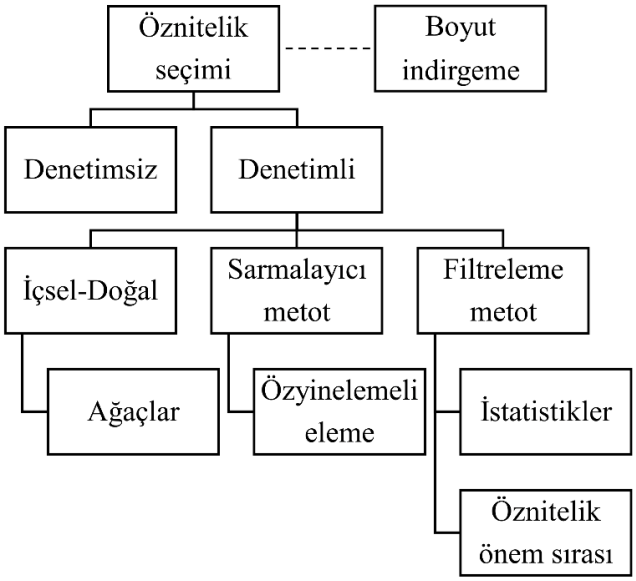
Gömülü yöntemlerde, öznitelik seçimi yöntemi bir öğrenme algoritmasına dâhil edilir ve bunun için optimize edilmiştir. Ayrıca filtre ve sarma yönteminin bir birleşimi olan hibrid model olarak da adlandırılır. Gömülü yöntemler, sarmalayıcı yöntemler tarafından yapılan farklı alt kümelerin yeniden sınıflandırılması için harcanan hesaplama süresini azaltır.

3.3. Mevcut Uygulamalarda Kullanılan Öznitelik Seçme Teknikleri

Yukarıda öznitelik seçim teknikleri kategorik olarak genelleştirilerek açıklandı. Bu yaklaşımlar dikkate alınmak kaydı ile makine öğrenmesi teknikleri açısından öznitelik seçim yaklaşımı, denetimli ve denetimsiz metotlar üzerinden ele alınabilir. Eğer öznitelik seçimleri yukarıda belirtilen yöntemler kullanılıyor ve veri setinde ki hedef/sonuç (sınıf) göz ardı ediliyorsa bu yaklaşım denetimsiz metotlarla yapılıyor

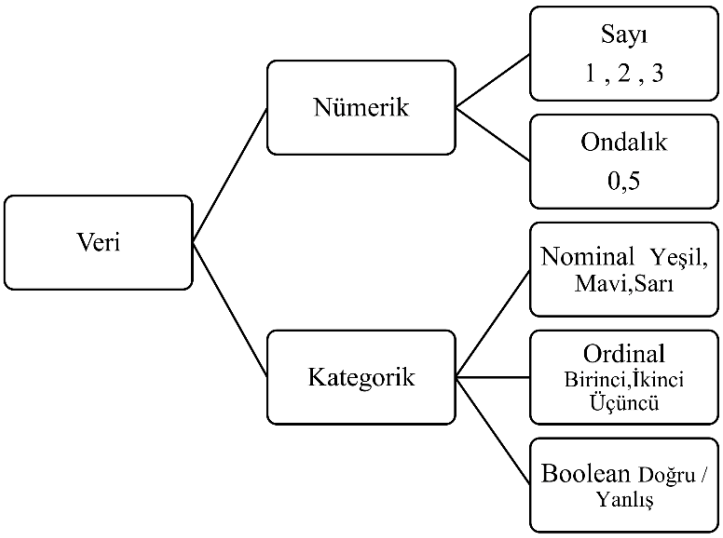
demektir. Denetimli öznitelik seçim yaklaşımı söz konusu olduğunda ise kullanılan algoritma hedef sütununu (sonuç sütunu) bağımlı değişken olarak ele alır ve diğer tüm girdi değişkenleri ile etkileşime sokarak en iyi öznitelik kombinasyonunu ortaya çıkarmaya çalışır.

Elde bulunan veriler içinden en uygun öznitelik sayılarını ortaya çıkarmak üzere Şekil 3.4'de akış şeması verilen bir yöntem izlenmiştir. Şekilde verilen bu şemaya göre veri seti denetimli (sonuç sütununun işin içinde bağımsız değişken olarak var olduğu durum) ya da denetimsiz (sonuç sütunu işlem dışı bırakılır ve sadece bağımsız değişkenler işleme tabi tutulur) olma durumuna göre çeşitli öznitelik seçme yöntemlerine tabi tutulabileceği gösterilmiştir.



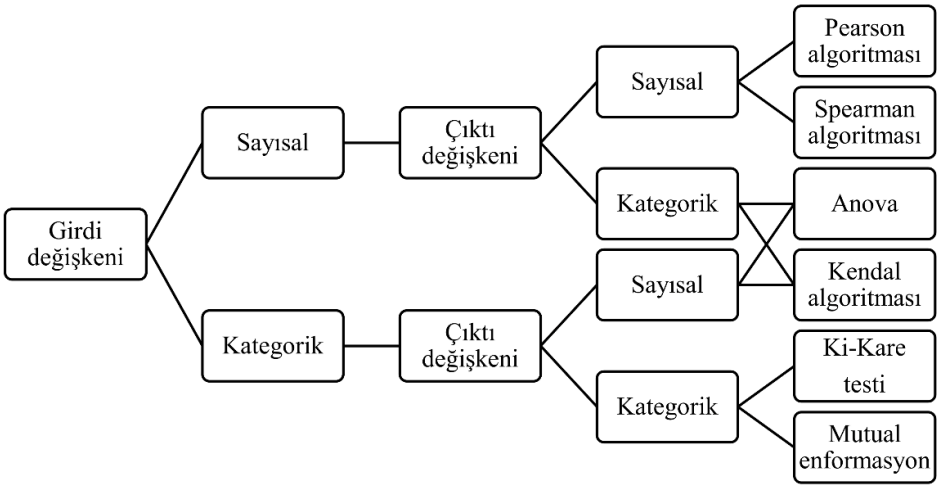
Şekil 3.4. Öznitelik seçim yöntemleri.

Eldeki tiroit kanseri hasta verileri üzeninden karar verilen sonuçlar hastaların kanserli olup olmadıkları üzerinden ikili (bi-nominal) verilerdir. Sonuçların (sınıf) belli olması denetimli metotlar üzerinden hareket edilmesine olanak sağlamaktadır. Denetimli yöntem ile başlamaya karar verildikten sonra hangi denetimli yöntemin ya da metodun kullanılacağı ise verilerin ne denli hazırlıklı olduğu ve veri tiplerinin bir arada nasıl entegre edildiklerine bağlı olacaktır. Veri seti içinde hem nümerik hem de kategorik veriler mevcuttur. Bu durumda verilerin veri tipleri, makine öğrenmesi algoritmalarının seçiminde önemli rol oynayacaklardır. Şekil 3.5 veri sınıflarını ve ihtiva ettikleri veri tiplerini göstermektedir.



Şekil 3.5. Veri seti için veri tipleri ve kategorileri.

Kullanılan tiroit kanseri veri setinde veri değişkenleri çeşitli tiplerdedir. Veriler arası korelasyon ve diğer farklı ilişkilerin ortaya çıkarılması aşamasında veri tipleri kendi aralarında uyumlu olmaları gerekmektedir. Şekil 3.6'da veri tiplerine en uygun öznitelik seçme algoritmaları için yaklaşımlar görülmektedir.



Şekil 3.6. Değişken veri tiplerine göre öznitelik seçme algoritmaları.

Şekil 3.6'ya göre hareket edilerek eldeki tiroit kanseri verileri sayısal ve kategorik olma durumları yanı sıra kesikli ve sürekli veri tipleri olarak tekrar oluşturularak çeşitli öznitelik seçme algoritmalarına ve testlerine tabi tutuldular.

Veri alt setleri oluşturulurken sınıflandırma modellerine en uygun olabilecek öznitelikleri seçmek için ki-kare istatistik (Chi2: stat), kramer değeri (Cramer's Value), müşterek malumat (Mutual Information), arttırım oranı (Gain Ratio), F:

istatistik (F: Stat), fisher puanı (Fisher Score), welch:istatistik (Welch: Stat), spearman:istatistik (Spearman: Stat), kendall : istatistik (Kendall: Stat), ki-kare:p-değeri (Chi2: p-Value), welch:p-değeri (Welch: p-Value), gini indeksi (Gini Index), f:p-değeri (F: p-Value), pearson:p-değeri(Pearson: p-Value), spearman:p-değeri (Spearman: p-Value), kendall:p-değeri (Kendall: p-Value) testleri kullanılmıştır.

3.4. Makine Öğrenmesi Model Yaklaşımı

Makine öğrenmesi, zeki kararlar almak amacına yönelik zeki karar destek sistemleri oluşturmak üzere yaklaşım tipi, metotlar ve tekniklerden oluşan bir bütünü oluşturur. Makine öğrenmesi söz konusu olduğunda akla gelen temel iki kavramdan biri *veri* iken diğeri verileri işleyen *model*'dir. Veriler makine öğrenmesi modellerine dahil edilerek modelin zeki kararlar vermesi sağlanmaktadır. Makine öğrenmesi modeli matematiksel olarak kısaca girdi değişkenler ile hedef (sınıf) değişken arasındaki ilişkiyi ortaya çıkarmaya çalışan bir fonksiyon olarak ifade edilebilir.

Bir fonksiyon olarak model, veri içinden kalıplar bulmaya, verileri anlamaya ve veriler üzerinde eğitimler uygulamaya çalışır. Hafızaya kaydedilen bu eğitim kalıpları üzerinden, makine öğrenmesi modelleri tahminler yürüterek karar vermede destek sunarlar.

Makine öğrenmesi modellerinde öğrenme işlemi etiketlenmiş bağımlı veri (labelled data) üzerinden oluyorsa denetimli öğrenme adını almaktadır. Denetimli öğrenme, sınıflandırma ve regresyon yöntemlerini içine almaktadır. Burada sınıflandırma yöntemi bir sınıfı yada ayrık bir değeri bulmaya odaklanırken (Doğru-yanlış, kadın-erkek, kanserli-kansersiz gibi), regresyon miktarlar yada sürekli değerleri bulma üzerine odaklanır (maaş, yaş, miktar, fiyat gibi). İlgili tüm tahlil ve tetkik sonuçları verilen hastanın kanserli yada kansersiz olduğunu ortaya çıkaran modeller denetimli iken, metrekareye düşen yağmur miktarını tahmin eden bir model regresyona örnek verilebilir.

Lojistik regresyon, destek vektör makineleri, karar ağacı, k-en yakın komşular, rastgele orman, naif-bayes sınıflandırma problemlerinde sıklıkla kullanılan sınıflandırıcılar iken, liner regresyon, lasso regresyon, çok terimli regresyon, destek vektör makine regresyon sağlayıcısı, rastgele orman regresyon sağlayıcısı, bayes doğrusal regresyon sağlayıcı algoritmaları regresyon algoritmaları olarak sıklıkla

kullanılmaktadırlar. Belirtilen bu sınıflandırıcılarla regresyon sağlayıcılarının tamamı makine öğrenmesi altında bulunan denetimli öğrenme modellerine aittirler.

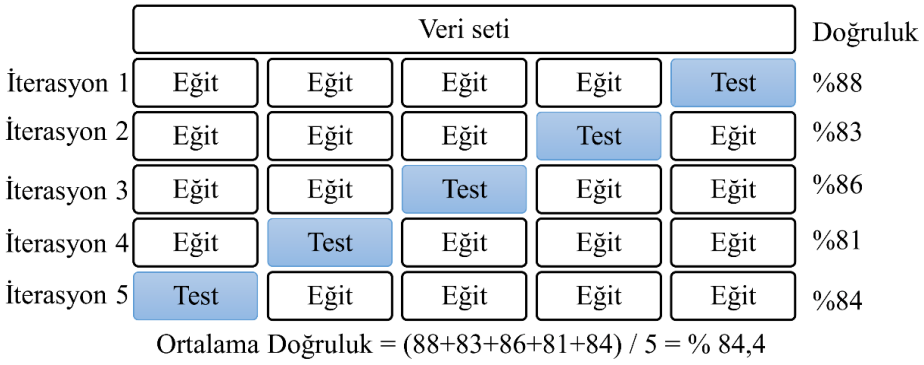
Makine öğrenmesinin bir parçası olan diğer bir öğrenme model yaklaşımı denetimsiz öğrenmedir. Denetimsiz öğrenmede algoritmalar bağımlı çıktı değişkeni (sınıf) olmaksızın sadece girdi-bağımsız değişkenler üzerinden öğrenme gerçekleştirirler ve kendi hedef çıktılarını oluştururlar. Kümeleme (Clustering) ve ilişkilendirme (Association) algoritmaları bu tür öğrenmede iki ana yaklaşımdır. Kümeleme (Clustering) benzer verileri aynı gruba dahil ederek gruplar oluştururken, ilişkilendirme (Association) algoritmaları ise veriler arasında önemli ilişkiler kurarak öğrenmesini gerçekleştirir.
K-means kümeleme, hiyerarşik kümeleme, temel bileşenler analizi ve apriori denetimsiz öğrenme modelleri için sıklıkla kullanılan algoritmalardır.

3.4.1. Çapraz doğrulama

Makine öğrenmesi modellerinde uygulamaya dahil olan veri setleri en uygun modeli hangi eğitim ve test setleri ile elde edeceği ile ilgili önceden bir sonuca varmak zordur. Eğitim ve test verilerinin modelin her çalıştırılmasında farklı kesitlerden alınması doğruluk değerini en çoklayacak sonuca götürebilir.

Makine öğrenmesi modellerinde çapraz doğrulama (ÇD) adıyla kullanılan bu fonksiyon farklı modeller ve farklı veri setleri üzerine uygulanarak en uygun veri setine uyan en iyi modelin bulunmasına yardımcı olur.

Şekil 3.7 ‘de temsili olarak verilen veri seti 5 parçaya ayrılmış ve her iterasyonda ayrı bir parçası test için kullanılırken diğer dört parçası birlikte eğitim seti olarak kullanılmıştır.



Şekil 3.7. Veri setinin farklı veri kesitleri ile çalıştırılmaları.

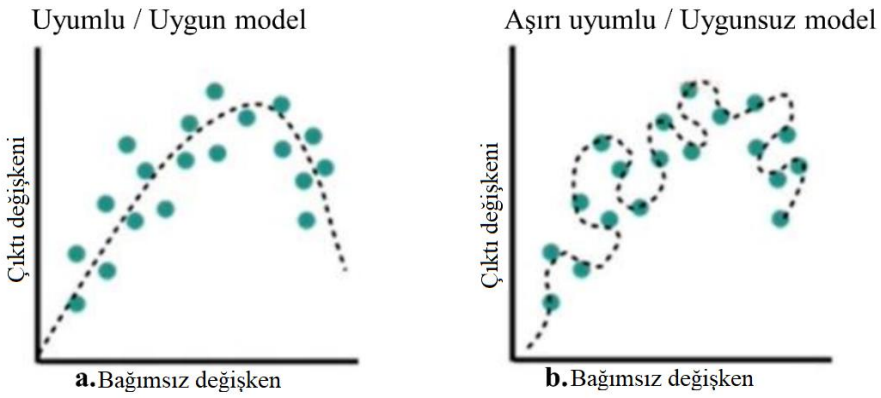
İterasyon sayısına göre elde edilen bu doğruluk değerlerinin ortalaması alınarak modelin ortalama doğruluk değeri elde edilir.

Buradaki amaç farklı eğitim ve test verileri kombinasyonlarının hangilerinin kullanılan makine öğrenmesi modeli üzerinden en iyi doğruluk sonucunu vereceğini bulmaktır. Çoğu durumlarda veri seti 5 parça olarak kullanılsa da 10 parça olarak kullanıldığında daha iyi sonuçlar verebilmesi mümkündür.

3.4.2. Veri seti ve model ilişkisinde aşırı uyum veya yetersiz uyum problemi

Aşırı uyum (AU) kavramı makine öğrenmesi modellerinde eğitilen verilerin beklenenden fazla ve çok yüksek bir doğruluk yüzdesi değeri ile sonuçlanması anlamına gelmektedir. Model veri setindeki tüm bilgiler ile birlikte gürültülü veriler dahil tüm detayları öğrenmesi aşırı uyum sorununu ortaya çıkarır ve bu durum modelin gerçek performansına olumsuz etkide bulunur. Bu duruma aşırı eğitilmiş model de denebilir.

Şekil 3.8 a'da görülen grafik veri seti için uygun ve yeterli bir modeli temsil ediyorken, Şekil 3.8 b ise her bir veri noktasını model fonksiyonuna dahil etmeye çalışan ve uygun olmayan bir modeli temsil etmekte ve istenmeyen bir sonucu göstermektedir ve aldatıcıdır.

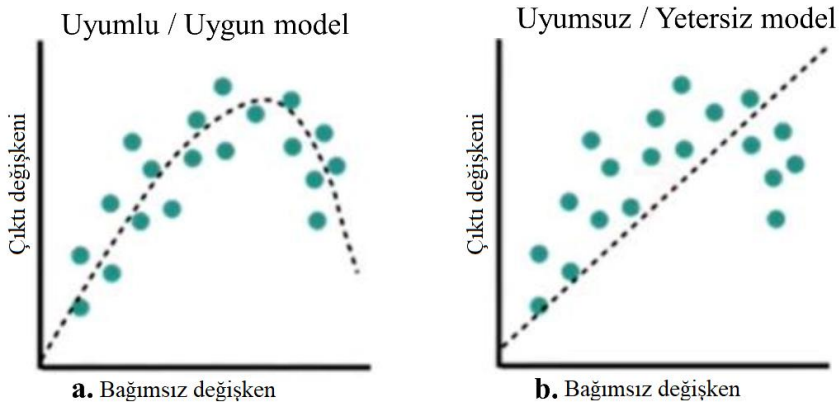


Şekil 3.8. a. Uygun model b. Aşırı uyumlu ve uygunsuz model

Veri seti, eğitim modelinde çok yüksek ancak test modelinde çok düşük doğruluk değerleri veriyorsa bu tür bir problemin varlığına işaret ediyor demektir.

Kullanılan modelin karmaşıklığı, kullanılan verinin azlığı ve yapay sinir ağları modelleri için fazla katmanın kullanılması gibi sebepler aşırı uyum pronlemini meydana getirebilir. Bu problemin üstesinden gelmek için fazla veri kullanılması,

yapay sinir ağı modelinde katmanların azaltılması ve sapma-varyans değişim tekniği kullanımı işlemleri gerçekleştirilebilir. Öte yandan kullanılan model verilerden yeterince öğrenemiyor olabilir. Bu durumlara ise yetersiz uyum problemi denmektedir.



Şekil 3.9. a. Uyumlu model b. Yetersiz model.

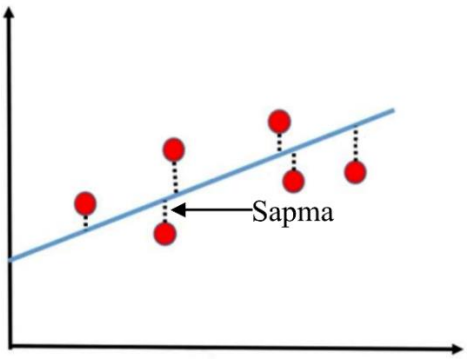
Şekil 3.9 a'da model yeterince eğitildiği halde Şekil 3.9 b'de veri setini eğitemeyen ve yetersiz kalan bir model grafiği mevcuttur. Bu tür problemlerde kullanılan model hem eğitim hemde test verisi için çok düşük doğruluk değeri üretecektir.

Yanlış model seçimi, modelin çok basit olması, düşük varyansın aksine sapmanin yüksek olması türünden nedenler yetersiz uyum problemini tetikler. Bu tür problemlerden uzak durmak için makine öğrenmesi modellerinin veri setine ve çözülmek istenen probleme uyumlu seçilmesi, modelin kompleksitesinin yeterince arttırılması, modele ait daha fazla parametrenin devreye alınması, sapma ve varyans değişim tekniği kullanımının gerçekleştirilmesi gerekmektedir.

3.4.3. Sapma-varyans değişimi ve kayıp fonksiyonu

Sapma kavramı, kullanılan model tarafından yapılan tahmin ile tahmin edilmeye çalışılan gerçek hedefler (çıktılar) arasındaki farkları ifade etmek için kullanılmaktadır.

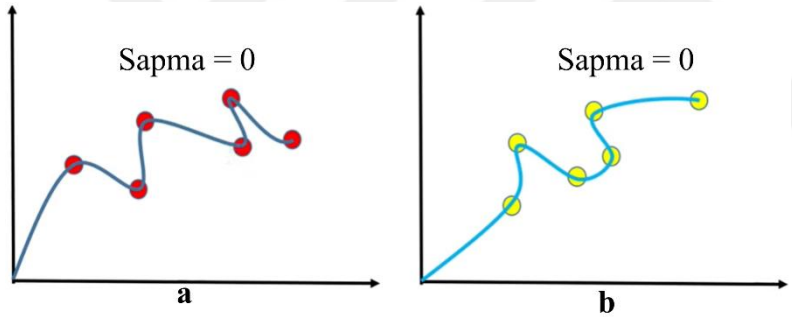
Şekil 3.10'da görüldüğü üzere veri noktalarının her birinin model fonksiyon çizgisine uzaklığı sapma olarak ifade edilmektedir. Sapma değeri az olduğu oranda model ile veri seti uyumu yüksek olacak ve bu nedenle modelin doğruluk değeri de yüksek olacaktır.



Şekil 3.10. Veri noktasının model fonksiyon çizgisine uzaklığı (Sapma).

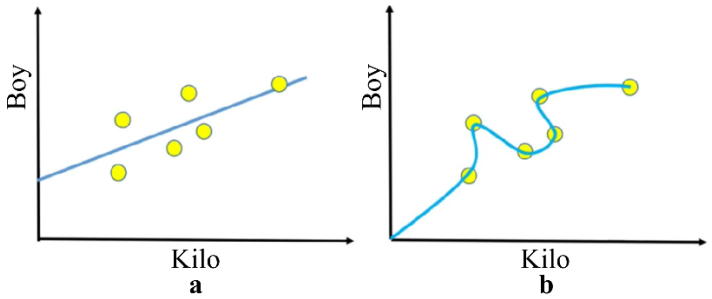
Varyans (variance) ise, farklı eğitim verisi kullanıldığında hedef fonksiyonun farklı miktarlarda hesapladığı değer anlamında kullanılmaktadır.

Şekil 3.11 a'da gösterildiği üzere modelin fonksiyon grafiği Şekil 3.11 b'de veriler değiştiğinde farklı bir fonksiyon grafiğine dönüşmüştür. Sapmanın sıfır olduğu durumlarda verinin değişmesi fonksiyonda değişikliğe neden olmaktadır.



Şekil 3.11. a. ve b. Sapmanın olmadığı durumlarda fonksiyon eğrisinin görünümü.

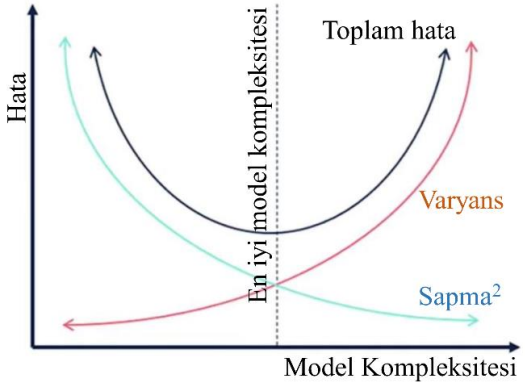
Sapmanın var olduğu ancak yetersiz uyum gösteren modeller (Şekil 3.12 a) ile sapmanın olmadığı ancak aşırı uyum gösteren modeller de (Şekil 3.12 b) mevcuttur.



Şekil 3.12. a. Yetersiz uyumlu b. Aşırı uyumlu modeller.

Şekil 3.12 a'da görülen grafik yetersiz uyum problemi gösteren, yüksek sapma ve düşük varyansa sahip bir modeli temsil ederken Şekil 3.12 b grafiği ise aşırı uyum

gösteren, düşük sapmali ve yüksek varyansa sahip olan bir modeli temsil etmektedir. En iyi modelin elde edilmesi ve verilerin aşırı uyumlu yada uyumsuz problemlerinden kurtarılabilmesi adına model fonksiyonu için en uygun kompleksitenin bulunması gereklidir. Kompleksite ile sapma ve varyans ilişkisini gösteren grafik Şekil 3.13'te verilmiştir.



Şekil 3.13. En iyi model kompleksitesinin gösterimi.

Makine öğrenmesi modellerinde hangi modelin daha iyi performans elde ettiğini gösteren bir diğer gösterge kayıp fonksiyonudur (KF). Model üzerinden hesap edilen değerlerin gerçek değerlerden ne kadar saptığını gösteren bir fonksiyondur. Hangi modelin ve hangi parametrelerin daha uygun olduğunu açığa çıkarmakta yardımcı olur. Cross entropy loss, sequared error loss, KL divergence adlarıyla farklı tipte kayıp fonksiyonları mevcuttur.

3.4.4. Model parametreleri ve hiperparametreleri

Makine öğrenmesi modellerinde kullanılan parametreler Model parametreleri ve hiperparametreleri olarak iki kategoride ele alınır. Model parametreleri eğitim verilerinin eğitilmesi sırasında kullanılan parametreler olup dahili parametreler olarak adlandırılılırlar. $Y = a\ X + e$ türünden bir doğrusal denklemde a ve e bu modelin parametreleri olarak adlandırılırlar. Çoklu girdi değişkenlerin varlığında ise yukarıdaki denklem $Y = a_1X_1 + a_2X_2 + a_3X_3 + e$ türünden bir denkleme dönüşür. Bu denklemlerde her X bağımsız girdi değişkeni için bulunan Y gerçek hedefe yakınsadığı en iyi durumlarda ortaya çıkan a ve e parametreleri bu model için en uygun parametreler olarak kayda geçerler. Bu modeller için dahili parametreler olan a katsayısı ve e sabit sayısı ağırlık ile sapma sayılarına karşılık gelirler. Burada ağırlık olarak a parametresi X girdi değişkeninin Y çıktı değişkeni üzerindeki etkisine karar verir. Model

denklemindeki *e* ise sapma değeri olarak modele verilen ofset değeridir. Burada kullanılan *e* sabit sayısı denklemi belirgin bir yöne yönlendirmek amacıyla kullanılır. Denklemdeki *Y* çıktı sonucu tüm girdi değişkenleri sıfır olduğunda sapma yerine geçen *e* sabit sayısına eşitlenir. Dolayısıyla model denkleminde ağırlık sayısı ile sapma sayısının en iyi durumu aynı zamanda modelin en iyi performansını ortaya koyar.

Model için harici parametreler olan hiperparametreler ise sahip olduğu değerler ile model işlem süreçlerini kontrol altında tutarlar. En iyi modelin araştırılmasında kullanılan bu parametreler değişken parametrelerdir. Öğrenme oranı, işlem döngü sayısı (iterasyon) hiperparametreler olarak modele dahil olurlar.

3.4.5. Makine öğrenmesinde gradyan iniş optimizasyon algoritması

Gradyan iniş makine öğrenmesi modellerinde en önemli konulardan biridir ve makine öğrenmesi modellerinden en iyisinin araştırılması sırasında kayıp fonksiyonunu, ağırlığı, sapmayı, dahili parametreleri ve hiperparametreleri birlikte çalıştıran bir optimizasyon algoritmasıdır.

Gradyan iniş optimizasyon algoritması ile, kullanılan modelin optimizasyonunu sağlayacak en iyi parametrelerin kullanılması ile kayıp fonksiyon sonucunun olabildiğince azaltılması ve böylelikle modelin doğruluk sonucunun en büyüklenmesi amaçlanmaktadır. Öğrenen modelin en iyi sonuçlar vermesine yarıyacak parametrelerini güncellemek için kullanılır. Güncellenen ağırlıkların (denklem 3.3) ve sapma değerlerin (denklem 3.4) hesaplamaları gösterilmektedir.

$$a = a - L * da \tag{3.3}$$

$$e = e - L * db \tag{3.4}$$

a --> Ağırlık

e --> Sapma

L --> Öğrenme oranı

*d*a --> a'ya göre kayıp fonksiyonunun kısmi türevi

*d*b --> b'ye göre kayıp fonksiyonunun kısmi türevi olmak üzere ağırlık (a) ve sapma (e), iterasyonlarda her seferinde hesaplanarak optimizasyon gerçekleştirilir.

4. ARAŞTIRMA VE BULGULAR

4.1. Doğru Makine Öğrenmesi Modelinin Araştırılması

Makine öğrenmesinde model seçimi, belirlenmiş bir problem için en uygun modelin seçilmesi sürecidir. Belirlenmiş bir tanı probleminin çözümü için bulunabilecek en uygun modelin seçilmesi farklı parametre ve hiperparametrelerle bir çok denemeyi gerektirmektedir. Makine öğrenmesi yaklaşımında eldeki verilerle en iyi sonuçları verebilecek modellerin araştırılması ve içlerinden en iyisinin seçilmesi hedeflenen son aşamadır.

En uygun model araştırması öncesi üzerinde durulması gerekli çeşitli faktörler vardır. Bunlar;

- Bir model seçimi için hedefe odaklanan mantıksal bir sebep
- Birden çok modelin ayrı ayrı performanslarının karşılaştırılması.

Modeller araştırlırken ve seçilirken belirli parametrelere başvurulur. Bu parametreler ve bağlı model seçme yaklaşımları birlikte Tablo 4.1'de bir arada verilmiştir.

Tablo 4.1. Makine öğrenme model tipleri ve seçme kriterleri.

Veri tipleri ve Model yaklaşımları		Sonuca odaklı ve işlem bazlı model yaklaşımları	
Görüntü ve Videolar	CNN	Sınıflandırma işlemleri	SVM, Lojistik Regresyon, Karar ağacı, Derin öğrenme, K-NN vb.
Metin yada Ses verileri	RNN	Regresyon İşlemleri	Liner regresyon, Rastgele orman, Polinomial regresyon vb.
Sayısal veriler	SVM, Lojistik Regresyon, Karar ağacı, Derin öğrenme, K-NN,Rastgeke Orman vb.	Kümeleme işlemleri	K-means kümeleme, Hiyerarşik kümeleme vb.

Eldeki tiroit kanser verileri bağımlı çıktı değişkeni olarak ikili (binominal) veri ihtiva ettiğinden makine öğrenmesine ait olan ikili sınıflandırma algortimaları tercih edilmiştir. İkili sınıflandırma algortimaları olarak adlandırılan bu denetimli makine öğrenmesi modelleri verdikleri sonuçlar açısından çeşitli avantaj ve dezavantajlara

sahiptirler. Mevcut ve sıklıkla kullanılmakta olan makine öğrenmesi modellerinin avantajlar ve dezavantajları Tablo 4.2'de bir arada verilmiştir.

Tablo 4.2. Makine öğrenme teknikleri, avantajları ve dezavantajları.

Makine Öğrenmesi Modelleri	Avantajlar	Dezavantajlar
Liner Regresyon	-Basittir ve İşleme dahil edilmesi kolaydır. -Verileri arasında Liner ilişkisi olan veri setleri için çok iyi sonuçlar verir.	-Verileri arasında doğrusal olmayan ilişkilere sahip veri setleri için uygun değildir. -Yetersiz uyum (Underfitting) problemi oluşturabilirler. -Aykırı değerlere çok duyarlıdırlar
Lojistik Regresyon	-Uygulanması kolaydır. -Veriler arası Liner ilişkileri olan setlerde iyi sonuçlar verebilir. -Küçük ölçekli veri setlerinde aşırı uyum gösterme problemlerine muhattap olmaz	-Büyük ölçekli veri setlerinde aşırı uyum gösterme problemlerinden kaçamaz. -Veriler arasındaki komplex ilişkileri çözümlemede zayıftır. -Aykırı verilere duyarlıdır. - Büyük ölçekli veri setlerine ihtiyaç duyarlar.
Destek Vektör Makineleri	-Oluşturulan sınıflar arasında net bir ayrım olduğunda daha iyi çalışır. -Yüksek boyutlu uzayda daha etkilidir. -Veri setinde ebatın örnek sayısından fazla olduğu durumlarda daha etkilidir. -Diğer modellere göre nispeten bellek verimlidir ve kısa sürede sonuç verir.	-Büyük veri kümeleri için uygun değildir. -Veri kümesi çok fazla gürültülü veriye sahip olduğunda, çok iyi performans göstermez. -Her bir veri noktası için öznitellik sayısının, eğitim veri örneği sayısını aştığı durumlarda, düşük performans gösterecektir.
Karar Ağacı	-Sınıflandırma ve Regresyon işlemleri olarak ayrı ayrı kullanılabilirler. -Değerlendirme ve yorumlaması kolaydır. -Standardazisyona yada Normalizasyona ihtiyaç duymaz. -Aykırı ve aşırı verilere duyarlı değildir, etkilenmez.	-Aşırı uyum problemi ile karşılaşabilirler. -Verileredeki en küçük bir değişiklik hiyerarşiyi dolayısıyla sonucu farklılaştırabilir. -Veri setinde Model eğitimi beklenenden daha fazla zaman alabilir.
Derin Öğrenme	-Öznitelikleri otomatik anlama ve kavrama yetisi. -Büyük ve karmaşık veri setleri ile baş edebilme kabiliyeti. -Gelişmiş bir performansa sahiptir. -Doğrusal olmayan ilişkileri rahatlıkla ele alabilir. -Yapılandırılmış yada yapılandırılmamış verileri işleyebilme. -Tahmine dayalı modellemeler için uygundur.	-Diğer modellere göre yüksek hesaplama maliyetlerine neden olabilir. -Aşırı Uyum Gösterme (Overfitting) problemlerine neden olabilirler. -Yorumlayıcı yazılımlar tarafından yorumlanmalarında eksiklikler çıkabilir. -Verimliliği veri seti verimliliğine bağımlıdır. -Eğitimleri Eğitildikleri veriler ile sınırlıdır.

Tablo 4.2. (Devamı) Makine öğrenme teknikleri, avantajları ve dezavantajları.

Makine Öğrenmesi Modelleri	Avantajlar	Dezavantajlar
En Yakın Komşu (KNN)	-Eğitim süresi kısadır. -Kolay uygulanabilir. -Doğrusal olmayan karar yapılarını anlayabilir.	-Büyük veri kümeleri ile iyi çalışmazlar -Çok fazla boyutlu (fazla öznitelik sayısı) veriler ile uyumlu çalışamazlar. -veri seti içindeki aykırı verilere duyarlıdırlar.
Rastgele Orman Algoritması	-Karar ağaçlarında aşırı uyum sorununu azaltır ve ayrıca varyansı azaltarak doğruluk değerini artırır. -Hem sınıflandırma hem de regresyon problemlerini çözer. -Hem kategorik hem de sürekli değişkenlerle iyi çalışır. -Eksik veri değerlerini ve aykırı verileri otomatik olarak işler -Öznitelikler için ölçeklendirme gerektirmez. -Doğrusal olmayan parametreleri verimli bir şekilde işler -Gürültülü verilerden az etkilenir.	-Karmaşıklık: Rastgele Orman çok sayıda ağaçtan oluşurlar (Karar ağacı yapılarında yalnızca bir ağacın aksine) ve bunların çıktıları birleştirilir. Varsayılan olarak Python sklearn kitaplığında 100 ağaç oluşturur ve bu durum hesaplama ve kaynak maliyetini arttırabilir. Tek bir karar ağacı basittir ve çok fazla sayıda ağaçtan oluşan Rastgele orman modeline nazaran çok fazla hesaplama kaynağına ihtiyaç duymaz. -Daha Uzun Eğitim Süresi: Rastgele Orman, çok sayıda ağaçtan oluştuğundan (karar ağacı model yapılarındaki bir ağaç yerine) ve oylamaların çoğunluğuna göre karar verdiğinden, karar ağaçlarına kıyasla çok daha fazla eğitim süresi gerektirmektedirler.
Yapay Sinir Ağları (YSA)	-Ağın tamamında bilgi depolaması yapabilir. -Eksik verilerle çalışabilme kabiliyeti vardır. -Nöronların bir kaçında oluşabilecek aksaklıkları telafi edebilir, bu durumu tolere edebilir. - Dağıtılmış belleğe sahiptirler. - Makine Öğrenmesi modelleri gibi işlem görebilirler. -Bir anda bir çok işi bir arada paralel yapma kabiliyetine sahiptirler.	-Donanıma bağlıdır. Kapasitesi yüksek donanımlarla çalışmak ister. -Açıklanması zor ağ davranışları sergileyebilir. -Uygun ağ yapısının var olmaması-uygulama ve denemeler neticesinde elde edilebilmesi. -Problemlerin nümerik değişkenler yoluyla ağa tanıtılması zorluğu. Veriler üzerinde nümerik çevrim çok iyi değilse pronlemin tanıtımı o denli zorlaşır.

Tablo 4.2'de bir arada verilen makine öğrenmesi teknikleri avantajları göz önünde bulundurularak eldeki veri üzerinde bazı teknikler her seferinde ayrı ayrı denenmiştir. Elde edilen sonuçlar doğruluk değerleri üzerinden değerlendirilip ve yorumlanmıştır.

Başlangıç olarak eldeki tiroit kanser verilerinde metin bazlı olanlar tamamı ile sayısal karşılıklarına çevrildi. Veriler ondalık sayılar ve tam sayılardan oluşan sütunlar

halinde bir araya getirildi. Tamsayılar 0 ve 1 den oluşan binominal sayılardan oluştuğu gibi 1'den 4'e kadar giden çoklu değerlerden (polynominal) de oluşmaktadır.

Veri ön hazırlık işlemlerinden geçirilen tiroit kanser veri seti farklı ve çok sayıda alt veri kümeleri şeklinde farklı veri setleri olarak düzenlenmiştir. Bu veri alt setleri oluşturulurken Ki-kare İstatistik (Chi2: stat), Kramer değeri (Cramer's Value), Müşterek malumat (Mutual Information), Arttırım oranı (Gain ratio), F: İstatistik (F: Stat), Fisher puanı (Fisher Score), Welch:İstatistik (Welch: Stat), Spearman:İstatistik (Spearman: Stat), Kendall : İstatistik (Kendall: Stat), p değerleri gibi testler kullanıldı.

Sıralamaları değişmekle birlikte 28 adet bağımsız öznitelik içinden (bu öznitelikler içinde sütun arttırma ile elde edilen öznitelikler dahildir.) öznitelik algoritmalarına göre uygun olan ilk 16 adet öznitelik seçimi yapıldı. Bu öznitelikler sarmalayıcı öznitelik seçim yaklaşımına göre 2'li, 3'lü, 4'lü ve en son 16'lı olacak şekilde farklı alt veri setlerine dönüştürüldü. Bunların yanı sıra makine öğrenmesi modellerinde denenmek üzere öznitelik algoritmalarına tabi tutmaksızın sadece tam sayılardan oluşan, sadece ondalık verilerden oluşan ve sadece ikili verilerden oluşan farklı alt veri setleri de ayrıca ayrı veri setleri olarak oluşturuldu. Her bir veri seti ayrı ayrı ön işlemlerden geçirilerek makine öğrenmesi modellerinde kullanılmak üzere hazır hale getirilmiştir.

Tüm senaryolar gözönünde bulundurularak tüm veriler farklı alt kümeler üzerinden (öznitelik seçim algoritmaları kullanılsın yada kullanılmasın) makine öğrenmesi sınıflandırma algoritmalarında doğruluk değerleri üzerinden davranışları incelenmiş ve karşılaştırmaları yapılmıştır. Bu bölümde kullanılan sınıflandırma modelleri makine öğrenmesi algoritmalarından hastalık tanısı için literatürde en çok kullanılanlardan oluşturulmuştur.

En uygun modelin hangisi olabileceği ile ilgili ilk işlem bir çok modelin bir arada işleme sokulduğu ve doğruluk değerleri açısından sınandığı işlem dizisidir. Bu ilk işlemler için *Python* yazılım kodları *Google Colab* arayüzü üzerinden kullanılmıştır. İkili sınıflandırma problemlerinde en çok tercih edilen makine öğrenmesi modelleri bir arada kullanıldı. Doğruluk değerlerinin elde edildiği python kod bloğu Tablo 4.3'te bir arada verilmiştir.

Tablo 4.3. Makine öğrenmesi modellerinde doğruluk oranları.

```
# Yazılımda ihtiyaç duyulan Kütüphanelerin içe aktarılması
import numpy as np
import pandas as pd
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
# Modeller için ilgili Kütüphanelerin içe aktarılması
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
# csv verisinin Pandas DataFrame veri çerçevesine aktarılması.
#Her seferinde  bir adet veri alt seti kullanılırken diğerleri devre dışı bırakılmıştır.
tiroit_veri_seti = pd.read_csv('/content/tiroitTamami_sayisalTam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-3.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-4.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_ondalik.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-HOT.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk2.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk3.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk4.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk5.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk6.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk7.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk8.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk9.csv')
# Veri setinin ilk beş satırının görüntülenmesi
tiroit_veri_seti.head()
# Veri setinde mevcut olan satır ve sütun sayısı.
tiroit_veri_seti.shape
# Kayıp verinin olup olmadığının kontrol edilmesi
tiroit_veri_seti.isnull().sum()
# İkili(binary) hedef çıktının dağılımı 0->144  1->90 Toplam:234
tiroit_veri_seti['postopSonuc'].value_counts()
#Bağımsız girdi ve Bağımlı çıktı verilerinin ayrılması
X = tiroit_veri_seti.drop(columns='postopSonuc', axis=1)
Y = tiroit_veri_seti['postopSonuc']
```

Tablo 4.3. (Devamı) Makine öğrenmesi modellerinde doğruluk oranları.

```
#Girdi ve Çıktı verilerinin ayrı ayrı görüntülenmesi
print(X)
print(Y)
#Girdi ve Çıktı verilerinin ayrı ayrı Nampay dizileri olarak aktarılması
 X=np.asarray(X)
 Y=np.asarray(Y)
```

Model Seçimi

1. Varsayılan hiperparametre değerleriyle çapraz doğrulama kullanarak yukarıdaki kodların devamı olarak makine öğrenmesi modelleri python kodları üzerinden karşılaştırıldı (Tablo 4.4).

Tablo 4.4. Model karşılaştırması için python kodları.

```
Çapraz doğrulama ile modellerin karşılaştırılması için python kodları
# Kullanılacak Modellerin modeller değişkenine atanması
modeller = [LogisticRegression(max_iter=10000), SVC(kernel='linear'),
        KNeighborsClassifier(), RandomForestClassifier(random_state=0)]
#Çapraz Doğrulama (cv) yöntemi kullanılarak yapılan Model karşılaştırmalarını yapan fonksiyon
def model_capraz_dogrulama_karsilastirma():
  for model in modeller:
    cd_puan = cross_val_score(model, X, Y, cv=10)
    mean_accuracy = sum(cd_puan)/len(cd_puan)
    mean_accuracy = mean_accuracy*100
    mean_accuracy = round(mean_accuracy, 2)
    print(model,'için Çapraz doğrulamalar ile Doğruluklar','=', cd_puan)
print(model,' için ortalama Doğruluk (Accuracy) değeri ','=',mean_accuracy,'%')
print('----------------------------------------------------------------')
# Fonsiyonun tetiklenmesi ve her bir model için ayrı ayrı Doğruluk #sonuçlarının elde edilmesi
model_capraz_dogrulama_karsilastirma()
```

2. Çapraz doğrulama yaklaşımıyla *GridSearchCV* nesnesi kullanarak farklı hiperparametrelerle makine öğrenmesi modelleri karşılaştırıldı.

Tablo 4.3 ve Tablo 4.4'te verilen kodların devamı olarak farklı hiperparametreler altında aynı modellerin verebileceği en iyi doğruluk sonuçları araştırılmak üzere Tablo 4.5'te gösterilen python kodları çalıştırıldı.

Tablo 4.5. Farklı hiperparametreler altında model araştırması.

Farklı hiperparametreler ve *GridSearchCV* nesnesi kullanımı ile birlikte

```
#Kullanılacak Modeller listesi
modeller_listesi=[LogisticRegression(max_iter=10000),SVC(),KNeighborsClassifier(),
RandomForestClassifier(random_state=0)]
 # Yukarıda belirtilen modeller için farklı hiperparametre değerleri #içeren sözlüğün oluşturulması.
model_hiperparametreleri = {
   'log_reg_hiperparametreleri': {
      'C' : [1,5,10,15,25,30,35,40,45]
    },
   'svc_hiperparametreleri': {
      'kernel' : ['linear','poly','rbf','sigmoid'],
      'C' : [1,10,15,20,25,30,35,40,45]
     },
   'KNN_hiperparametreleri' : {
      'n_neighbors' : [1,5,10,15,20,25,30,35,40,45]
     },
    'random_forest_hiperparametreleri' : {
     'n_estimators' : [1, 5, 10, 15, 20, 25, 30,35,40,45]
     }
}
# Her bir modelin hiperparametrelerini ayrı ayrı çağırmak üzere
#anahtarlama listesinin oluşturulması
model_anahtarlari = list(model_hiperparametreleri.keys())
# Model seçim fonksiyonu
def Model_Secimi(modeller_listesi, hiperparametreler_sozlugu):
  result = []
  i = 0
  for model in modeller_listesi:
    anahtar = model_anahtarlari[i]
    parametreler = hiperparametreler_sozlugu[anahtar]
    i += 1
    print(model)
    print(parametreler)
    print('----------------------------------')
  #GridSearchCV nesnesinin uygulanması
  siniflandirici = GridSearchCV(model, parametreler, cv=10)
```

Tablo 4.5. (Devamı) Farklı hiperparametreler altında model araştırması.

Farklı hiperparametreler ve *GridSearchCV* nesnesi kullanımı ile birlikte

```
# Verinin Sınıflandırıcıya verilmesi
  siniflandirici.fit(X,Y)
  result.append({
     'Kullanılan Model' : model,
     'En yüksek Sonuç' : siniflandirici.best_score_,
     'En iyi hiperparametreler' : siniflandirici.best_params_
   })
  sonuc = pd.DataFrame(result, columns = ['Kullanılan Model','En yüksek Sonuç','En iyi hiperparametreler'])
  return sonuc
# Model seçim fonksiyonunun tetiklenmesi- SON SATIR
Model_Secimi(modeller_listesi, model_hiperparametreleri)
```

Yukarıdaki Tablo 4.3'te gösterilen Python kodları aynı veri setinden çıkarılmış farklı alt veri setleri ile çalıştırıldığında farklı doğruluk değerleri ortaya çıkmaktadır. Buna göre veri tiplerini ayırmaksızın tüm veriler bir arada kullanılmasının yanı sıra, ondalık sayılardan ve ikili verilerden oluşan ayrı ve farklı veri setleri de makine öğrenmesi modellerinde denenmiştir. Kullanılan veri setlerinin içerikleri Tablo 4.6'da verilmiştir. Kullanılan modeller var sayılan mevcut parametreler üzerinden çalıştırıldığında elde edilen sonuçlar ile farklı hiperparametrelerle çalıştırılan modeller üzerinden elde edilen doğruluk sonuçları Tablo 4.7'de bir arada verilmiştir.

Tablo 4.6. Veri setleri ve içerdikleri girdi değişkenleri.

Veri setleri	Girdi Değişkenleri
VS1	Sitoloji_sonucu_iyi_huylu USG_LAP
VS2	USG_iyi_huylu USG_LAP Sitoloji_sonucu_iyi_huylu
VS3	USG_olasi_iyi_huylu USG_iyi_huylu USG_LAP Sitoloji_sonucu_iyi_huylu
VS4	USG_olasi_iyi_huylu USG_iyi_huylu USG_LAP Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olasi_kotu_huylu

Tablo 4.6. (Devamı) Veri setleri ve içerdikleri girdi değişkenleri.

Veri setleri	Girdi Değişkenleri
VS5	USG_olasi_iyi_huylu USG_iyi_huylu USG_olasi_kotu_huylu USG_LAP Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olasi_kotu_huylu
VS6	USG_olasi_iyi_huylu USG_iyi_huylu USG_olasi_kotu_huylu USG_LAP Nodul_Ekosu_izoekoik Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olasi_kotu_huylu
VS7	USG_olasi_iyi_huylu USG_iyi_huylu USG_olasi_kotu_huylu USG_LAP Nodul_Ekosu_izoekoik mikrokalsifikasyon Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olasi_kotu_huylu
VS8	USG_olasi_iyi_huylu USG_iyi_huylu USG_olasi_kotu_huylu USG_LAP Nodul_Ekosu_izoekoik Nodul_Ekosu_hipoekoik Mikrokalsifikasyon Sitoloji_sonucu_iyi_huylu Sitoloji_sonucu_olasi_kotu_huylu
VS9	İkili tipe dönüştürülmüş tüm tamsayılar
VS10	Sürekli (tam sayılı) veri tipine sahip tüm veriler
VS11	Tüm ondalık veriler

Tablo 4.7. Farklı parametreler ve çapraz doğrulama ile doğruluklar.

Çapraz doğrulama= 5 veya 10	Lojistik regresyon doğruluk(%)		DVM doğruluk(%)		K-en yakın komşu doğruluk(%)		Rastgele orman doğruluk(%)	
Kullanılan veri setleri	Varsayılan parametre	Farklı Hiper-parametre	Varsayılan parametre	Farklı Hiper-parametre	Varsayılan parametre	Farklı Hiper-parametre	Varsayılan parametre	Farklı Hiper-parametre
VS1	76,63	76,63 C:1	76,21	78,35 C:1 sigmoid	69,26	72,21 neighbor:25	**77,08**	77,51 estimator:1
VS2	75,76	76,19 C:15	**76,21**	77,52C1 linear	73,95	73,95 neighbor:5	76,21	77,52 estimator:1
VS3	74,86	74,85 C:1	74,91	**77,52** C:1 linear	71,45	79,07 n:40 cv:5	76,2	77,39 estimator:5
VS4	**77,43**	**77,43 C:1**	73,17	76,13 C:1 sigmoid cv:5	73,13	77,84 neighbor:10	71,92	75,24 estimator:5
VS5	76,12	76,56 C:25	71,11	**79,08** C:40 sigmoid	74,44	77,83 neighbors: 40	70,98 cv:5	73,57 estimator:5
VS6	75,71	75,71 C:1	68,5	75,78 C:1 linear	72,34	76,97 neighbors: 45	69,75	74,06 estimator:5
VS7	75,71	75,71 C:1	74,91	76,97 C:10 kernel: rbf	**75,74**	76,52 neighbor:40	76,56	78,71 estimator:1
VS8	75,71	75,71 C:1	75,33	77,39 C:1 poly	74,87	76,50 neighbor: 45 (cv:5)	76,99	78,30 estimator:1
VS9	74,93	74,93 C:1	71,14	74,49 C :1 linear	71,88	75,33 neighbor :10	74,38 (cv:5)	73,95 estimator: 20 (cv:5)
VS10	74,4 (cv:5)	74,40 C:1 (cv:5)	76,21	76,21 C : 1 rbf	74,44	77,38 neighbor :45 (cv:5)	73,54 (cv:5)	75,25 estimators: 30 (cv:5)
VS11	61,52	61,52 C1	61,52	62,39 C20, rbf (cv:5)	55,51	62,41 neighbor:35	59,0	61,12 estimator:40

4.2. Sonuçların Değerlendirilmesi

Tablo 4.7'ye göre dört ayrı modelle (varsayılan hiperparametreler ile) elde edilen sonuçlar içinden en iyi sonuçları tam sayılı değişkenlerin bulunduğu veri setleri vermiştir. Ondalık verilerin bulunduğu veri setleri kullanılan tüm makine öğrenmesi modellerinde düşük sonuçlar vermiştir. Tüm verilerin içinde bulunduğu veri setinin yüksek doğruluk sonucuna ulaşmasının ise içinde tam sayılı değişkenlerin varlığına borçlu olduğu söylenebilir. Tablo 4.7'de varsayılan parametrelerle elde edilen bazı sonuçların, farklı hiperparametreler üzerinden ele edilen sonuçlardan daha yüksek çıkması varsayılan parametrelerin uygulanan bazı modeller ve bazı veri setleri için daha uygun olduğu sonucunu vermektedir. Ayrıca farklı veri tiplerine sahip veri setlerinin modeller üzerinde farklı sonuçlar vermesi, sonuçların veri tiplerine bağımlı olduğunu da göstermektedir. Buna göre tam sayıların (çoğunlukla 0 yada 1 olarak iki sayı) olduğu veri setinin sınıflandırma algoritmalarında daha iyi sonuçlar verdiği -en azından eldeki veri seti açısından- söylenebilir.

Tablo 4.6'da belirtilen veri setlerine ek olarak tam sayılı veri seti tekrar bir ön işlemden geçirilerek farklı hiperparametreli makine öğrenmesi modellerinde denenmiştir. Tam sayı veri tiplerine sahip olan veri setinin tüm girdi değişkenleri, ikili (binary) veri tiplerine one-hot-codding yöntemi kullanarak dönüştürülmüş ve modeller üzerinde çalıştırılmıştır.

Tablo 4.7'te gösterilen sonuçlar karşılaştırıldığında; tam sayılı veri tipine sahip olan veri setinin one-hot-codding yöntemi sonrası elde edilen veri setinin ulaştığı sonuçlar LogisticRegression, SVM ve KNeighborsClassifier modellerinde artarken Rastgele Orman sınıflandırma modelinde azalma eğilimine girmiştir. Veri setleri bir arada değerlendirilerek yorumlandığında hangi modelin hangi veri seti ile uyumlu çalıştığı ve sonuçlar açısından hangi modelin daha iyi olduğu ortaya konulabilir. Buna göre tam sayılı veri setinin hem one-hot-codding işlemi öncesi hem de sonrası tüm makine öğrenmesi modellerinde daha iyi sonuçlar verdiği rahatlıkla söylenebilir.

Bu aşamadan sonraki adımlar, tam sayılı verilerin bulunduğu veri setinin ön işlemler ve diğer iyileştirmeler yapıldıktan sonra seçilen makine öğrenmesi modellerinde denenmesi ve en uygun özniteliklerden oluşan veri seti ile en uygun modelin ortaya çıkarılması olacaktır.

4.3. Lojistik Regresyon Tekniği ile Tiroit Kanseri Tanısı

Bu bölümde makine öğrenmesi tekniklerinden biri olan lojistik regresyon model yaklaşımı üzerinde durulacaktır. Lojistik regresyon tekniğinin yapısı ve regresyon öncesi kullanılacak veri setinden en uygun öznitelik sütunlarının seçimi için kullanılan en iyi algoritmalar ortaya konulacaktır. Lojistik regresyon tekniğinin genelde hangi alanlarda ve hangi veri tiplerinde verimli olduğu üzerinde durulacak ve eldeki tiroit kanser veri seti üzerinden elde edilen sonuçlar yorumlanacaktır.

Eldeki veri seti Sakarya Üniversitesi Araştırma hastanesi, Genel cerrahi bölümünden alınan tiroit kanseri hasta verilerinden oluşmaktadır. Öznitelik seçimi dışında tüm diğer veri ön hazırlık işlemleri verilerin elle alınarak veri seti dosyasına aktarılması sırasında gerçekleştirilmiştir. Tiroit kanseri hasta veri seti, kullanılacak makine öğrenmesi modelleri üzerinden beklenen en iyi sonucu vermesi için hazır hale getirilmiştir. Doğruluk yüzdesinin en iyilenmesi için gerekli ve önemli öznitelikler çeşitli yöntemlerle seçilerek modellerde kullanılacak şekilde hazır hale getirilmiştir. Eksik veri olarak göze çarpan beş adet boş veri hücresi *mean-ortalama* kullanılarak tamamlandı. Excel dosyası içinde oluşturulan verilerin öz nitelik adları ve ifade ettikleri anlamları Tablo 3.1'de verilmiştir.

4.3.1. Lojistik regresyon model yaklaşımı

Regresyon analizi, iki ya da daha fazla nicel değişken arasında pozitif ya da negatif ilişkiyi ölçmek için kullanılan bir analiz yöntemidir. Değişkenler arasında -var ise- ilişkinin şiddetini yada büyüklüğünü ortaya koyar [80].

Lojistik regresyon, istatistiksel yöntemler kullanılarak bağımsız değişkenlerin ikili bağımlı değişkenlerle ilişki düzeylerini ihtimaller üzerinden belirleyen bir yaklaşımdır. Sonucu veren bağımlı değişken ikili ise sonuca hangi ihtimalle varıldığını analiz etmek üzere lojistik regresyon en iyi yöntemdir [81].

İlişki üzerinden elde edilen ihtimaller bağımsız değişkenlerin bağımlı değişkenin iki cevabından birine ihtimaller üzerinden ne kadar yakınsa o oranda sınıflandırılırlar. Örneğin tiroit kanseri olan hasta için 1 ve Kanser olmayan hasta için 0 sonucu bağımlı değişken olarak ele alındığında, bağımsız değişkenler lojistik regresyon sonucu %50'nin üzerinde bir ihtimalle bağımlı değişkeni pozitif etkiliyorsa hastanın kanserli, etkilemiyorsa kansersiz olduğu kanısına varılabilir. Ancak burada en doğru yaklaşım bir hastanın hangi ihtimalle kanser olduğu ya da hangi ihtimalle kanser olmadığıdır.

Lojistik regresyon analizlerinde veriler arası ilişki doğrusal ilişki olarak ele alınmazlar. Şekil 4.2'de gösterildiği üzere kanserli ya da kansersiz olma durumları 1 ya da 0 ile gösterilse de her bir sonuç için her hastaya ait bağımsız değişkenler farklıdır. Farklı bağımsız değişkenler aynı sonucu verse de aynı sonuçları farklı ihtimaller üzerinden vermektedirler. Örneğin tamamı ile farklı verilere sahip olsalar da 0.60 ve 0.70 kanserli olmadıkları sonucuna varılan A ve B şahısları farklı ihtimallerle de olsa kanserli olmadıkları sonucuna varılır. Bu durum aynı zamanda A şahsının 0.40 ihtimalle ve B şahsının da 0.30 ihtimalle kanserli oldukları anlamına da gelmektedir.

Lojistik regresyon bir algoritma çeşididir ve doğrusal logaritmaları sınıflanırmaya yarar. Matematiksel ifadesi Denklem 4.1 ve Denklem 4.2'de gösterilmektedir.

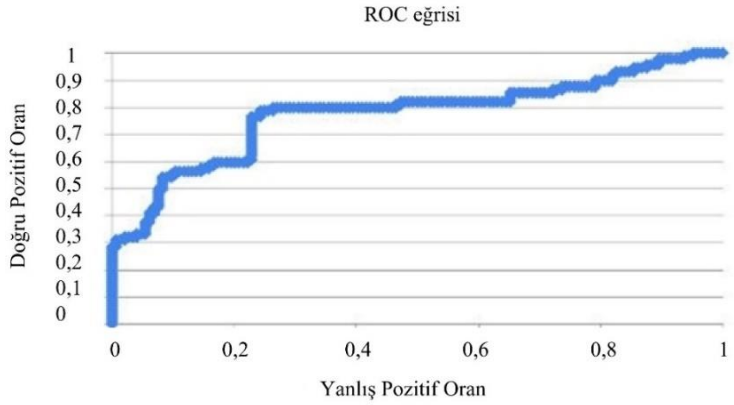
$$Y = \frac{exp\ (w.x + b)}{1 + exp\ (w.x + b)} \tag{4.1}$$

$$Y = \frac{1}{1 + exp\ (w.x + b)} \tag{4.2}$$

Bu denklemlerde *x* rasyonel sayılar kümesine ait olmak kaydıyla girdi değişkenini temsil eder. *Y* ise *0* yada *1* sonucunu veren çıktı (bağımlı) değişkenidir. *w* kullanılan ağırlık katsayısı iken *b* sapma ofset değeridir ve *w.x* ise matrislerin çarpımıdır.

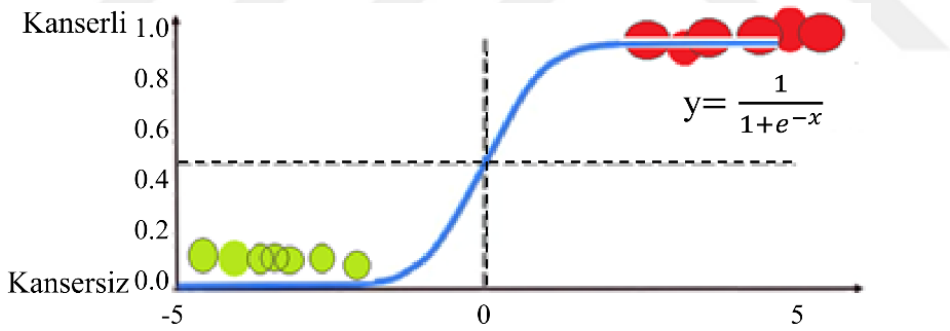
Lojistik regresyon elde ettiği sonuçları her seferinde iki olasılık değerini karşılaştırır ve *x*'i en yüksek olasılık değerine sahip olan kategoriye atar. Şekil 4.2 grafik olarak lojistik regresyon model yaklaşımını temsil etmektedir.

Şekil 4.1'de gösterildiği üzere eldeki veriler doğrusal bir grafik üzerinde iyi bir sınıflamaya dahil değilken lojistik regresyon grafiği üzerinde anlamlı bir sınıflamaya tabi tutulabilmektedir. Sonuçlar, doğrusal olmayan bir yol izleyerek 1'e yada 0'a ne kadar yakınsadığını ortaya koyan ihtimaller üzerinden gösterilmektedir.



Şekil 4.1. %61 sınırlılık düzeyinde ölçümlenmiş olasılık eğrisi.

İçinde regresyon kelimesi geçse de bu analiz bir sınıflandırma algoritması olarak kullanılmaktadır. Özellikle bağımlı değişkenin (sınıf) iki sayılı (binominal) olduğu durumlarda (evet-hayır, 1-0 gibi) sonuçların 1 ya da 0 sonuçlarından birine yakınlığını yüzdelik olarak hesaplayarak ait olduğu sınıfa yakınlık derecesini ölçer. Ölçümde genellikle 0.50 sınırı kullanılsa da bazı durumlarda hedeflenen netice sonuçlardan birine odaklanıyor ise bu %50 sınır değiştirilebilir. Bu dağılım yaklaşımı temsilen Şekil 4.2'de gösterilmiştir.



Şekil 4.2. Lojistik regresyonda ikili sınıdlandırma

Lojistik regresyon analiz modeli ile çeşitli çalışmalar ve denemeler yapılmıştır. Covid-19 hastalığı nedeniyle uygulanan karantina önlemleri sonrası ortaya çıkan riskler üzerinde yapılan çalışmalar [82-85], ile karantina sırasında ortaya çıkan kaygı ve tüm diğer faktörler arasındaki ilişkiyi ortaya çıkarmak için yapılan diğer çalışmalarda [86-89], lojistik regresyon modelleri kullanılmıştır.

4.3.2. Lojistik regresyon modeli için öznitelik seçimi

Sınıflandırma ve kümeleme yaklaşımları ile üzerinde çalışılan modellerde genellikle ham veriler kullanılmamalıdır. Bunun nedenleri şu şekilde sıralanabilir [90]:

- Makine öğrenme algoritmaları verilerin nümerik olması şartını ortaya koymaktadır.
- Bazı makine öğrenme algoritmaları bunun dışında özel bazı şartlar ortaya koymaktadır.
- İstatistiksel olarak gürültülü ve hatalı verilerin ayıklanmış olması gerekmektedir.
- Karmaşık ve/veya doğrusal olmayan bazı verilerin veri setinden çıkarılması gerekebilir.

Zaman ve maliyet göz önüne alındığında, çok fazla sayıda özniteliğe sahip olan veri setlerini temsilen daha az sayıda özniteliğin seçilmesi elzemdir. Bu gereklilik sağlanırken makine öğrenmesinde kullanılacak model için en iyi sonucu verecek öznitelik kombinasyonu iyi seçilmelidir.

Robust Scaling, Absolute Maximum Scaling, Min-Max scaling, Normalization (Mean kullanarak), Standardization gibi ölçeklendirmelerin ardından hazır hale getirilen tiroit kanseri veri seti sıklıkla kullanılan çeşitli öznitelik seçme testlerine tabi tutulmuştur.

Kullanılan testler ve test kuralları Tablo 4.8'de gösterilmiştir. Bu tabloya göre elde edilen tüm öznitelik sıralamasında kurallara uyan ilk 6 öznitelik seçilerek yeni veri setleri oluşturulmuştur. Daha fazla özniteliğin set içinde var olmasının beklenen sonuçlara olumlu katkı sunmadığı deney sonucunda ortaya konmuştur. Elde edilen alt veri setleri lojistik regresyon modelinde kullanılarak en iyi sonuçları vermesi hedeflenmiştir.

Tablo 4.8. Öznitelik önem sırası belirten testler ve test kuralları.

Testler	Kurallar
Chi2:stat,Cramer's, Mutual Information, Gain ratio, F:Stat, Fisher Score, Welch:Stat, Pearson:Stat, Spearman:Stat, Kendall:Stat	Testlerin kuralı gereği ortaya çıkan sonuçlar yukarıdan aşağıya büyükten küçüğe doğru sıralanmıştır.
Chi2:p-Value, Welch: p-Value ve Gini Index, F: p-Value, Pearson:p-Value, Spearman:p-Value, Kendall:p-Value	Testin kuralı gereği ortaya çıkan sonuçlar yukarıdan aşağıya küçükten büyüğe doğru sıralanmıştır.

Excel'de *AnalyticSolver* eklentisi kullanılarak Tablo 4.8'te gösterilen bütün testler yapılmış ve belirtilen kurallara uyan 6 adet önemli özniteliklerden oluşan veri seti ile birlikte 2'li, 3'lü,4'lü,5'li ve 6'lı alt veri setleri oluşturulmuştur. Öznitelik seçimi için kullanılan test sonuçlarını gösteren tablolardan bir kesit Tablo 4.9'de gösterilmiştir.

Tablo 4.9. Öznitelik seçiminde kullanılan testler ve önem sıralaması.

Öznitelik Seçimi: İstatistiksel					
Değişkenler	Ki-kare değeri	p değeri	Cramer değeri	Müşterek Malumat	Artış oranı
Sitoloji_sonucu_iyi huylu	65,5900	5,55174E-16	0,529433184	0,209117541	0,213642934
USG Lap	39,6385	3,05594E-10	0,411576781	0,125465891	0,213770520
Sitoloji_sonucu_kötü huylu	39,1518	3,92083E-10	0,409042528	0,131733398	0,266677578
USG_kötü huylu	29,5356	5,48971E-08	0,355275650	0,098380931	0,233591809
USG_olası_iyi huylu	24,9354	5,92818E-07	0,326438077	0,080519135	0,082261604
Sitoloji_sonucu_olası_kötü huylu	19,8616	8,32554E-06	0,291339470	0,050456966	0,057891672
USG_olası_kötü huylu	16,9218	3,89500E-05	0,268915863	0,051604071	0,056461887
Nodül_ekosu_izoekoik	15,1604	9,87504E-05	0,254535336	0,049291212	0,053931305
Mikrokalsifikasyon	14,8850	0,000114264	0,252213031	0,044832649	0,059900717
Nodül_ekosu_hipoekoik	14,0086	0,000181974	0.244675091	0,044873330	0,047133856
Yaş	13,1995	0,153781590	0,237504633	0,041148317	0,013988721
Nodül özelliği	12,6401	0,000377544	0,232417420	0,042537586	0,055662569
Neutrophil	10,3074	0,244108097	0,209878139	0,035305603	0,015928567
Lenfosit	08,5661	0,478250535	0,191330266	0,027049031	0,009841453
Platelet	08,3090	0,503326869	0,188437631	0,027880116	0,010433369
Şüpheli_nodülün_çapı_1	07,8023	0,005217922	0,182601328	0,025373259	0,031678594
MCHC	07,1342	0,522223805	0,174608397	0,025303551	0,011154391
USG_iyi huylu	06,9476	0,008393041	0,172309958	0,023128914	0,035051724

4.3.3. Lojistik regresyon modeli uygulaması

Verilerin ham halleri üzerinden veri setlerinin hangi makine öğrenmesi modeline uygun olacağı ile ilgili net bir cevap yoktur. Ancak en iyi cevabın verilmesi, verilerin uygun veri setleri haline gelmeleri aşamalarında çeşitli veri tiplerinin dönüştürülmeleri, ölçeklendirilmeleri, özniteliklerin seçimi ve makine öğrenmesi modelleri/teknikleri ile defalarca denenmesi sonucunda mümkündür. Ortaya konulacak en iyi sonuç, verilerin elde edilmesinden sonuca kadar takip edilen adımlar içinde kullanılan çeşitli testlere, algoritmalara ve deneylere tabidir. Makine öğrenmesi

altında hangi model kullanılırsa kullanılsın, verilerin hazırlık süreci sonuçları mutlaka etkileyecektir. Veri tipi dönüşümlerinin yapılması, birbirleri ile uyumlu hale getirilmeleri ve çeşitli testlere tabi tutulmaları gerekmektedir. Veri setlerinde elde edilmek istenen sonuç (bağımlı değişken) evet/hayır, kanserli/ kansersiz, 1/0 gibi değişkenlerden oluşuyorsa, analizlerde kullanılacak algoritmaların doğrusal yerine doğrusal olmayan yöntemler kullanması uygun olacaktır.

Eldeki tiroit kanseri verileri, kategorik verilerin nümeriğe dönüştürülmesinden sonra sürekli ve kesikli veri tiplerinden oluşmaktadır. Sonucun ikili değer alması nedeniyle doğrusal olmayan yöntemlerden biri olan lojistik regresyon modelinin kullanılması yerinde olacaktır.

Lojistik regresyon modeli kullanılmadan önce hazır hale getirilen veri seti çeşitli öznitelik seçme algoritmalarına tabi tutuldular. Bu algoritmaların tamamında ortak öznitelikler ortaya çıksa da önem sıralaması değişebilmektedir.

Oluşturulan farklı alt veri kümeler varsayılan parametreler ve hiper-parametreler üzerinden lojistik regresyon denemelerine tabi tutuldu. Ortaya çıkan sonuçlar karışıklık matrisi tablosu üzerinden değerlendirmeye tabi tutularak sonuçlar doğruluk, kesinlik veya pozitif tahmin, anımsama, duyarlılık veya gerçek pozitif, spesifiklik/özgüllük oranları üzerinden hesaplanmıştır.

Bağımsız değişkenlerden en uygun öznitelikler olan *Sitoloji_sonucu_iyi_huylu* ve USG_LAP kullanılmış ve bu öznitelikler üzerinden lojistik regresyon algoritması çalıştırılarak tahmin yüzdeleri oluşturulmuştur.

Sonuçlar için sınırlılığın varsayılan olarak 0,50 alındığında elde edilen sayısal sonuçların tamamı Tablo 4.10'daki karışıklık matrisinde gösterilmiştir. *0,61* sınırlılığında elde edilen en iyi *Spesifiklik* sonucunu veren karışıklık matrisi ise Tablo 4.11'de verilmiştir. Tablo 4.11'de verilen oranlar elde edilirken ortaya çıkan sonuca yönelik olasılıklar tablosunun bir kesiti Tablo 4.12'de verilmiştir.

Tablo 4.10. 0,50 sınırlılığında oluşan sonuçlar için karışıklık matrisi tablosu.

	Başarılı-Gerçek (Suc-Obs)	Başarısız-Gerçek (Fail-Obs)	Toplamlar
Başarılı-Tahmin (Suc-Pred)	69 (TP)	33 (FP)	102 (PP)
Başarısız-Tahmin (Fail-Pred)	21 (FN)	111 (TN)	132 (PN)
	90 (OP)	144 (ON)	234 (TOPLAM)
Cutoff: 0.50	0.766667	0.770833	0.769231
tahmini pozitif	tahmini negatif	gözlemlenen pozitif	gözlemlenen negatif
(PP = TP + FP)	(PN = FN + TN)	(OP= TP+FN)	(ON: FP+TN)
TP: Doğru pozitif	TN: Doğru negatif	FN: Yanlış negatif	FP: Yanlış pozitif

Tablo 4.11. 0,61 sınırlılığında oluşan sonuçlar için karışıklık matrisi tablosu.

	Başarılı-Gerçek (Suc-Obs)	Başarısız-Gerçek (Fail-Obs)	Toplamlar
Başarılı-Tahmin (Suc-Pred)	27 (TP)	1 (FP)	28 (PP)
Başarısız-Tahmin (Fail-Pred)	63 (FN)	143 (TN)	206 (PN)
	90 (OP)	144 (ON)	234 (TOPLAM)
Cutoff: 0.61	0.30	0.993055556	0.726495726
tahmini pozitif	tahmini negatif	gözlemlenen pozitif	gözlemlenen negatif
(PP = TP + FP)	(PN = FN + TN)	(OP= TP+FN)	(ON: FP+TN)
TP: Doğru pozitif	TN: Doğru negatif	FN: Yanlış negatif	FP: Yanlış pozitif

Tablo 4.12. 0,61 sınırı ile sonuçlanan olasılıklı regresyon tablosu.

A	B	C	D	E	F
Bağımlı değişken (Post-op)	p-tahmin (p-Pred)	Başarılı tahmin (Suc-Pred)	Başarısız tahmin (Fail-Pred)	LL	% Doğruluk (%Correct)
1	0.152314935	0.152314935	0.84768506	-1.8818	0
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.594414738	0.594414738	0.40558526	-0.90242	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.936449935	0.936449935	0.06355007	-2.75593	0
1	0.152314935	0.152314935	0.84768506	-1.8818	0

Tablo 4.12. (**Devamı**) 0,61 sınırı ile sonuçlanan olasılıklı regresyon tablosu.

A	B	C	D	E	F
Bağımlı değişken (Post-op)	p-tahmin (p-Pred)	Başarılı tahmin (Suc-Pred)	Başarısız tahmin (Fail-Pred)	LL	% Doğruluk (%Correct)
0	0.152314935	0.152314935	0.84768506	-0.16525	100
1	0.594414738	0.594414738	0.40558526	-0.52018	0
1	0.936449935	0.936449935	0.06355007	-0.06566	100
0	0.594414738	0.594414738	0.40558526	-0.90242	100
1	0.594414738	0.594414738	0.40558526	-0.52018	0
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
0	0.152314935	0.152314935	0.84768506	-0.16525	100
1	0.602632139	0.602632139	0.39736786	-0.50645	0
…..	……..	……	……	……	…….

%50 sınırlılık değeri üzerinden hesaplanan ve Tablo 4.10'da karışıklık matris tablosunda gösterilen sonuçlara göre lojistik regresyon performans ölçüm değerleri; *Doğruluk oranı (Accuracy rate)* = (TP+TN) / (TP+TN+FP+FN) = (69 + 111) / 234 = *0.7692*

Kesinlik veya Pozitif Öngörü Değeri (Precision or the Positive Predictive Value) =TP/(TP+FP) = 69 / 102 = *0,6765*

Anımsama, Duyarlılık veya Gerçek Pozitif Oran (Recall or Sensitivity or True Positive Rates)=TP/(TP+FN) = 69 / 90=*0,7667*

Spesifiklik/Özgüllük veya gerçek negatif oran (Specificity or True Negative rate) = TN/(TN+FP) = 111/144=*0,7708* olarak hesaplanmıştır.

Benzer şekilde lojistik regresyon algoritmasında 0,61 sınırlılık değeri üzerinden hesaplanan ve Tablo 4.11'de gösterilen sonuçlara göre performans ölçüm değerleri;

Doğruluk oranı (Accuracy rate) = (TP+TN) / (TP+TN+FP+FN) = (27 + 143) / 234 = *0,7265*

Kesinlik veya Pozitif Öngörü Değeri (Precision or the Positive Predictive Value) =TP/(TP+FP) = 27 / 28 = *0,9643*

Anımsama, Duyarlılık veya Gerçek Pozitif Oran (Recall or Sensitivity or True Positive Rates)=TP/(TP+FN) = 27/90= *0,30*

Spesifiklik/Özgüllük veya gerçek negatif oran (Specificity or True Negative rate) = TN/(TN+FP) = 143/144= *0,9931* olarak hesaplanmıştır. Uygulama sonucu ortaya çıkan en yüksek *Özgüllük* (Specificity) değeri *0,61* sınırlılık ile elde edilmiş ve Tablo 4.13'de son sütunda gösterilmiştir.

Tablo 4.13. 0,61 sınırlılık ile elde edilen özgüllük sonuçları.

Normalizasyon	Kullanılan testler	En önemli öznitelikler	En iyi öznitelikler	Özgüllük (sınır:%61)
Absulute-Max Min-Max Normalization Robust Scaling Standardization	Chi2:Stat, p-Value, Welch: p-Value, F: Stat, p-Value, Fisher Score Gini Index, Pearson: Stat, p-Value Kendall: Stat, p-Value Spearman: Stat, p-Value	Sitoloji_sonucu_iyi huylu USG_LAP_Mevcut Sitoloji_sonucu_kötü_huylu USG_kötü_huylu USG_olası_iyi_huylu Sitoloji_sonucu_olası kötü huylu	Sitoloji_sonucu_iyi huylu USG_LAP_Mevcut	%99,31
Absulute-Max Min-Max Normalization Robust Scaling Standardization	Chi2:Stat, p-Value, Welch: p-Value, F: Stat, p-Value, Fisher Score Gini Index, Pearson: Stat, p-Value Kendall: Stat, p-Value Spearman: Stat, p-Value	Sitoloji_sonucu_iyi huylu USG LAP_Mevcut Sitoloji_sonucu_kötü huylu USG_kötü huylu USG_olası_iyi huylu Sitoloji_sonucu_olası kötü huylu	Sitoloji_sonucu_iyi huylu USG_LAP_Mevcut	%99,31
Absulute Max Min-Max Normalization Robust Scaling Standardization	Mutual Information	Sitoloji_sonucu_iyi huylu Sitoloji_sonucu_kötü huylu USG LAP_Mevcut USG_kötü huylu USG_olası_iyi huylu Sitoloji_sonucu_olası_kötü huylu	Sitoloji_sonucu_iyi huylu Sitoloji_sonucu_kötü huylu	%99,31
Absulute Max Min-Max Normalization Robust Scaling	Gain Ratio	Sitoloji_sonucu_kötü huylu USG_kötü huylu USG LAP_Mevcut Sitoloji_sonucu_iyi huylu USG_olası_iyi huylu Sitoloji_sonucu_olası_kötü huylu	Sitoloji_sonucu_kötü huylu USG_kötü huylu	%98,61

4.3.4. Deney sonuçlarının yorumlanması

Ele alınan tiroit kanser verilerinin lojistik regresyon modeline uygulanmasında Excel kullanılmıştır. Excel hücre içi formülasyon özelliği nedeni ile sınırlılık değişimlerinde diğer hücreleri otomatik olarak güncellemesi, modelin her seferinde baştan çalıştılmasına gerek duymaması, veri madenciliği ve makine öğrenmesi uygulamaları için gerekli uygun eklentilerle uyumlu çalışması ve kısa sürede sonuç vermesi nedenleri ile tercih edilmiştir. Öznitelik seçimleri için sıralı ve ileri adımlı seçim (Squential Forward Selection) yöntemleri kullanılmıştır.

Model uygulamalarına uygun hale getirilen tiroit kanseri veri setine uygun en iyi makine öğrenmesi algoritmalarının seçimi için literatür taraması yapılmıştır. Şekil 2.3'te gösterilen en önemli makine öğrenmesi sıralaması göz önüne alınmış ve ikili sonuç ihtiva eden veri setleri için uygun olan lojistik regresyon modeli eldeki veri seti için de uygun olduğu kararlaştırılmıştır. Lojistik regresyon model uygulamasında elde edilen sonuçlar karışıklık matris tabloları olarak bir önceki alt bölümde Tablo 4.10 ve Tablo 4.11'de verilmiştir.

Bir modelin uygun model olup olmadığını değerlendirmenin birçok yöntemi vardır. Lojistik regresyon modelinde, sonuç üzerinde yapılan değerlendirmelerden biri de Spesifiklik (Specificity) analizidir. Üzerinde çalışılan model bağlamında spesifiklik, belirli bir duruma ait olma anlamında ele alınmıştır. Bu yaklaşımla spesifiklik, kansersiz (0) olma sonuçları üzerinden ele alınmıştır.

Spesifiklik ("gerçek negatif oran" olarak da adlandırılır), doğru bir şekilde tanımlanan negatiflerin (kansersiz-0) oranını ölçer (örn. hastalığa sahip olmadığı doğru şekilde tanımlanan ve gereksiz ameliyat olan sağlıklı insanların yüzdesi).

Lojistik regresyon ile elde edilen sonuçlar karışıklık matris tablosu üzerinden yorumlanmadan önce, belirtilmelidir ki çıktı sonucu olan *postopSonuc* bağımlı değişkeni, ameliyat gerçekleştikten sonra alınan parçaların patoloji sonuçlarıdır. Diğer bir değişle *postopSonuc* bağımlı değişken sonucu 1 yada 0 olup olmamasına bakılmaksızın veri setinde ki her bir hasta hali hazırda ameliyat olmuş durumdadır. Buna göre patoloji test sonucu (postopSonuc) kansersiz (0) çıkan hastaların ameliyat olmaması gereken hastalar olduğu ancak ameliyat edildiği, dolayısı ile verilen ameliyat kararının yersiz olduğu ortaya çıkmaktadır. Bu yaklaşımla bakıldığında karışıklık matrisindeki metriklerden biri olan spesifiklik göz önünde bulundurularak

kansersiz (0) veriler spesifik olarak ele alınmalı ve bu veriler üzerinden yorumlamalar yapılmalıdır.

Sonucu kanserli ya da kansersiz olarak tayin eden lojistik regresyon algoritması sınıflandırma tayinini olasılıklar üzerinden yapmaktadır. Olasılığı yüzde olarak açığa çıkaran bu model, kanserli ya da kansersiz vakaların hangi yüzdelik dilimine girdiğine karar verir. Bu kararın sonucu, tayin edilecek yüzdeliğin sınırına bağlı olarak değişecektir. Yüzdelik sınırının iyi tayini ve özniteliklerin uygun seçilmiş olması, göz önüne alınan ölçümler açısından yüksek sonuçlar elde edilmesini sağlayacaktır. Lojistik regresyon ile elde edilen ihtimaller belirlenen sınırlılığın (varsayılan sınırlılık parametresi 0,50) altında ise kansersiz (postopSonuc kansersiz ise), belirlenen sınırlılığın üzerinde ise kanserli (1) sonucu verecektir.

Tiroit kanser vakaları özelinden gidersek, burada amaç gereksiz ameliyatların önüne geçmek olacağından bu modelde kansersiz (kanser olmama) sonuçları üzerinden gidilmiştir. Bu nedenle, her bir veri seti ve farklı sayıda öznitelik kombinasyonları kullanılarak sınırlılık %50 sınırından başlatılmış ve birer yükseltilerek her seferinde model çalıştırılmış ve sonuçlar elde edilmiştir. Elde edilen sonuçlar Tablo 4.10'da gösterilmiştir. %61 sınırlılık (cut-off) ile elde edilen olasılıklı sonuçların bir kısmı Tablo 4.12'de kesit olarak verilmiştir. Tablo 4.12'deki verilere göre F sütunundaki değerler elde edilirken, A sütunundaki bağımlı değişken kanserli (1) ve B sütunu değeri 0,61 sınırlılık oranının altında ise F sütununda karşılık gelen %Doğru (%Correct) değeri %0'dır ve tahmin doğru çıkmamış demektir ve eğer B sütunu değeri 0,61 sınırlılık oranının üstünde ise %Doğru (%Correct) değeri %100'dür ve tahmin doğru çıkmış demektir. Benzer şekilde A sütunundaki bağımlı değişken kansersiz (0) ve D sütun değeri 0,61 sınırlılık oranının altında ise F sütununda karşılık gelen %Doğruluk (%Correct) değeri %0 ve tahmin doğru çıkmamış ve eğer D sütunu 0,61 sınırlılık oranının üstünde ise F sütununda karşılık gelen %Doğruluk (%Correct) değeri %100 ve tahmin doğru çıkmış demektir. 0,61 sınırlılığında tüm satırlara uygulanan aynı kurallar sonrası ortaya çıkan tahmin sonuçları Tablo 4.11'da karışıklık matrisi tablosunda (Confusion matrix table) gösterilmiştir. Karışıklık matrisi göz önüne alınarak, model çeşitli metrikler ile yorumlanabilir. Bu metrikler anlamları ve sonuçları ile birlikte karışıklık matris tablosundan hemen sonrasından itibaren verilerek yorumlanmıştır.

Spesifiklik = gerçek negatifler / (gerçek negatif + yanlış pozitifler) denklemi ile hesaplanır. Tablo 4.11'te gösterilen sonuç verilerinden hareketle;

Spesifiklik = TN/ON = TN/(TN+FP) = 143 / (143+1) = *0,993055556* (yaklaşık %99,31).

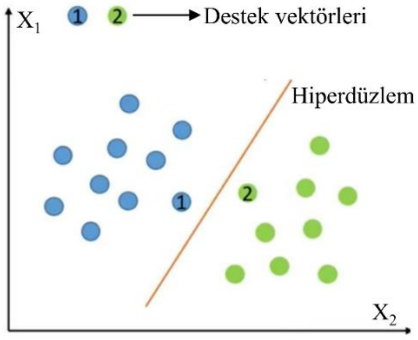
Buna göre lojistik regresyon modeli kullanılarak elde edilen 6 adet en iyi öznitelik arasından, denemelerle ortaya çıkarılan *Sitoloji_sonucu_iyi_huylu* ve *USG LAP* öznitelik ikilisi %99,31 ile en iyi spesifiklik sonucunu vermektedir. Bunun anlamı şudur: Eldeki veriler üzerinden tiroit kanser ön teşhisi konulanlar arasından postopSonuc (Son Operasyon) sonuçlarına göre her 100 kişiden 99,31i operasyon geçirmiş ancak kanserli olduğu yönünde ön-tanısının yanlış konulduğu ortaya çıkmıştır. Bu sonuç, ameliyat edilen 144 kanser olmayan hastanın 143'üne denk gelmektedir. Bunun sonucunda, verilere göre, kanser olmama ön tanısı, kullanılan lojistik regresyon modeli üzerinden sadece yaklaşık 1 kişide doğru çıkmamıştır. Diğer bir değişle model verileri kullanarak kanser olmayan 144 kişiden 143'ünü kansersiz olarak tespit etmiştir. Elde edilen sonucun yüksekliği yanında uzman görüşü alınarak seçilen özniteliklerin doğruluğu teyit edilmiştir. Buna göre Sakarya Üniversitesi Araştırma Hastanesi Genel Cerrahi Bölümü hocaları ile yapılan sonuç değerlendirmesinde elde edilen Sitoloji_sonucu_iyi huylu ve USG LAP öznitelikleri, ameliyat öncesi İİA testi yapma konusunda karar almada göz önüne aldıkları girdi değişkenleri olduğu teyit edilmiştir.

4.4. Destek Vektör Makineleri Sınıflandırıcısına Dayalı Tiroit Kanseri Tanısı

Destek vektör makineleri modeli denetimli öğrenme algoritmalarına dahil edilen makine öğrenmesi modellerinden biridir. Sınıflandırma ve regresyon problemlerinde kullanılmaya uygun bir algoritmaya sahip bu model tıbbi tanı problemlerinde de sıklıkla başvurulan modellerden biridir.

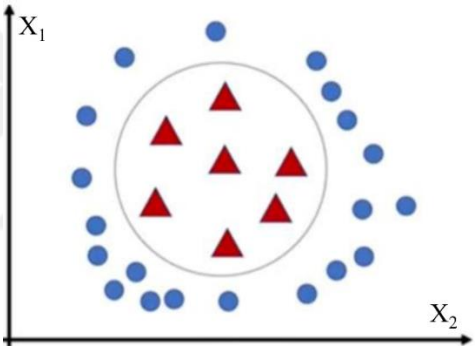
4.4.1. Destek vektör makineleri model yapısı

Destek vektör makineleri algoritması söz konusu olduğunda önemli iki kavram öne çıkmaktadır. Hiperdüzlem (Hyperplane) ve destek vektörleri (Support Vektors) adları verilen bu kavramlar destek vektör makineleri algoritmasını açıklamada önemli yer tutarlar. Şekil 4.3'de iki boyutlu düzlemde iki değişkenli verinin hiperdüzlem kullanılarak sınıflandırılması temsili olarak ifade edilmiştir.

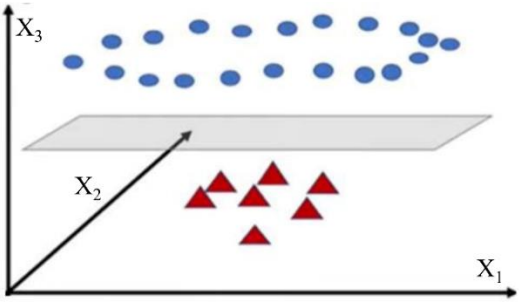


Şekil 4.3. Destek vektörlerinin hiperdüzlem ile sınıflandırılmaları.

Söz konusu veriler her zaman Şekil 4.3'de gösterildiği gibi dağılım göstermez. Şekil 4.4'te gösterildiği üzere iki boyutlu düzlemde bu türden dağılım gösteren veriler hiperdüzlem ile uyumlu bir seyir izleyemeyeceğinden üç boyutlu düzlemlerde temsil edilerek Şekil 4.5'te gösterildiği gibi sınıflandırmaya tabi tutulurlar. Üç boyutlu ortamda kullanılan hiperdüzlem de üç boyutlu olacaktır.



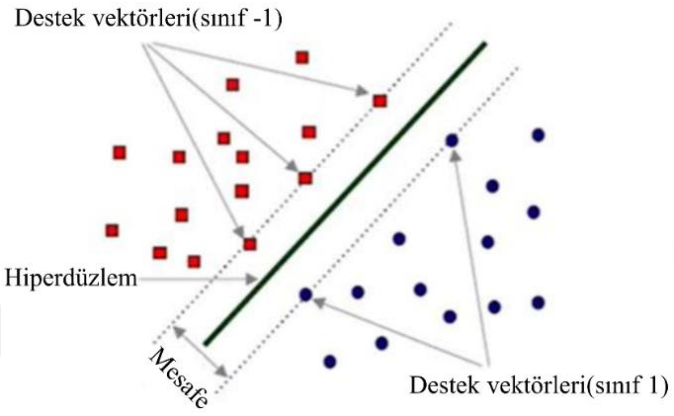
Şekil 4.4. İki boyutlu düzlemde destek vektör makineleri görünümü.



Şekil 4.5. Destek vektör makineleri modelinin üç boyutlu düzlemde temsili.

Hiperdüzlem Şekil 4.3'te gösterildiği gibi veri noktalarını iki ayrı sınıfa ayırmak üzere iki boyutlu düzlem için kullanılan temsili bir çizgiden ibarettir. Sözkonusu üç boyut olduğunda bu çizgi yüzeysel bir düzleme dönüşür.

Destek vektörleri ise hiperdüzlemin en yakınında bulunan veri noktalarını temsil ederler. Verilerin temsil ettikleri veri noktaları düzlemde yer değiştirdikçe hiperdüzlem de pozisyonunu değiştirir. Şekil 4.6'da temsil edilen ve oklarla gösterilen destek vektörleri (veri noktaları) hiperdüzleme olan mesafeleri gözönüne alınarak , hiperdüzleme yakın olan veriler üzerinden sınıflarını oluşturur ve diğer veri noktalarını da ilk sınıflandırdığı bu verilere nispeten ilgili sınıflara dahil eder.



Şekil 4.6. Veri noktaları ve hiperdüzlem ilişkisi.

Destek vektör makineleri algoritması kullanılarak oluşturulan model tercihlerinde dikkat tedilmesi gereken noktalar vardır. Veri setlerinin veri yapıları, öznitelik sayıları ve veriler arası ilişkilerin kompleks olup olmamasına bağlı olarak Vektör destek makineleri modelinin avantajları ve dezavantajları mevcuttur. Veri setinin küçük olduğu, verilerin net olarak bir birinden ayrıştırılabildiği ve öznitelik sayısının fazla olduğu durumlar için avantajlıdır ve tercih edilebilir. Diğer taraftan eğitim sürelerinin uzaması sebebi ile büyük boyutlu ve gürültülü veriler içeren veri setleri için uygun değildir.

4.4.1.1. Destek vektör makineleri modelinin matematiksel ifadesi

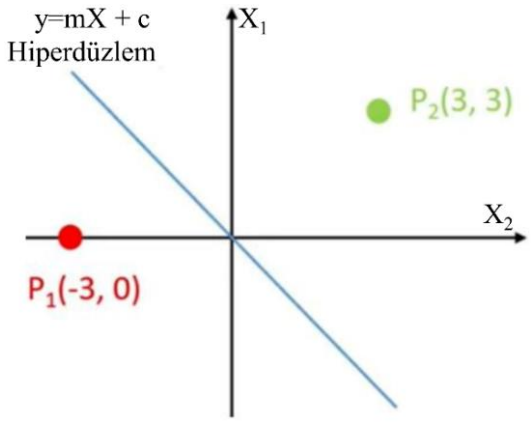
Destek vektör makineleri algoritmasında kullanılan veri noktaları her zaman doğrusal bir çizgi ile sınıflara ayrılmayabilirler. Bir önceki başlık altında verilen ve Şekil 4.3'te gösterilen doğrusal hiperdüzlem çizgisi gerçek hayat senaryolarında genellikle var olan bir durum değildir. Bu nedenle Şekil 4.4'te temsil edilen veri dağılımları Şekil 4.5'te gösterilen üç boyutlu ortama dönüştürülerek verilerin ayrıştırılması ve farklı sınıflandırmalara tabi tutulmaları sağlanır**.** Şekil 4.4'te gösterilen model durumundan Şekil 4.5'te temsil edilen çok boyutlu duruma geçiş için çekirdek (kernel) adı verilen matematiksel fonksiyon kullanılır. Böylelikle verilerin doğrusal yaklaşımlarla

ayrışmaları mümkün olmadığı durumlarda çekirdek fonksiyonu kullanılarak, öznitelikler çok boyutlu düzlemde sınıflara ayrılırlar. Burada çekirdek fonksiyonun yaptığı işlem iki boyutlu ortamı matematiksel olarak çok boyutlu ortama dönüştürmesidir.

Destek vektör makinelerinde hiperdüzlem çizgisi, destek vektörleri, destek vektörler arasındaki mesafe ve destek vektörlerin dışındaki ayrıştırılabilir diğer tüm veriler arasındaki ilişkiler matematiksel olarak ifade edilebilir. Burada hiperdüzlem çizgisi mesafe uzunluğunun tam ortasına tekabül edecek şekilde yerini alır. Şekil 4.6'da gösterildiği üzere verileri temsilen verilen dörtgen şekilli veri noktalarından hiper düzleme en yakın veriler üzerinden hiperdüzleme bir paralel çizgi çekilir. Aynı şekilde hiperdüzlemin diğer tarafındaki veri noktalarından kendisine en yakın olan veri/veriler üzerinden paralel bir çizgi çizilir. Çizilen bu paralel çizgilerden üst çizginin üstünde kalan veri noktaları -1'e ve diğer paralel çizginin altında kalan veri noktaları ise 1'e ait olacak şekilde matematiksel sınıflandırmaya tabi tutulurlar.

Veri noktalarının bulunduğu alan grafik düzlemde doğrusal bir denklemle ifade edilecekse, $y = mx + c$ şeklinde ifade edilebilir. Bu denklem iki boyutlu düzlemde hiperdüzlem çizgisini ifade eder ve doğrusal bir yapıya karşılık gelir.

Şekil 4.7'de gösterilen doğrusal denklem çizgisinin eğimi (m) -1 ve sabit sayı c'ye karşılık gelen değer ise sıfırdır. Denklem çizgisinin parametrelerini w ile ifade edersek, w'ye ait parametreler olan m ile c -1 ve 0 olarak ele alınabilir. Buna göre w parametreleri $(m,c) = (-1,0)$'dir ve çeşitli matematiksel algoritmalarda kullanılan ağırlığa karşılık gelecektir.

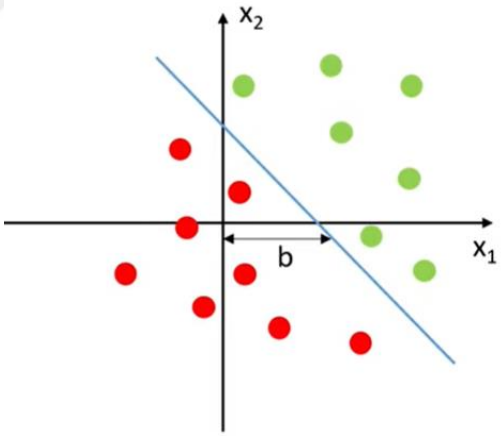


Şekil 4.7. İki boyutlu hiperdüzlem çizgisini temsil eden denklem gösterimi.

Grafik düzleminde verilen iki ayrı noktanın verilen denklem çizgisine nazaran durumunu irdeleyelim. Buna göre *(-3,0)* noktası ile ağırlık olarak kullanılan (-1,0) parametre işleme dahil edildiğinde;

$W^TX = \begin{bmatrix} -1 \\ 0 \end{bmatrix} [-3 \quad 0]$ ilişkisi oluşturulur. Burada dikkat edilmesi gereken iki önemli kuraldan biri parametrelerin matrise dönüştürülmeleri ve matris işlemlerinde çarpma işleminin yapılabilmesi için denklem çizgisi parametrelerinin matrise dönüştürülüp devriğinin alınması gerekmektedir.

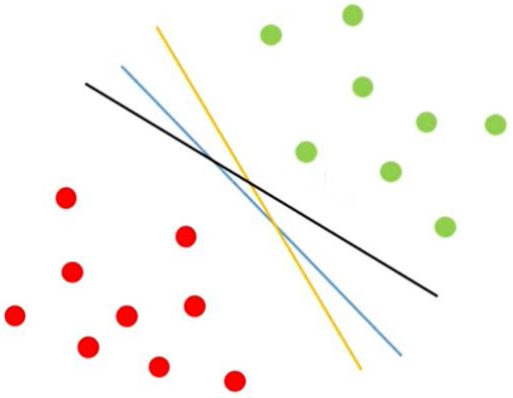
$W^TX = \begin{bmatrix} -1 \\ 0 \end{bmatrix} [-3 \quad 0]$ işlemi gerçekleştirildiğinde, W^TX sonucu 3 (pozitif sayı) verir. Buradan hareketle, Şekil 4.7'te verilen denklem çizgisinin solunda bulunacak tüm noktalar yapılan matris işlemi ile pozitif bir sonuç verecektir. Benzer şekilde denklem çizgisinin sağında var olan her hangi bir nokta işleme sokulduğunda ortaya çıkacak sonuç negatif olacaktır. Bu sonuçlar denklem çizgisinin orijinden (0,0) geçmesi halinde elde edilirken Şekil 4.8'deki denklem çizgisinde ise meydana çıkan *b*(sapma) W^TX sonucuna eklenir.



Şekil 4.8. Sapmanın denkleme eklenmesi.

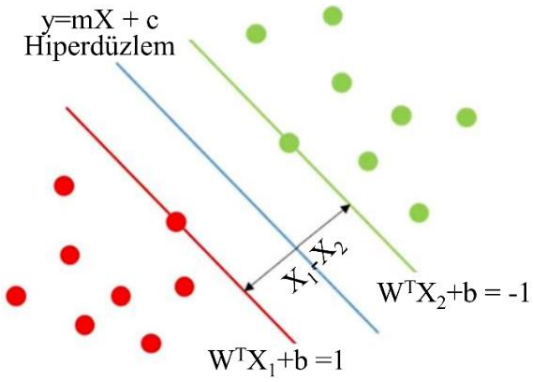
Denklem çizgisinin (Hiperdüzlem) hangi pozisyonda olduğuna göre hesaplanan sonuçlar her bir nokta için farklı hiperdüzlemler üzerinde farklı sonuçlar verecektir. Bu sonuçlar eldeki veri setinde sınıflara (sonuçlar) karşılık geldiği oranda başarılı olacaktır. Bu durumda veri noktalarının her birinin sınıflara denk gelmesi için en uygun hiperdüzlemin (denklem çizgisi) hangisi olduğu araştırılır. Her seferinde eldeki hiperdüzlem ile veri noktaları denklem içinde reaksiyona sokulur ve bu denklem çizgilerinden en optimum olanına ulaşana kadar işlemler yürütülür. En iyi hiper

düzlemin bulunması sırasında mesafe'nin (margin) olabildiğince maksimize edilmesi sonucu en iyi duruma getirecektir (Şekil 4.9).



Şekil 4.9. Araştırılan en uygun hiperdüzlemin farklı pozisyonlarda gösterimi.

Hiperdüzlemin pozisyonu, işleme giren veri noktaları üzerinden çizilen paralel çizgiler arası mesafenin maksimize edilmesi ile elde edilen orta çizgi ile belirlenir. Buna göre bir sınıflandırma problemi olarak veri noktalarının denklem üzerinde yapılan işlemleri ile elde edilen sonuçlar ya 1 yada -1 sınıfına ait olacaktır ve tüm sınıflandırma işlemleri bitirildiğinde ortaya çıkan en iyi sonucu belirleyen düzlem denklemi temsil eden en iyi hiperdüzlem çizgisi olacaktır (Şekil 4.10).



Şekil 4.10. Hiperdüzlem araştırılmasında sınıflandırma ve mesafe ilişkisi.

Şekil 4.10'da mesafeyi (margin) temsilen hesaplanan $X_1 - X_2$ sonucu $W^TX_1+b = 1$ denkleminden $W^TX_2+b = -1$ denkleminin çıkarılması ile elde edilir. Yapılan çıkarma işlemi sonucu elde edilen $W^T (X_1 - X_2) = 2$ eşitliğinde bir vektör olan W^T 'yi eşitliğin sol tarafından kaldırılabilmek için eşitliğin her iki tarafı vektörün sayısal değerine bölünmelidir. Bu durumda elde edilen sonuç Denklem 4.3'te gösterildiği gibi

olacaktır. Destek vektör makineleri algoritmasında sonucu en iyi verecek hiperdüzlem tespitinde bu marjin sonucunun olabildiğince maksimum yapılması gereklidir.

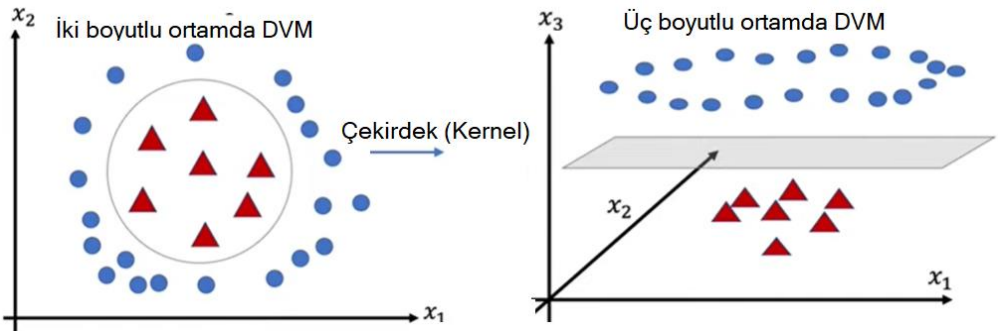
$$(\mathrm{X1} - \mathrm{X2}) = \frac{2}{||\mathrm{W}||} \tag{4.3}$$

Veri setindeki tüm veri noktalarının işleme alınması ile elde edilen sonuçlar aynı zamanda hangi sınıfa ait olduğunu belirleyecektir. Denklemde yerine konularak yapılan hesaplamada -1 den daha küçük sonuçlar için sınıf -1 iken sonucun 1 ve daha büyük bir sayısal sonuca tekabül etmesi halinde sınıfın 1'e denk düştüğünü söyleyebiliriz (denklem 4.4).

$$y_i = \begin{cases} -1, \ W^T X_1 + b \leq -1 \\ 1, \ W^T X_1 + b \geq 1 \end{cases} \tag{4.4}$$

4.4.1.2. Destek vektör makineleri algoritmasında Çekirdek (Kernel) işlevi

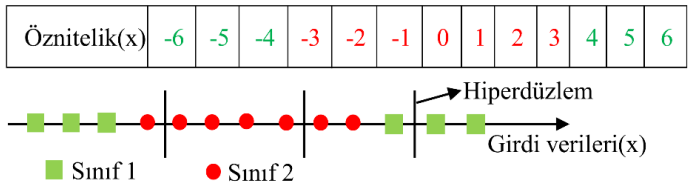
Destek vektör mekineleri algoritması uygulamalarında ele alınan ikili sınıflandırmalar her zaman doğrusal hiperparametre çizgisi ile sonuçlanmaz. Hiperparamere düzlemi sınıflandırma yapılacak girdi değişkenlerinin dağılımına göre düzlem yapısını üç boyutlu düzleme taşımak zorunda kalabilir. Düzlem değişikliği doğrusal olan yapıdan doğrusal olmayan yapıya geçişlerde *Çekirdek* (kernel) adı verilen matematiksel fonksiyonlar kullanırlar. Bu fonksiyon, veriler doğrusal bir hiperdüzlem denklem çizgisinin sağ ve solunda dağılım gösterdiğinde uygun bir sınıflandırmayı gerçekleştiremediği durumlarda, farklı sınıflara ait verilerin grafikte altta ve üstte olacak şekilde üç boyutlu bir düzleme taşınması işlevini gerçekleştirir ve bu işlev sonucunda sınıflandırma kolaylaşacaktır (Şekil 4.11).



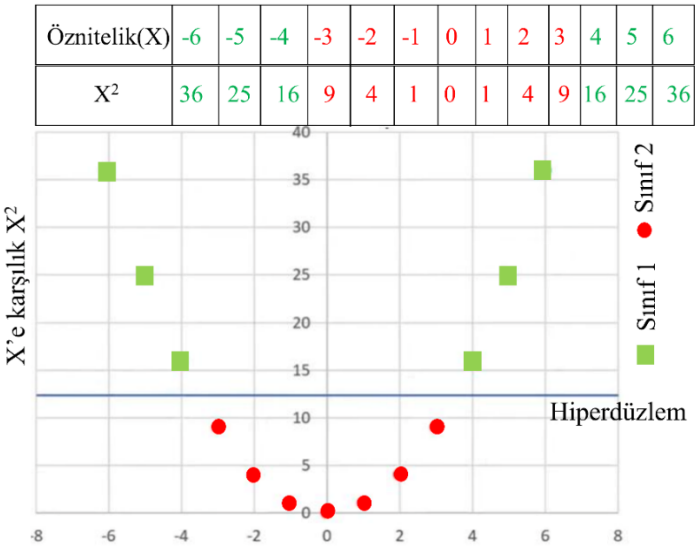
Şekil 4.11. Destek vektör makineleri algoritmasında çekirdek işlevi.

Çekirdek fonksiyonu genellikle eğitim veri setini dönüştürür, böylece doğrusal olmayan bir karar yüzeyi daha fazla sayıda boyut alanında doğrusal bir denkleme dönüştürülebilir. Standart bir girdi verisi boyutunda iki nokta arasındaki fonksiyonel işlem sonucunu döndürür. Bu fonksiyonel işlem iki vektör arasındaki matemetiksel büyüklük ve aralarındaki açının kosünüsüne bağlı olarak sonuç verir.

Şekil 4.12 üzerinde verilen girdi bağımsız değişkenleri temsilen kare şeklindeki veri noktalarının Sınıf 1'e ve yuvarlak şeklindeki veri noktalarının Sınıf 2'ye ait olduklarını varsayalım. Bu durumda doğrusal olarak hiperdüzlem çizgisi nereden çizilirse çizilsin kare ve yuvarlak şekilli veri noktaları bir birlerinden tamamı ile ayrılamazlar. Bu durumda her bir girdi verisinin karesi alınarak oluşturulan yeni girdi değerleri ile oluşacak veri noktaları grafiksel gösterimde doğrusal olmayan bir dizilişe sahip olacaklardır. Şekil 4.13 üzerinde gösterildiği gibi kareleri alınan veri noktaları kare ve yuvarlak şekilli noktalar olarak farklı bölümlerde var olacaklar ve hiperdüzlem çizgisinin bir levha şeklinde rahatlıkla çizilmesine olanak sağlayacaklardır. Böylelikle hiperdüzlem çizgisi her iki sınıfa ait olan girdi verilerini tamamı ile ve rahatlıkla ayırabilecek bir koordinatta var olabilecektir.



Şekil 4.12. Girdi veri noktaları ve hiperdüzlem ilişkisi.



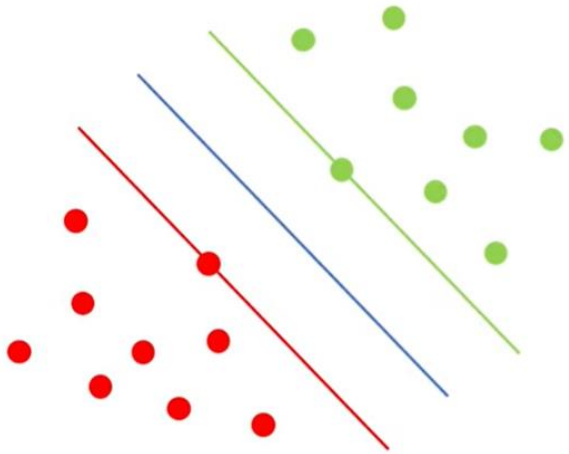
Şekil 4.13. Girdi verilerinin kareleri ve grafiksel düzlemde dizilişleri.

Farklı matematiksel formüllerle oluşturulan çeşitli DVM çekirdek fonksiyonları mevcuttur. Bunlar *Doğrusal* (Linear), *Polinom-Çok terimli* (Polynomial), *Radyal tabanlı fonksiyon* (Radial basis function) ve *Sigmoid* fonksiyonları olarak adlandırılırlar.

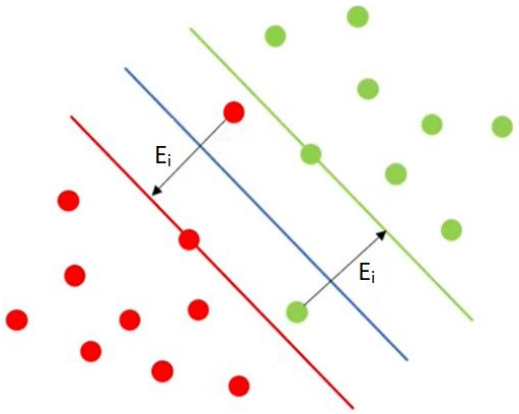
Veri setinde bulunan özniteliklerin karmaşıklığı, sayısal çokluğu, birbirleri ile olan ilişkilerinin durumu göz önüne alınarak DVM çekirdek fonksiyonu seçilebilir. Veri setlerinin doğrusal çekirdek fonksiyondan başlanarak tüm çekirdek fonksiyonlarının denenmesi ve sonuçların doğruluk (Accuracy) oranları açısından yorumlanması tavsiye edilebilir.

4.4.1.3. Destek vektör makineleri algoritmalarında kayıp fonksiyon işlevi

Denklem 4.3'te verilen marjin değerinin maksimum tutulması kaydı ile veri noktalarının Denklem 4.4'te yerlerine konulması ile elde edilen sonuçlar sınıfların tayinini sağlamaktadır. Ancak bazen veri noktaları hiperdüzlem çizgisinin iki tarafında düzgün dağılım göstermezler ve hiperçizginin her iki yanında çizilen çizgiler arasındaki marjin mesafesi içinde her biri diğer sınıfa ait veri noktaları mevcut olabilir. Şekil 4.14'te temsil edilen veri noktaları hiperdüzlemin her iki tarafında ayrı ayrı kümelenirken Şekil 4.15'te ayrı veri kümelerine ait olan verilerden bazıları marjin çizgilerini aşarak E_i hata mesafesi ile çizgiler arasında yer almışlardır.



Şekil 4.14. Verilerin hiperdüzlemin her iki tarafında kümelenmeleri.



Şekil 4.15. Marjin çizgisini aşan veri ve E_i hata mesafesi.

Katı marjin olarak adlandırlılan Şekil 4.14'deki mesafe hesabı gerçek hayat senaryolarında mümkün olmayabilir. Veri noktalarının bu kadar ayrışarak dağıldığını ve %100 oranla düzgün dağılımlarla belirgin sınıflara ayrıştıkları varsayılamaz. Bu türden bir marjin üzerinden model çalıştırıldığında aşırı uyum gösterme problemi ile karşılaşılabilir. Bu durumlarda marjin hesaplamalarında marjinle ilgili hataya tolerans gösterilir. E_i hata oranı gerçek hayat senaryolarında marjin çizgileri arasına düşebilecek veri noktalarının olabileceğini ve bu veriler üzerinden oluşacak hatalı mesafelerin ortaya çıkabileceğini gösterir (Şekil 4.15). E_i hata oranı dahil edilerek elde edilen marjin, *yumuşak marjin* olarak ifade edilir ve bu tür hesaplamalarda *kayıp fonksiyonu* kullanılır.

Kayıp fonksiyonu, tahmin edilen değerin gerçek değerden ne kadar uzakta olduğunu ölçer. Hangi modelin daha iyi performans gösterdiğini veya hangi parametrelerin daha iyi olabileceğini belirlemek için kullanılır. Kayıp fonksiyonun vereceği sonucun minimum olma düzeyi kullanılan modelin doğruluk oranının yüksekliğini ve dolayısı ile veri seti için en iyi sonucu veren modelin hangisi olduğunu ortaya koyacaktır.

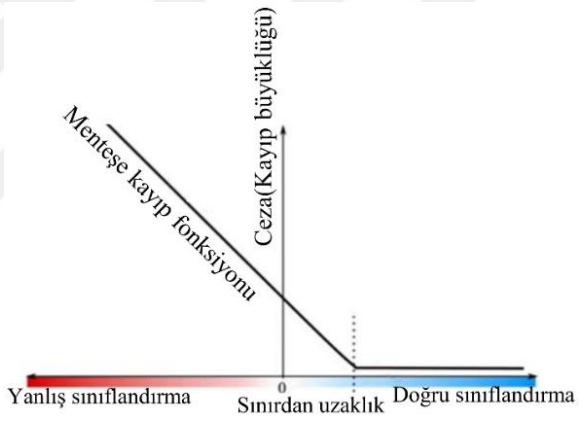
Bir çok doğrusal regreyon problemlerinde denklem 4.5'te verilen kayıp fonksiyonu denklemi destek vektör makineleri modellerinde Menteşe kayıp fonksiyonu olarak kullanılır. Kayıp fonksiyon çeşitlerinden biri olan Menteşe kayıp fonksiyonu maksimum marjin sınıflandırmalı modellerde kullanılır. Bu fonksiyonun kullanılma nedeni marjin mesafesinin maksimize edilme isteğidir.

$$Kayıp = \frac{1}{n}\sum_{i=1}^{n}(Yi - Yi')^2 \tag{4.5}$$

Menteşe kayıp fonksiyonu, sınıflandırma sınırından (hiperdüzlem çizgisi) bir marjı veya mesafeyi kayıp hesaplamasına dahil eder. Yeni gözlemler doğru şekilde sınıflandırılmış olsa bile, karar sınırından gelen marj yeterince büyük değilse bir ceza değeri ortaya çıkar. Denklem 4.6'da gösterildiği üzere menteşe kayıp fonksiyon *(MKF)* ifadesinde 0 yada 1 rakamlarından her biri ayrı ayrı kullanılmak üzere Y_i eldeki gerçek sonuçları temsil ederken $(W^T X_i + b)$ ifadesi ise $Y^{\wedge}_i$ olarak tahmin edilen sonuçtur. Burada 0 rakamı doğru sınıflandırma için kullanılırken 1 rakamı yanlış sınıflandırma için kullanılmaktadır.

$$MKF = (0{,}1 - y_i\,(W^T X_i + b\,)) \tag{4.6}$$

Şekil 4.16'da gösterildiği gibi kayıp fonksiyon denklemi doğru tahmin sınıfına yaklaştıkça hata büyüklüğünde azalma olurken tersine fonsiyon yanlış tahmin bölgesinde yüksek hata büyüklüklerine sahip olmaktadır.



Şekil 4.16. Fonksiyon ve tahmin ilişkisi ve hata büyüklükleri.

Gerçek sonuçların (Y_i) ile tahmini sonuçların ($Y^{\wedge}_i$) birlikte Denklem 4.6 işleme alındığında ortaya çıkan kayıp fonksiyon değerleri Tablo 4.14'da verilmiştir.

Tablo 4.14. Menteşe kayıp fonksiyonu ve ortaya çıkan değerler.

Yanlış Sınıflandırma		Doğru Sınıflandırma	
Y_i=1 ve $Y^{\wedge}_i$=-1 durumunda	Y_i=-1 ve $Y^{\wedge}_i$=1 durumunda	Y_i =1 ve $Y^{\wedge}_i$ =1 durumunda	Y_i=-1 ve $Y^{\wedge}_i$=-1 durumunda
K= (1-(1)(-1))	K= (1-(-1)(1))	K= (0-(1)(1))	K= (0-(-1)(-1))
K= (1 + 1)	K= (1 + 1)	K= (0 - 1)	K= (0 - 1)
K=2 (Yüksek kayıp değeri)	K=2 (Yüksek kayıp değeri)	K = -1 (Düşük kayıp değeri)	K=-1 (Düşük kayıp değeri)

Tablo 4.14'e göre gerçek sonuç 1 ve tahmin sonucu -1 yada gerçek sonuç -1 iken tahmini sonuç 1 ise sınıflandırma yanlış yapılmış demektir. Gerçek değerler ile tahmini değerler kayıp fonksiyonunda yerlerine yazıldığında yüksek bir sonuç değeri elde edilir. Bu sonuç aynı zamanda yanlış sınıflandırma yapılmış olduğunun da diğer bir ifadesidir. Benzer şekilde doğru sınıflandırmada fonksiyon işleme alındığında kayıp değeri düşük çıkmaktadır. Kayıp fonksiyon sonucunun düşük çıktığı oranda sınıflandırma için doğruluk değeri o denli yüksek çıkacaktır.

4.4.1.4. Destek vektör makinelerinde gradyan iniş fonksiyonu

Gradyan iniş fonksiyonu en iyi doğruluk değerini bulmak amacıyla destek vektör makineleri algoritmasında kullanılacak en iyi parametreleri bulmak için kullanılır. Bu parametreler *ağırlık* ve *sapma* olarak adlandırılırlar.

Gradyan iniş, çeşitli makine öğrenmesi algoritmalarında maliyet fonksiyonunu en aza indirmek için kullanılan bir optimizasyon algoritmasıdır (Şekil 4.17). Öğrenme adım değeri (Learning rate) parametrelerini güncellemek için de kullanılır.

w --> Ağırlık (Weights)

b --> Sapma (Bias)

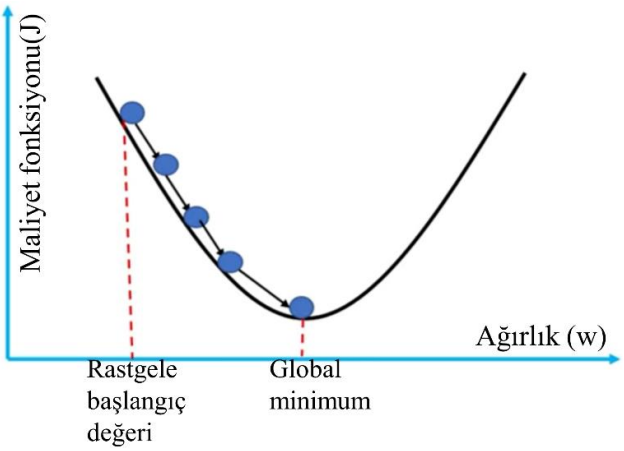
L --> Öğrenme adım değeri (Learning rate) olarak ele alındığında her iterasyonda kullanılacak ağırlıklar ve sapma değerleri Denklem 4.7 ve Denklem 4.8 üzerinden hesaplanırlar.

$$w_i = w_{i-1} - L * \frac{dJ}{dw} \tag{4.7}$$

$$b_i = \mathrm{b}_{i-1} - L * \frac{dJ}{db} \tag{4.8}$$

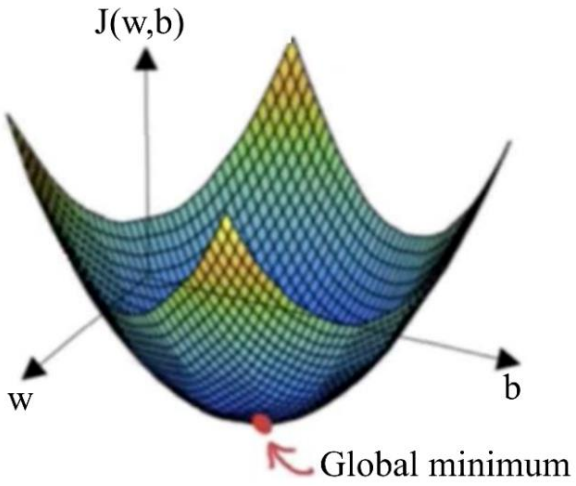
$\frac{dJ}{dw}$ --> *w'* ya göre maliyet fonksiyonunun kısmi türevidir.

$\frac{dJ}{db}$ --> *b*'ye göre maliyet fonksiyonunun kısmi türevidir.



Şekil 4.17. Maliyet fonksiyonu ve optimizasyon algoritması ile ilişkisi.

Gradyan iniş işleminde amaç en iyi sonuçlar için en iyi parametreleri bulmak olduğundan Şekil 4.17'de verilen grafiğe göre *X* düzleminde seçilen *w* noktasına nazaran ortaya çıkan maliyet fonksiyon değeri en aza indirgenene kadar *w* değeri güncellenir. Şekil 4.18'de ise *w* değeri ile birlikte *b* sapma değerinin güncellenmesi ile ortaya çıkan maliyet fonksiyonu gösterilmiştir. Buradaki maliyet fonksiyonu veri setinden alınan her bir veri noktası için oluşturulan kayıp fonksiyon değerlerinin ortalamasını temsil eder ve *w* ile *b* parametrelerinin tekrarlı şekilde değiştirilmesi ile farklı sonuçlar verir. Maliyet fonksiyonunun en az değer ile temsil edildiği noktada var olan parametre değerleri en iyi sonucu içeren en uygun değerleri temsil ederler (Global minimum).



Şekil 4.18. w ve b değerlerinin güncellenmesi ile ortaya çıkan maliyet.

Destek vektör makineleri algoritmaları için Gradyan iniş işlevi çalıştırıldığında iki duruma göre işlem yapılmaktadır. Hesaplamalarda $(y_i . (w.x + b) >= 1)$ ise $dJ/dw =$

$2\lambda w$ olurken $dJ/db = 0$ olmaktadır. Diğer yandan $(y_i . (w.x + b) < 1)$ ise $dJ/dw = 2\lambda w - y_i . x_i$ olarak hesaplanırken $dJ/db = y_i$ olarak alınır. Buradaki λ değeri rastgele bir değerdir.

Bu iki durum üzerinden hesaplanan türev sonuçları her iterasyonda Denklem 4.7 ve Denklem 4.8'de işleme alınarak her seferinde yeni ağırlıklar ve sapmalar hesaplanır. Modelde en iyi ağırlık ve sapma değeri ile en iyi doğruluk oranına erişene kadar bu işlemlere devam edilerek her seferinde ağırlık ve sapma değerleri değiştirilir.

4.4.2. Tiroit Kanser tanısında destek vektör makineleri modelinin kullanımı

Destek vektör makineleri sınıflandırma modellerinin çalıştırılması sırasında önemli işlevler bir araya getirilmelidir. Bundan önceki alt bölümlerde detaylıca açıklanan Hiperdüzlem çizgi denkleminin tanımlanması ve bu denklemin çalıştırılması sırasında gerçekleştirilen her bir iterasyon için denklem içinde bulunan ağırlıkların, sapmaların ve öğrenme adım büyüklüklerinin güncellenmeleri gereklidir. Tüm iterasyonlarda parametrelerin güncellenmesinin amacı doğruluk (Accuracy) oranının yükseltilmesi ve dolayısıyla kayıp fonksiyon sonucunun olabildiğince küçültülmesidir.

Detek vektör makineleri sınıflandırma modeli ile tiroit kanser tanısı koymak üzere 234 adet kanser ameliyatı yapılmış hasta verileri kullanılmıştır. Python yazılım kodlamaları ile Google Colab ortamı içinde tahmin modeli oluşturulmuştur. Python kodları kullanılarak oluşturulan modelde temel adımlar;

$Y = wX - b$ ile ifade edilen hiperdüzlem denkleminin oluşturulması adımı ile başlanmıştır. Buradaki Y elde etmeye çalışılan en iyi tahmini temsil ederken, w girdi verileri çarpanı olarak kullanılan ağırlığı ve b ise sapmayı temsil etmektedir.

Modelde önemli ikinci parça ise her iterasyonda w, b ve a (L-öğrenme oranı) parametrelerinin güncellenmesi için Gradyan iniş fonksiyonunun çalıştırılması adımıdır. Hiperdüzlem denkleminde parametre güncellemeleri için ayrıca yazılımda fonksiyon oluşturulmuştur. (Tablo 4.15).

Tablo 4.15. Tiroit kanser tanısında DVM kullanımı.

```
import numpy as np   # numpy kütüphanesinin girilmesi
#Destek Vektör makinesi Sınıflandırıcısının Fonksiyon olarak tanımlanması
class DVM_Siniflandirici():
  # Parametrelerin başlangıç atamaları
  def __init__(self, ogrenme_orani, iterasyon_sayisi, lambda_parametresi):
```

Tablo 4.15. (Devamı) Tiroit kanser tanısında DVM kullanımı.

```
    self.ogrenme_orani = ogrenme_orani
    self.iterasyon_sayisi = iterasyon_sayisi
    self.lambda_parametresi = lambda_parametresi
    def atama(self, X, Y):  # Veri setinin DVM sınıflandırıcısına atanması
    # m  --> Veri noktalarının sayısı --> Satır sayısı  # n  --> Öznitelik sayısı --> Veri setindeki sütun sayısı
    self.m, self.n = X.shape
    # Ağırlık ve sapma değerlerinin atanması
    self.w = np.zeros(self.n)
    self.b = 0
    self.X = X
    self.Y = Y
# Optimizasyon için Gradyan iniş algoritmasının işleme alınması
    for i in range(self.iterasyon_sayisi):
      self.agirlik_guncellemesi()
def agirlik_guncellemesi(self):  # Ağırlık ve Sapma değerlerini güncelleyen fonksiyon
  y_sonuc = np.where(self.Y <= 0, -1, 1)   # Sonuç sütununun koda dönüştürülmesi
    #Kısmi türevler alınarak Gradyanların oluşturulması( dw, db)
    for index, x_i in enumerate(self.X):
      durum = y_sonuc[index] * (np.dot(x_i, self.w) - self.b) >= 1
      if (durum == True):
        dw = 2 * self.lambda_parametresi * self.w
        db = 0
      else:
        dw = 2 * self.lambda_parametresi * self.w - np.dot(x_i, y_sonuc[index])
        db = y_sonuc[index]
      self.w = self.w - self.ogrenme_orani * dw
      self.b = self.b - self.ogrenme_orani * db
  # Verilen girdi verisine göre tahmin sonucu veren fonksiyon
  def tahmin(self, X):
    cikti = np.dot(X, self.w) - self.b
    tahmini_sonuclar = np.sign(cikti)
    y_hat = np.where(tahmini_sonuclar <= -1, 0, 1)
    return y_hat
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Tablo 4.15. (Devamı) Tiroit kanser tanısında DVM kullanımı.

```
# Veri setinin Pandas veri çerçevesine aktarılması
#tiroit_veri_seti = pd.read_csv('/content/tiroit_3ondalikPLUS-2tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_3ondalikPLUS-3tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_5ondalikPLUS-3tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_6ondalikPLUS-3tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_6ondalikPLUS-6tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_6ondalikPLUS-10tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_SOMEondalikPLUS-SOME-tam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTamami_sayisalTam.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-3.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroit_ondalik.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-HOT.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-HOT-3.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-HOT-4.csv')
#tiroit_veri_seti = pd.read_csv('/content/tiroitTAMSAYI-HOT-5.csv')
#tiroit_veri_seti = pd.read_csv('/content/model8.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk3.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk4.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk5.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk6.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk7.csv')
#tiroit_veri_seti = pd.read_csv('/content/ilk8.csv')
tiroit_veri_seti = pd.read_csv('/content/ilk9.csv')
# Veri setinin ilk 5 satırının görüntülenmesi
tiroit_veri_seti.head()
tiroit_veri_seti.shape  # Veri setinde var olan satır ve sütun sayıları
tiroit_veri_seti.describe()  # Veri setinin teşitli istatistiksel bilgileri
tiroit_veri_seti['postopSonuc'].value_counts()       # 0 --> Kansersiz 1 --> Kanserli
# Bağımsız girdi verileri ile bağımlı tahmin sütununun birbirinden ayrılması
girdiler = tiroit_veri_seti.drop(columns='postopSonuc', axis=1)
hedef_sinif = tiroit_veri_seti['postopSonuc']
print(girdiler)
print(hedef_sinif)
olceklendirici = StandardScaler() # Verilerin Standartlaştırılması
olceklendirici.fit(girdiler)
standartlastirilmis_veri = olceklendirici.transform(girdiler)
```

Tablo 4.15. (Devamı) Tiroit kanser tanısında DVM kullanımı.

```
print(standartlastirilmis_veri)
girdiler = standartlastirilmis_veri
hedef_sinif = tiroit_veri_seti['postopSonuc']
print(girdiler)
print(hedef_sinif)
# Veri setinin Eğitim ve Test verileri şeklinde ayrılması
X_egitim, X_test, Y_egitim, Y_test = train_test_split(girdiler, hedef_sinif, test_size=0.2, random_state = 2)
print(girdiler.shape, X_egitim.shape, X_test.shape)
# Destek vektör makinesi sınıflandırıcısı kullanılarak modelin eğitilmesi
siniflandirici=DVM_Siniflandirici(ogrenme_orani=0.001,iterasyon_sayisi=1000,lambda_parametresi=0.01)
# DVM sınıflandırıcısının Eğitim verileri kullanılarak eğitilmesi
siniflandirici.atama(X_egitim, Y_egitim)
# Doğruluk (Accuracy) değeri üzerinden Modelin değerlendirilmesi
# Eğitim verileri üzerinden Doğruluk oranı
X_egitim_tahmin = siniflandirici.tahmin(X_egitim)
egitim_verisinde_dogruluk = accuracy_score(Y_egitim, X_egitim_tahmin)
print('Accuracy score on training data = ', egitim_verisinde_dogruluk)
# Test verileri üzerinden Doğruluk oranı
X_test_tahmin = siniflandirici.tahmin(X_test)
test_verisinde_dogruluk = accuracy_score(Y_test, X_test_tahmin)
print('Accuracy score on test data = ', test_verisinde_dogruluk)
```

Destek vektör makineleri sınıflandırıcısı kullanılarak en iyi parametreler üzerinden elde edilmeye çalışılan doğruluk oranları veren python kodları Tablo 4.15'de satırlar halinde bir arada verilmiştir. Bu kod satırları üzerinden amaç kayıp fonksiyon sonucunu en aza indirmek ve doğruluk oranını en yükseğe çıkarmaktır. En iyi sonuç içinse *w*, *b* ve *L* parametrelerini her iterasyonda güncellemek hedefiyle Gradyan iniş işlemleri yapmak üzere fonksiyonlar oluşturulmuştur. Bu tablodaki kod satırlarına göre sırasıyla;

Python yazılım dilinde nümerik işlemler sözkonusu olduğunda başvurulan kütüphane olan numpy kütüphanesi yazılıma dahil edilmiştir.

Destek vektör makineleri sınıflandırıcısı sınıfı tanımlanmış ve bu sınıf içerisine her iterasyonda *w* (ağırlık), *b* (sapma), *L* (öğrenme oranı) ve λ (adım büyüklüğü) parametrelerinin güncellenmesi için fonksiyonlar tanımlanmıştır.

Öğrenme oranı, iterasyon sayısı ve menteşe gradyan iniş işlevi için gerekli lambda parametrelerinin ilk atamalarını yapan *_init_ (initial)* başlangıç fonksiyonu tanımlanmıştır.

atama adıyla yeni bir fonksiyon tanımlanmış ve bu fonksiyon ile eldeki veri seti girdi bağımsız değişkenler *X* ve bağımlı değişkenler *Y* simgeleri kullanılarak satırı *m* ve sütunu *n* ile ifade edilen matris türünden atanmıştır.

Ağırlık, sapma, bağımsız ve bağımlı değişkenlerin atamaları yapılmıştır ve gradyan iniş işlevi for döngüsü ile fonksiyon içindeki yerini almıştır.

Ağırlık parametrelerini güncelleyen bir fonksiyon tanımlanmıştır. Bu fonksiyona göre bağımlı değişken sonucunun 0,1 yada -1 olma durumuna göre farklı hesaplamalar yapılmaktadır. Gradyan iniş işlevi çalıştırıldığında iki duruma göre işlem yapar. Eğer hesaplamalarda $(y_i . (w.x + b) >= 1)$ *ise* $dJ/dw = 2\lambda w$ olurken $dJ/db = 0$ olmaktadır. Diğer yandan $(y_i . (w.x + b) < 1)$ ise $dJ/dw = 2\lambda w - y_i . x_i$ olarak hesaplanırken $dJ/db = y_i$ olarak hesaplanır. Buradaki λ değeri rastgele bir adım büyüklüğü değerdir. Duruma göre elde edilen *w* ve *b* parametreleri sonraki satırlarda yerlerine konularak yeni *w* ve *b* parametreleri elde edilir.

Yeni *w, b* parametreleri ve girdi verilerinin kullanımı ile hesaplanan $y=w*X-b$ denklem sonuçlarının her biri -1 den küçük yada -1'e eşitse, sınıf olarak 0'ı değilse sınıf sonucu olarak 1'i seçmesini sağlayan tahmin fonksiyonu oluşturulmuştur.

Python dilinin önemli kütüphanelerinden biri olan *pandas* kod satırlarına dahil edilmiştir. Pandas veri kümeleriyle çalışmak için kullanılan bir Python kütüphanesidir. Verileri analiz etmek, temizlemek, keşfetmek ve işlemek için çeşitli fonksiyonlara sahiptir. Devamında sırasıyla verilerin standartlaştırılması için *StandardScaler,* verilerin eğitim ve test amaçlı kullanımı için ayrıştırılmaları amacıyla *train_test_split* ve sonuçların doğruluk oranı açısından elde edilmesini sağlayan *accuracy_score* bileşenleri kütüphane kodları olarak eklenmiştir.

Çeşitli öznitelik seçim algoritmaları ile elde edilmiş alt veri seti kümelerinin girdi olarak uygulanması için farklı çok sayıda veri setleri tanımlanmış ve dosya olarak girdisi sağlanmıştır. Her seferinde bir adet veri seti işleme dahil edilebildiğinden biri hariç diğerleri kullanım dışı bırakılmak üzere geçici olarak açıklama satırı şeklinde kodlanmıştır.

Uygulanan veri setinin ilk beş satırını ve öznitelik adlarını ekrana yansıtan kod satırı

yazılmıştır. Kısa bir ön bilgi için gerekli olabilir. Veri setinin satır ve sütun sayılarını ekrana yansıtan kod satırı eklenmiştir. Istatistiksel anlamda veri seti hakkında daha fazla bilgi edinmeyi sağlayan ve veri setindeki bağımlı değişkenlerin değişken sayısını guruplama ve sayı olarak ekrana yansıtma işlemlerini yapan kod satırları da eklenmiştir.

Bağımsız girdi verileri ile bağımlı tahmin sütununun birbirinden ayrılması işlemi için kod satırı eklenmiştir. Bu ayrım, bağımsız girdi değişkenlerinin bağımlı değişkenlerden ayrıştırılarak eğitim ve test verileri olarak matrislerde gösterilmeleri için gereklidir. Ayrıştırılan veri setinde var olan sınıfların (bağımlı değişkenler) farklı bir değişkene atanması sağlanmıştır.

Veri setinde bulunan veriler aynı ölçekte değil ise sonuca olumlu etki göstermesi amacıyla ölçeklendirilmesi gerekebilir. Bu işlemi yapan ilgili satırlar girdi bağımsız değişkenlerin standartlaştırılması için yazılıma dahil edilmiştir.

Veri setinin eğitim ve test verileri şeklinde ayrıştırılması için gerekli satır eklenmiş ve destek vektör makinesi sınıflandırıcısı kullanılarak modelin eğitilmesi sağlanmıştır. Parametre olarak kullanılan rakamlar başlangıçtaki parametreler olup her iterasyonda değişmektedir. Bu parametreler *atama* fonksiyonuna gönderilerek işlem başlatılmaktadır.

Yazılımın son kısmında eğitim girdi verileri ile girdi test verileri *DVM_Siniflandirici* sınıfında bulunan *tahmin* fonksiyonuna gönderilerek *Doğruluk* (Accuracy) yüzdeleri elde edilmiştir.

Tablo 4.15‘de python kodları verilen yazılım bloğunda girdi olarak verilen veri setleri kullanılarak sonuçlar elde edilmiş ve bu sonuçlar farklı veri seti alt kümeleri için *Doğruluk* ölçüsü olarak Tablo 4.16’te bir arada verilmiştir.

Tablo 4.16. Tiroit kanseri alt veri setleri ile elde edilen doğruluk değerleri.

Veri Setleri	Egitim_dogruluk_tahmin	Test_dogruluk_tahmin
tiroit_3ondalikPLUS-2tam.csv	0.7860	0.7659
tiroit_3ondalikPLUS-3tam.csv	0.7754	0.7659
tiroit_5ondalikPLUS-3tam.csv	0.7860	0.7659
tiroit_6ondalikPLUS-3tam.csv	0.7754	0.7446
tiroit_6ondalikPLUS-4tam.csv	0.7807	0.7659
tiroit_6ondalikPLUS-6tam.csv	0.775	0.7659

Tablo 4.16. (Devamı) Tiroit kanseri alt veri setleri ile elde edilen doğruluk değerleri.

Veri Setleri	Egitim_dogruluk_tahmin	Test_dogruluk_tahmin
tiroit_6ondalikPLUS-10tam.csv	0.7967	0.7872
tiroit_SOMEondalikPLUS-SOME-tam.csv	0.7967	0.7872
tiroitTamami_sayisalTam.csv	**0.8502**	0.7659
tiroitTAMSAYI.csv	0.7807	0.7659
tiroit_ondalik.csv	0.6363	0.5744
tiroitTAMSAYI-HOT.csv	0.8021	0.7234
tiroitTAMSAYI-HOT-3.csv	0.7754	0.7659
tiroitTAMSAYI-HOT-4.csv	0.7754	0.7659
tiroitTAMSAYI-HOT-5.csv	0.7754	0.7659
ilk3.csv	0.7754	0.7659
ilk4.csv	0.7754	0.7659
ilk5.csv	0.7754	0.7659
ilk6.csv	0.7754	0.7659
ilk7.csv	0.7754	0.7659
ilk8.csv	0.7754	0.7659
ilk9.csv	0.7754	0.7659

4.4.3. Sonuçların değerlendirilmesi

Destek vektör makineleri sınıflandırıcısı kullanılarak elde edilen farklı doğruluk değerlerinin çokluğu farklı veri seti alt kümelerinden oluşan farklı veri setlerinin kullanılmasından kaynaklanmaktadır. Eldeki veriler nümerik veri tiplerine dönüştürüldükten sonra tamsayılar, ondalık sayılar ve ikili sayılardan oluşan farklı veri setlerine dönüştürülmüşlerdir. Ayrıca polinominal sayılar yeniden kodlanarak (one hot encodding) binominal veri tipine dönüştürülmüş ve binominal veri tiplerinden oluşan yeni veri seti alt kümeleri oluşturulmuştur. Tablo 4.8'de gösterilen öznitelik seçme testleri kullanılarak elde edilen öznitelik küme ve alt küme veri setleri yanı sıra benzer veri tiplerinden oluşan veri alt kümeleri de oluşturularak modelde veri seti olarak kullanılmıştır. Diğer makine öğrenme algotirmalarında kullanıldıkları gibi destek vektör makineleri modelinde de tüm bu önemli veri alt setleri kullanılmıştır. Belirlenen öznitelikler dışında özellikle ondalık sayılardan oluşan veri tiplerinin sonuç üzerindeki etkilerini görebilmek için oluşturulan sadece ondalık verilerden oluşan *tiroit_ondalik.csv* veri alt kümesi modelde denenmiş ve doğruluk açısından sonucun en az olduğu gösterilmiştir (Tablo 4.16). Sadece tam sayılı yada ikili veri tiplerinden oluşan veri alt kümelerinin modelde ulaştığı sonuçlar doğruluk açısından daha

yüksektir. Tam sayılı yada ikili veri tiplerinden oluşan veri alt setlerine dahil olan ondalık verilerin sonuçlara olumlu anlamda yeterince etki etmedikleri elde edilen doğruluk değerleri açısından görünmektedir. Bu modelde tüm veri tiplerinin ve tüm özniteliklerin dahil olduğu veri seti eğitim verileri açısından % 85 ile doğruluk değerleri içinde en yükseğini verirken aynı veri seti test verileri açısından eğitim verilerine nazaran %77 ile daha düşük doğruluk değerine ulaşmıştır. Eğitim ile test verileri üzerinden ayrı ayrı alınan doğruluk değerleri arasında fark olması aşırı uyumluluk (overfitting) problemini akla getirse de farkın çok fazla olmaması kullanılan model için yeterli ve uygun sonuçlara varıldığına işaret etmektedir. Bu yaklaşımla eğitim ve test verilerinde ortaya çıkan sonuçların bir birine yakın olacak bir oranla aynı büyüklükte olması en iyi hedeftir. Buna göre Tablo 4.16'deki tüm diğer veri setleri eğitim ve test verilerinden elde edilen doğruluk değerlerinin bir birine yakın olmaları açısından daha tutarlıdırlar. Diğerlerine nazaran en az öznitelikle aşırı uyumluluk sorunu şüphesi olmaksızın bir birlerine yakın doğruluk değerlerini veren veri setleri model8.csv ve ilk3.csv olmuştur.

5. KANSER TANISI WEB UYGULAMASI

5.1. Giriş

Bu bölümde, destek vektör makineleri sınıflandırıcısı kullanılarak elde edilen doğruluk sonucu üzerinden tanı koyan bir web kullanıcı arayüzü oluşturulmuştur. Bu ara yüzün hedefi, modele girdi olarak verilen hasta verilerinin bir web formu üzerinden DVM modeline dahil edilmesi ve sınıflandırıcı üzerinden elde edilen sonucun (Kanserli-1 yada Kansersiz-0) bir tanı tahmini olarak ekrana yansıtılmasıdır. Önceki bölümlerde kullanılan sınıflandırma yöntemleri ile elde edilen doğruluk sonuçları ortaya konmuştu. Dördüncü bölümde kullanılan DVM sınıflandırıcı üzerinden en iyi sonucu eğitim verileri için %85 ve test verileri için %75 ile doğruluk değerleri elde edilmişti. Tüm veri seti ve veri alt kümeleri üzerinden elde edilen bu en iyi sonuç bir model olarak *.sav* uzantılı bir dosyada saklanarak web sayfasında işleme alınmıştır.

5.2. Gerekli Araçlar ve Uygulama

Tiroit kanser tanısı koyan Web uygulaması için, özelde makine öğrenmesi genelde ise veri bilimleri alanında web ara yüzleri oluşturmaya yarayan ve Python'a ait olan *streamlit* kütüphanesi kullanılmıştır. Bununla birlikte kullanılan tüm araçlara bir platform sunan Anaconda navigator, kod yazmak için spyder uygulaması ve Python için numpy, Scikit-learn, matplatlib ve panda kütüphanelerinin kurulumları gerçekleştirilmiş ve kullanılmıştır. Tüm bu işlem adımları ve adımlara karşılık gelen ilgili şekil ve tablo numaraları Tablo 5.1'de bir arada verilmiştir.

Tablo 5.1. İşlem adımları ve karşılık gelen ilgili çıktılar.

İşlem Adımları	İşlem Çıktıları
1- DMV sınıflandırıcısı ile Doğruluk(Accuracy) değerlerinin elde edilmesi	Tablo 5.2
2- Anaconda Navigatorun kurulması ve açılması (Yönetici olarak başlatılmalıdır)	Şekil 5.1
3-Anaconda Navigator'da yeni bir ortamın (Environments) oluşturulması	Şekil 5.1
4- Oluşturulan ortam seçili halde iken Python kütüphanelerinin yüklenmesi	Şekil 5.2
5- Oluşturulan ortam seçili iken aynı ortamda bulunan Spyder'ın kurulumu	Şekil 5.1
6- Oluşturulan ortamın terminali üzerinden streamlit kütüphanesinin kurulması	Şekil 5.2

Tablo 5.1. (Devamı) İşlem adımları ve karşılık gelen ilgili çıktılar.

İşlem Adımları	İşlem Çıktıları
7- 1. Adımda elde edilen sonuçları veren eğitilmiş modelin paket haline getirilmesi	Tablo 5.2
8- Spyder üzerinde eğitilmiş modeli işleten fonksiyonların oluşturulması	Tablo 5.3
9- streamlit run "Kod dosyası" ile web uygulamasının başlatılması	Şekil 5.3

Tablo 5.2'de verilen python kodları girdi olarak aldığı Troit kanser hasta verilerini destek vektör makineleri sınıflandırıcısında eğitime tabi tutmakta ve bir model oluşturmaktadır. 1. Satırda girdi olarak yazılan kütüphanelerle başlayan kod satırları bir tahmin modeli oluşturmuş ve bu model 35. satırda *pickle* ile bir *.sav* uzantılı dosyaya dönüştürülerek tanıyı gerçekleştiren fonksiyonlarda kullanılmıştır. Elde edilen ve dosya halinde bir model olarak saklanan tanı modeli Tablo 5.3' te verilen ve *Spyder* derleyicisinde yazılan kodlar içerisine girdi olarak dahil edilmiştir ve *Tiroit Kanser Tahmini.py* dosya adıyla kaydedilmiştir.

Tablo 5.3'teki kod satırları ile, hasta verilerini kullanıcıdan ve kullanacağı sınıflandırıcı modelini ise 4. satırdaki *.sav* uzantılı dosyadan alan bir tiroit kanser tanı fonksiyonu oluşturulmuştur. Tablo 5.3'te tüm kodları verilen *.py* uzantılı dosya ile Tablo 5.2'de kodları verilen *DVM ile WEB TANI UYGULAMASI-Streamlit.ipynb* adlı dosya aynı dizinde tutulmuştur. Aynı dizine *.sav* uzantılı model dosyası da dahil edilmiştir.

Anaconda Navigator'de oluşturulan ortam terminali açılarak *streamlit* ile Tablo 5.3'te kodları verilen *.py* uzantılı pyhon dosyası çalıştırılmıştır. Streamlit üzerinde çalışan Kanser tanısı programı web tarayıcısında otomatik açılarak ekrana Şekil 5.3'te olduğu gibi yansıtılmıştır.

Tablo 5.2. DVM ile modelin oluşturulması.

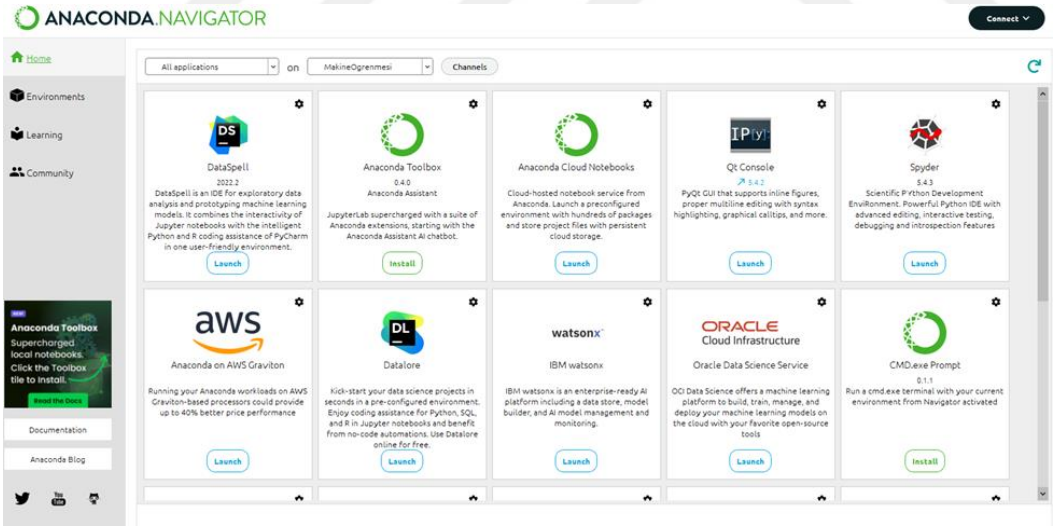
```
import numpy as np   #numpy kütüphanelerin girilmesi
import pandas as pd   #pandas kütüphanesinin girilmesi
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import accuracy_score
#Veri seti elde edilmesi ve Analizi  #Tiroit kanseri veri seti
# pandas DataFrame özelliğine bağlı olarak tiroit kanseri veri setinin yüklenmesi
tiroit_veriseti = pd.read_csv('/content/tiroit6tam.csv')
tiroit_veriseti.head()     #Veri Setinin ilk 5 satırının görüntülenmesi
```

Tablo 5.2. (Devamı) DVM ile modelin oluşturulması.

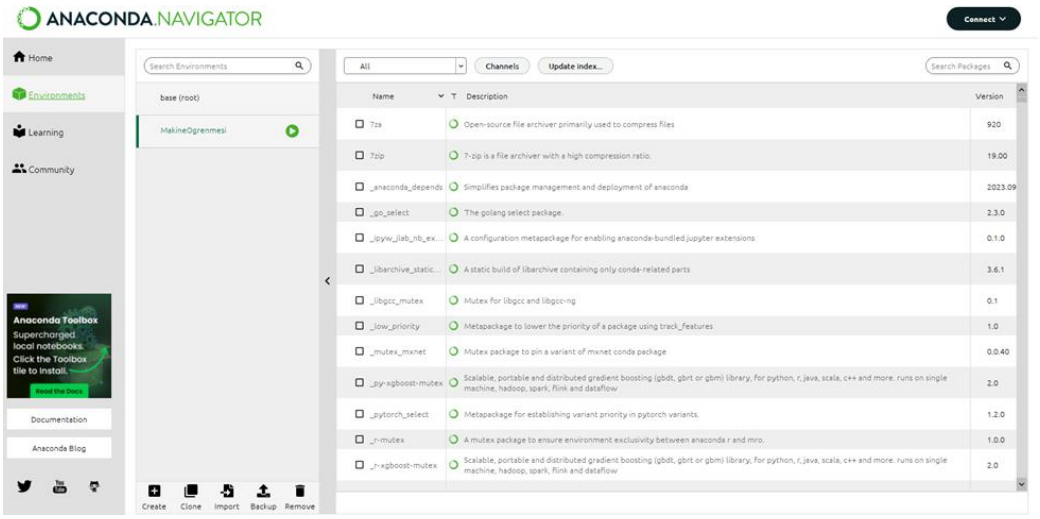
```
tiroit_veriseti.shape      #Veri setindeki  satır (örnek) ve sütun (öznitelik) sayıları
# Verilerin istatistiksel ölçümlerinin görüntülenmesi
tiroit_veriseti.describe()
tiroit_veriseti['postopSonuc'].value_counts() # Sonuç : 0 --> Kanserli  Değil  1 --> Kanserli
tiroit_veriseti.groupby('postopSonuc').mean()
# Girdi verileri ile Çıktı verilerinin ayrıştırılması
X = tiroit_veriseti.drop(columns = 'postopSonuc', axis=1)
Y = tiroit_veriseti['postopSonuc']
print(X)
print(Y)
#Verilerin Eğitim ve Test olarak parçalara ayrılması
X_egitim, X_test, Y_egitim, Y_test = train_test_split(X,Y, test_size = 0.2, stratify=Y, random_state=2)
print(X.shape, X_egitim.shape, X_test.shape)
#Modelin eğitilmesi
siniflandirici = svm.SVC(kernel='linear')
# Destek Vektör Makineleri Sınıflandırıcısının eğitilmesi
siniflandirici.fit(X_egitim, Y_egitim)
#Model Çalıştırılması ve Değerlendirilmesi
#Doğruluk (Accuracy) değeri
# Eğitim verilerinde Doğruluk değeri
X_egitim_tahmini = siniflandirici.predict(X_egitim)
egitim_verileri_Dogruluk = accuracy_score(X_egitim_tahmini, Y_egitim)
print('Egitim verileri için Dogruluk : ', egitim_verileri_Dogruluk)
X_test_tahmini = siniflandirici.predict(X_test)   # Test verilerinde doğruluk değeri
test_verileri_Dogruluk = accuracy_score(X_test_tahmini, Y_test)
print('Test verileri için Dogruluk : ', test_verileri_Dogruluk)
girdi_verileri = (1,61,0,0,0,0)    #Tahmin Sisteminin Oluşturulması
# girdi_verileri'nin numpy dizisine dönüştürülmesi
numpy_dizisi_olarak_girdi_verileri = np.asarray(girdi_verileri)
# Bir örnek için Kanser tahmini yapılması sırasında dizinin tekrar şekillendirilmesi
yeni_bicimli_girdi_verileri = numpy_dizisi_olarak_girdi_verileri.reshape(1,-1)
tahmin = siniflandirici.predict(yeni_bicimli_girdi_verileri)
print(tahmin)
if (tahmin[0] == 0):
  print('Kişi Kanserli Değildir')
else:
  print('Kişi Kanserlidir')
```

Tablo 5.2. (Devamı) DVM ile modelin oluşturulması.

```
#Eğitilmiş Modelin Kayıt Altına Alınması
import pickle
filename = 'egitimli_modelT.sav'
pickle.dump(siniflandirici, open(filename, 'wb'))
# Kaydedilen Modelin değişkene aktarılması
yuklenen_model = pickle.load(open('egitimli_modelT.sav', 'rb'))
girdi_verileri = (1,61,0,0,0,0)
# girdi_verileri'nin numpy dizisine dönüştürülmesi
numpy_dizisi_olarak_girdi_verileri = np.asarray(girdi_verileri)
# Bir örnek için kanser tahmini yapılması sırasında dizinin tekrar şekillendirilmesi
yeni_bicimli_girdi_verileri = numpy_dizisi_olarak_girdi_verileri.reshape(1,-1)
tahmin = yuklenen_model.predict(yeni_bicimli_girdi_verileri)
print(tahmin)
if (tahmin[0] == 0):
  print('Kişi Kanserli Değildir')
else:
  print('Kişi Kanserlidir')
```



Şekil 5.1. Anaconda Navigator kurulumu sonrası çalışma alanı.



Şekil 5.2. Ortam oluşumu ve python kütüphanelerinin yüklenmesi.

Tablo 5.3. Tahmin dosyasının spyder kullanılarak oluşturulması.

Tiroit Kanser Tahmini.py dosyasının oluşturulması

```
import numpy as np
import pickle
import streamlit as st
# Kaydedilmiş eğitimli modelin yüklenmesi
yuklenen_model = pickle.load(open('D:/Kanser Tahmin Otomasyon/egitimli_model.sav', 'rb'))
def tiroit_kanser_tahmini(girdi_verileri):          # Tahmin için Fonksiyon oluşturulması
    # girdi_verileri'nin numpy dizisi olarak değiştirilmeleri
    numpy_dizisi_olarak_girdi_verileri = np.asarray(girdi_verileri)
# Bir örnek için Kanser tahmini yapılması sırasında dizinin tekrar şekillendirilmesi
    yeni_bicimli_girdi_verileri = numpy_dizisi_olarak_girdi_verileri.reshape(1,-1)
    tahmin = yuklenen_model.predict(yeni_bicimli_girdi_verileri)
    print(tahmin)
    if (tahmin[0] == 0):
      return 'Kişi Kanserli Değildir'
    else:
      return 'Kişi Kanserlidir'
def main():
    # Başlığın Verilmesi
    st.title('Tiroit Kanser Tahmini')
# Girdi verilerinin kullanıcıdan elde edilmesi
    cinsiyet = st.text_input('cinsiyet')
    YAS = st.text_input('YAS')
    USG = st.text_input('USG değeri')
```

Tablo 5.3. (Devamı) Tahmin dosyasının spyder kullanılarak oluşturulması.

Tiroit Kanser Tahmini.py dosyasının oluşturulması

```
    USG_LAP = st.text_input('USG_LAP değeri')
    nodul_ozelligi = st.text_input('nodul_ozelligi')
    Sitoloji_sonucu = st.text_input('Sitoloji_sonucu')
    # Tanı tahmini için değişken oluşturma
    tani = ''
    # Tahmin için sonuç butonunun oluşturulması
    if st.button('Kanser Tanı Sonucu'):
        tani = tiroit_kanser_tahmini([cinsiyet,YAS,USG,USG_LAP,nodul_ozelligi,Sitoloji_sonucu])
    st.success(tani)
if __name__ == '__main__':
    main()
```

Şekil 5.3. Kodların streamlit üzerinden çalıştırılması ile ortaya çıkan form.

5.3. Sonuçların Değerlendirilmesi

Önceki bölümlerde tiroit kanseri hasta verileri çeşitli öznitelik seçim algoritmalarına tabi tutularak veri setleri oluşturulmuştur. Çeşitli sınıflandırıcılar ve farklı veri alt kümeleri kullanılarak en iyi doğruluk oranı bulmaya çalışılmıştır. Bu bölümde kullanılan DVM sınıflandırıcısına girdi olarak verilen bir çok veri alt seti içinden verilerin tamamını barındıran veri seti %85 doğruluk oranı ile en iyi sonucu vermiştir. Buna göre her 100 kişi için yapılan kanser tanısında 85 kişiye doğru tanı koyan bir model oluşturulmuştur. Oluşturulan web arayüzü tiroit kanseri hasta verilerini form üzerinden alarak modele girdi olarak vermekte ve hastanın kanserli olup olmadığı tahminini ekrana yansıtmaktadır.

Gerçek hayat senaryosunda eldeki cerrahi operasyon kararı verilen 234 hasta verisi üzerinden 144 hastaya kanserli olduğu tanısının yanlışlıkla konulduğu ve bu sayının 144/234 oranı alındığında %61'e denk geldiği ve 100 hastadan 61'ine yanlış, 39'una doğru tanı konulduğu ortaya çıkmaktadır. Geliştirilen web arayüzü ile kullanılan sınıflandırma modeli üzerinden %85 doğru ve sadece %15 yanlış tanı konulmuştur. Bu sonuçlara göre poliklinikler üzerinden verilmiş olan %61 yanlış tanı oranı azaltılarak %15'e düşürülmüştür. Daha ileri çalışmalar yapılarak bu oranın daha da düşürülüp düşürülemeyeceği araştırılabilir.

Bu model doktorlara, tiroit kanser tanısı koymak için, bir karar destek sistemi olarak yardımcı olabilir. Bu bölümde oluşturulan model sınıflandırıcısı diğer ikili sonuç veren hastalık tanılarında da kullanılabilir.

6. TARTIŞMA VE İLERİ ÇALIŞMALAR

Eldeki tiroit kanseri hasta veri seti farklı veri tiplerinden oluşmakta ve sınıf olarak ikili (binary) veri tipi ihtiva etmektedir. Çeşitli şikayetlerle ön muayeneye alınan hastalar daha sonra çeşitli görüntüleme işlemlerine tabi tutulmuşlar ve kan testine yönlendirilmişlerdir. Ortaya çıkan veriler 25'ten fazla farklı veri çeşidine sahiptir. Bu verilerin bir arada değerlendirilmeleri ya mümkün olmamış yada tanı için sadece bir yada bir kaçı kullanılmıştır.

Eldeki veriler üzerinden uzmanlar tarafından yapılan değerlendirmeler neticesinde Troit kanseri tanısı ile ameliyatlarına karar verilen kadın ve erkeklerden oluşan 234 kişinin ameliyat sonrası 144 'ünün patoloji sonuçları temiz çıkmış ve aslında kanser olmadıkları ve bu nedenle ameliyat olmalarına gerek olmadığı ortaya çıkmıştır. Gereksiz ameliyat edilen 144 kişinin toplam hastaya oranı alındığında % 61 gibi yüksek bir oran çıkmaktadır. Doğru teşhis konulan ve dolayısıyla ameliyatları yerini bulan sadece %39 oranında kalmıştır. Bunun anlamı tanının yüksek oranda yanlış konulduğudur.

Buna bağlı olarak makine öğrenmesi algoritmaları aynı verileri kullanarak daha fazla doğru tanı koyma yetisine sahiptir. Bu tezde kullanılan veri setinden çıkarılan bir çok veri alt kümesi destek vektör makineleri ve lojistik regresyon modellerinde %70'in üzerinde doğruluk paylarına ulaşmışlardır. Bu tezde 5. Bölümde oluşturulan sınıflandırma modeli ile tüm veri seti üzerinden %85 doğruluk oranı elde edilmiştir. Bunun anlamı kanser değerlendirmesi yapılan her 100 kişiden 85'ine doğru tanı konulduğu anlamına gelir ve bu sonuç makine öğrenmesi olmaksızın bilinen yöntemlerle elde edilen sonuçlarından daha iyi netice vermiştir.

Lojistik regresyon modeli çalıştırılarak elde edilen sonuçlar hata matris tablosu üzerinden çeşitli metriclerle yorumlanmıştır. Özellikle tıbbi tanı için elde edilen doğruluk sonuçları yorumlamalr için yeterli olmayabilir. Tiroit kanseri tanısında amaç kansersiz vakaları tespit ederek gerekisz ameliyatların önüne geçmek olduğundan özellikle Kansersiz (0 olma durumu) üzerinde yoğunlaşmak gerekmektedir. Veri setindeki 144 Kansersiz olma durumu lojistik regresyon ile %99,31 Özgüllük oranı ile 143 kişide tespit edilmiştir. Spesifiklik ölçümü ile elde edilen sonuca göre gerçekte

kanser olmayan 144 hasta verileri girdi olarak lojistik regresyon modeline dahil edildiğinde bu hastaların 143'ünün ameliyat edilmemesi gerektiği sonucuna varılmıştır.

Bu tezin amaçları göz önüne alındığında varılan sonuçlardan biri, makine öğrenme algoritmalarından oluşturulan modeller üzerinden daha iyi tanı sonuçları çıkartılabileceği iken bir diğeri, genelde ondalık sayılar olarak verilen tahlil sonuçlarının tanı koymada, kullanılan modeller çerçevesinde, önemli oranlarda iyi sonuçlar veremediği bilgisinin ortaya çıkmasıdır.

Mevcut veri setindeki diğer test verilerinin (özellikle kan testlerinden oluşan ondalık veri tipine sahip özniteliklerin) kanser teşhisi kararlarında ne kadar yol gösterici olabileceği şimdilik, lojistik regresyon ve destek vektör makinesi modelleri üzerinden gösterilememiştir. İlgili uzman doktorların özelikle talep ettikleri sonuç, non-invaziv yöntemlerle karar verebilmek için, kan ve hormon testlerinin ne kadar etkili olduğu ile ilgilidir. Bu durum farklı makine öğrenmesi algoritmaları ile test edilerek ayrıca araştırılabilir. Tüm bunlarla birlikte, son yıllarda görüntüleme tekniklerinin modernleşmesi neticesinde elde edilen görüntü verileri üzerinden de makine öğrenmesi modelleri kullanılabilir. Bu durumda, hastalardan istenen birçok teste gerek kalmaksızın görüntü verileri üzerinden tiroit kanser tanısı konulabilir.

Burada, ayrıca gözden kaçması muhtemel çok önemli bir yaklaşım da ayrıca belirtilmelidir. İstatistik ve matematik elbette birçok alanda çözümler sunmaktadır ancak söz konusu tıbbi tahliller, tanı ve tedavi olduğunda, klinik analizler ile istatistiki analizler arasındaki farklılıklar ve benzerlikler iyi kavranmalı ve klinik analizleri istatistiki olanlara kurban vermemelidir. Tüm bu sonuçlara, eldeki verilere uygulanan istatistiki testler üzerinden kullanılan algoritmalarla varıldığı göz önünde bulundurulmalıdır.

KAYNAKLAR

[1] Fett, M. J. (2000). Technology, Health and Health Care. Department of Health and Aged Care publishing.

[2] Saluvan, M. (2013). Sağlık hizmetlerinin kalitesini iyileştirmede bilgi sistemlerinin rolü. Ankara Sağlık Bilimleri Dergisi, 2(123), 25–39. https://doi.org/10.1501/Asbd_0000000040

[3] Aronow, D.B., Coltin, K. L.(1993). The Joint Commission Journal on Quality Improvement. Using Information Technology to Improve the Quality of Health Care, 19(9), 403–415. https://doi.org/10.1016/S1070-3241(16)30024-4

[4] Derrington, D. (2017). Artificial Intelligence for Health and Health Care. JASON The Mitre Corporation publishing.

[5] Larosa, E., Danks D., (2018). Impacts on Trust of Healthcare AI Roles for Healthcare AI. Proceedings of the 2018 AAAI/ACM Conference on AI, U.S.A, 210-215. https://doi.org/10.1145/3278721.3278771

[6] Türk Toraks Derneği (2019, 20 Eylül). Halk sağlığı. https://www.toraks.org.tr/halk/News.aspx?detail=4013 adresinden 14 Temmuz 2020 tarihinde alınmıştır.

[7] World Health Organization (2019, 13 Ekim). Kanser. https://www.who.int/news-room/fact-sheets/detail/cancer adresinden 14 Temmuz 2020 tarihinde alınmıştır.

[8] World Health Organization. (2006). Constitution Of The World Health Organization, Basic Documents, United Nations, 45, 1-18.

[9] Health Impact Assessment (2023, 20 Ekim). The determinants of health. https://web.archive.org/web/20191214054600/https://www.who.int/hia/evidence/doh/en/ adresinden 5 Aralık 2023 tarihinde alınmıştır.

[10] Lammert, E. & Zeeb, M. (2014). Metabolism of Human Diseases. Springer.

[11] Niederhuber, J. E., Armitage, J. O., Doroshow, J.H., Kastan M. B., & Tepper J. E. (2020). Abeloffs Clinical Oncology, 6th edition, USA, Elsevier Publishing.

[12] Sherr, C. J. (2004). Principles of Tumor Suppression. Howard Hughes Medical Institute, Department of Genetics and Tumor Cell Biology, Cell, 116, 235–246. https://doi.org/10.1016/s0092-8674(03)01075-4

[13] Gurunathan, S., Kang, M. H., Qasim, M. & Kim, J. H. (2018). Nanoparticle-Mediated Combination Therapy: Two-in-One Approach for Cancer. International Journal of Molecular Sciences, 19(10), 32-64. https://doi.org/10.3390/ijms19103264)

[14] World Health Organization (2023, 22 Kasım). Cancer. https://www.who.int/en/news-room/fact-sheets/detail/cancer adresinden 5 Aralık 2023 tarihinde alınmıştır.

[15] National Cancer Institute (2017, 18 Temmuz). Thyroid Cancer. https://www.cancer.gov/types/thyroid/patient/thyroid-treatment-pdq#section/all adresinden 24 Ağustos 2019 tarihinde alınmıştır.

[16] Russell, S. J., & Norvig, P. (2009). Artificial Intelligence: A Modern Approach. Pearson Education.

[17] White Paper: On Artificial Intelligence (2020). A European approach to excellence and trust (PDF). European Commission. Brussels

[18] Luca M., Kleinberg J., & Mullainathan S., (2016). Algorithms Need Managers Too. Harvard Business Review. https://hbr.org/2016/01/algorithms-need-managers-too adresinden 16 Haziran 2020 tarihinde alınmıştır.

[19] Wikipedia Orginization (2023, 21 Şubat). Artificial intelligence in healthcare. https://en.wikipedia.org/wiki/Artificial_intelligence_in_healthcare adresinden 15 Haziran 2023 tarihinde alınmıştır.

[20] Power, B. (2015). Artificial Intelligence Is Almost Ready for Business. Harvard Business Review. https://hbr.org/2015/03/artificial-intelligence-is-almost-ready-for-business adresinden 21 Kasım 2022 tarihinde alınmıştır.

[21] Bahl, M., Barzilay R., Yedidia, A. B., Locascio, N. J., Yu L., & Lehman C. D. (2018). High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision. Radiology, 286 (3), 810–818. https://doi.org/10.1148/radiol.2017170549

[22] Bloch, B. S. (2016). NHS using Google technology to treat patients. BBC News. https://www.bbc.com/news/health-38055509 adresinden 21 Kasım 2023 tarihinde alınmıştır.

[23] Lorenzetti, L. (2016). Here's How IBM Watson Health is Transforming the Health Care Industry. Fortune. https://fortune.com/longform/ibm-watson-health-business-strategy/ adresinden 21 Kasım 2023 tarihinde alınmıştır.

[24] Bloch, B. S. (2016). NHS using Google technology to treat patients. BBC News. https://www.bbc.com/news/health-38055509 adresinden 21 Kasım 2023 tarihinde alınmıştır.

[25] Kent, J. (2018). Providers Embrace Predictive Analytics for Clinical, Financial Benefits. HealthITAnalytics. https://healthitanalytics.com/news/providers-embrace-predictive-analytics-for-clinical-financial-benefits adresinden 22 Kasım 2023 tarihinde alınmıştır.

[26] Lee, K. (2016). Predictive analytics in healthcare helps improve OR utilization. SearchHealthIT. https://searchhealthit.techtarget.com/feature/Predictive-analytics-in-healthcare-helps-improve-OR-utilization adresinden 22 Kasım 2023 tarihinde alınmıştır.

[27] Wang, H., Zu, Q., Chen, J., Yang, Z., & Ahmed, M. A. (2021). Application of Artificial Intelligence in Acute Coronary Syndrome: A Brief Literature Review. Advances in Therapy, 38(10), 5078–5086. https://doi.org/10.1007/s12325-021-01908-2

[28] Infante, T., Cavaliere, C., Punzo, B., Grimaldi, V., Salvatore M., & Napoli C. (2021). Radiogenomics and Artificial Intelligence Approaches Applied to Cardiac Computed Tomography Angiography and Cardiac Magnetic Resonance for Precision Medicine in Coronary Heart Disease: A Systematic Review. Circulation: Cardiovascular Imaging, 14(12), 1133–1146. https://doi.org/ 10.1161/CIRCIMAGING.121.013025

[29] Stewart, J., Lu, J., Goudie, A., Bennamoun, M., Sprivulis, P., Sanfillipo, F., & Dwivedi G. (2021). Applications of machine learning to undifferentiated chest pain in the emergency department: A systematic review. Plos One, 16 (8), e0252612. https://doi.org/10.1371/journal.pone.0252612

[30] Chen, W., Sun, Q., Chen, X., Xie, G., Wu, H., & Xu, C. (2021). Deep Learning Methods for Heart Sounds Classification: A Systematic Review. Entropy, 23(6), 667-679. https://doi.org/10.3390/e23060667

[31] Chan, S., Reddy, V., Myers, B., Thibodeaux, Q., Brownstone, N., & Liao W. (2020). Machine Learning in Dermatology: Current Applications, Opportunities, and Limitations. Dermatology and Therapy, 10 (3), 365–386. https://doi.org/ 10.1007/s13555-020-00372-0

[32] Cao, J. S., Lu, Z. Y., Chen, M. Y., Zhang, B., Juengpanich, S., & Hu J. H. (2021). Artificial intelligence in gastroenterology and hepatology: Status and challenges. World Journal of Gastroenterology, 27(16), 1664–1690. https://doi.org/ 10.3748/wjg.v27.i16.1664

[33] Tran, N. K., Albahra, S., May, L., Waldman, S., Crabtree, S., Bainbridge, S., & Rashidi, H. (2021). Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing. Clinical Chemistry, 68(1), 125–133. https://doi.org/ 10.1093/clinchem/hvab239

[34] MIT Technology Review (2021). AI could make health care fairer by helping us believe what patients say. https://www.technologyreview.com/2021/01/22/1016577/ai-fairer-healthcare-patient-outcomes/ adresinden 6 Haziran 2023 tarihinde alınmıştır.

[35] Sarakhsi, M. K., Haghighi, S. S., Ghomi, S. F., & Marchiori, E. (2022). Deep learning for Alzheimer's disease diagnosis: A survey. Artificial Intelligence in Medicine, 130, 933–3657. https://doi.org/10.1016/j.artmed.2022.102332

[36] Bhinder, B., Gilvary, C., Madhukar, N.S., & Elemento, O. (April 2021). Artificial Intelligence in Cancer Research and Precision Medicine. Cancer Discovery, 11 (4), 900–915. https://doi.org/10.1158/2159-8290.CD-21-0090

[37] Ravindran, S. (2019). How artificial intelligence is helping to prevent blindness. Nature. https://doi.org/10.1038/d41586-019-01111-y

[38] Cao, J. S., Lu, Z. Y., Chen, M. Y., Zhang, B., Juengpanich, S., & Hu, J. H. (2021). Artificial intelligence in gastroenterology and hepatology: Status and challenges. World Journal of Gastroenterology. 27(16), 1664–1690. https://doi.org/10.3748/wjg.v27.i16.1664

[39] Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., & Fokkema, M. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), 154–170. https://doi.org/ 10.1002/wps.20882

[40] Pisarchik, A. N., Maksimenko, V. A., & Hramov, A. E. (2019). From Novel Technology to Novel Applications: Comment on An Integrated Brain-Machine Interface Platform With Thousands of Channels by Elon Musk and Neuralink. Journal of Medical Internet Research, 21(10). https://doi.org/10.2196/16356

[41] Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., & Ma, S., (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2 (4), 230–243. https://doi.org/10.1136/svn-2017-000101

[42] Shayea, A. (2011). Artificial neural networks in medical diagnosis. International Journal of Computer Science. https://www.researchgate.net/publication/285912467_Artificial_Neural_Networks_in_Medical_Diagnosis

[43] Prasath, V., Lakshmi, N., Nathiya, M., Bharathan, N., & Neetha, N. P. (2013). A survey on the applications of fuzzy logic in medical diagnosis. International Journal of Science and Engineering, 4(4), 1199–1203. https://api.semanticscholar.org/CorpusID:17561242

[44] Janardhanan, P., Heena, L., & Sabika, F. (2015). Effectiveness of support vector machines in medical data mining. Journal of Communication and Software and Systems, 11(1), 25–30. https://doi.org/10.24138/jcomss.v11i1.114

[45] Mourya, A.K., Tyagi, P., & Bhatnagar, A. (2016). Genetic algorithm and their applicability in medical diagnostic: a survey. International Journal of Science and Engineering, 7(12), 1143–1145. https://doi.org/10.14299/ijser.2016.12.007

[46] Wikipedia Organization (2023, 24 Kasım). Machine learning and relation to artificial intelligence. https://en.wikipedia.org/wiki/Machine_learning Relation_to_artificial_intelligence adresinden 5 Aralık 2023 tarihinde alınmıştır.

[47] Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowl Inf Syst, 14, 1–37. https://doi.org/10.1007/s10115-007-0114-2

[48] McCallum, A. Graphical Models, Lecture2: Bayesian Network Represention(PDF). https://people.cs.umass.edu/~mccallum/courses/gm2011/02-bn-rep.pdf adresinden 1 Temmuz 2020 tarihinde alınmıştır.

[49] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artifitial Intelligence and Medication, 23(1), 89–109. http://refhub.elsevier.com/S0010-4825(22)00250-5/sref1 https://doi.org/10.1016/S0933-3657(01)00077-X

[50] Li, M., & Zhou, Z. H. (2007). Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Trans. Syst. Man Cybern. Syst. Hum., 37(6), 1088–1098. http://refhub.elsevier.com/S0010-4825(22)00250-5/sref80 https://doi.org/10.1109/TSMCA.2007.904745

[51] Wei, L., Yang, Y., Nishikawa, R.M., & Jiang, Y. (2005). A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. IEEE Trans. Med. Imag., 24(3), 371–380. https://doi.org/ 10.1109/tmi.2004.842457

[52] Ing, E., Su, W., Schonlau, M., & Torun, N. (2019). Support Vector Machines and logistic regression to predict temporal artery biopsy outcomes Canadian Journal of Ophthalmology, 54(1), 116-118, https://doi.org/10.1016/j.jcjo.2018.05.006

[53] Ambrish, G., Ganesh B., Ganesh A., Srinivas C., & Kiran, D. (2022). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, India, 3, 127–130. https://doi.org/ https://doi.org/10.1016/j.gltp.2022.04.008

[54] Khaleel, F. A., & Al-Bakry, A. M., (2023). Diagnosis of diabetes using machine learning algorithms. Materials Today: Proceedings, 80(3), 3200-3203. https://doi.org/ https://doi.org/10.1016/j.matpr.2021.07.196

[55] Wozniak, J. M., Jain, R., Balaprakash, P., Ozik, J., Collier, N.T., Bauer, J., Xia, F., Brettin, T., Stevens, R., Mohd, J.Y.,. Cardona, C.G., Essen, B.V., & Baughman, M. (2018). Candle/supervisor: a workflow framework for machine learning applied to cancer research, BMC Bioinf., 19(18), 491. http://refhub.elsevier.com/S0010-4825(22)00250-5/sref91 https://doi.org/10.1186/s12859-018-2508-4

[56] Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013). Using deep learning to enhance cancer diagnosis and classification. Proceedings of the International Conference on Machine Learning, USA, 28, http://refhub.elsevier.com/S0010-4825(22)00250-5/sref92

[57] Podolsky, M. D., Barchuk, A. A., Kuznetcov, V. I., Gusarova, N. F., Gaidukov, V.S., & Tarakanov, S. A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels, Asian Pac. Journal of Cancer Prev., 17 (2), 835–838. https://doi.org/10.7314/APJCP.2016.17.2.835

[58] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl., 36(2), 3240–3247. https://www.sciencedirect.com/science/article/abs/pii/S0957417408000912 https://doi.org/10.1016/j.eswa.2008.01.009

[59] Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694- 701. https://doi.org/10.1016/j.dsp.2006.10.008

[60] Chen, Y. C., Ke, W. C., & Chiu, H. W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med, 48, 1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006

[61] Chang, S. W., Kareem, S. A., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinforma, 14, 170-182. https://doi.org/10.1186/1471-2105-14-170

[62] Rosado, P., Lequerica, F. P., Villallaín, L., Peña, I., Sanchez, L. F., & de Vicente, J. C. (2013). Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines. Expert Syst Appl., 40(6), 4770. https://doi.org/10.1016/j.eswa.2013.02.032

[63] Sunday, O., Olatunji, S. A., Ebtisam, A., Zainab, A., Yasmeen, A., Rahaf, A., Mona, A., Mohammed, I. B. A., Farooqui, M., & Alhiyafi, J. (2021). Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset. Computers in Biology and Medicine, 28, 131-142, https://doi.org/10.1016/j.compbiomed.2021.104267

[64] Zachariah, R., Samarasena, J., Luba, D., Duh, E., Dao, T., Requa, J., Ninh, A., & Karnes, W. (2020). Prediction of polyp pathology using convolutional neural networks achieves "resect and discard" thresholds. Am. J. Gastroenterol, 115(1), 138-144. https://doi.org/10.14309/ajg.0000000000000429

[65] Murugan, A., Anu, S., Nair, H., Angelin, A., Peace, P., & Kumar, S. (2021). Diagnosis of skin cancer using machine learning techniques. Microprocessors and Microsystems, 81(1), https://doi.org/10.1016/j.micpro.2020.103727

[66] Minnoora, M., & Baths, V. (2023). Diagnosis of Breast Cancer Using Random Forest. International Conference on Machine Learning and Data Engineering Procedia Computer Science, 218, 429–437. https://doi.org/ 10.1016/j.procs.2023.01.025

[67] Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L., & Liotta, A. (2022). Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. Computer Methods and Programs in Biomedicine, 223(4), 106951. https://doi.org/10.1016/j.cmpb.2022.106951

[68] Maksood, F.Z. (2016). Analysis of Data Mining Techniques and its Applications. International Journal of Computer Applications, 140(3), 6–14. https://doi.org/10.5120/ijca2016909249

[69] Ramageri, M. (2010). Data Mining Techniques And Applications. Indian Journal of Computer Science and Engineering, 1(4), 301–305. https://api.semanticscholar.org/CorpusID:1705805

[70] Petre, R., (2013). Data Mining Solutions for the Business Environment. Database Systems Journal, 4(4), 21–29. https://api.semanticscholar.org/CorpusID:12617008

[71] Tubitak (2019, 1 Ocak). Sağlık projeleri. http://www.tubitak.gov.tr/tr/destekler/sanayi/ulusal-destek-programlari/1511/icerik-bit adresinden 21 Aralık 2020 tarihinde alınmıştır.

[72] Crockett B. D., & Eliason, B. (2017). What is Data Mining in Healthcare ?. HealthCatalyst. https://www.healthcatalyst.com/wp-content/uploads/2014/06/What-is-data-mining-in-healthcare.pdf

[73] Jackson, J. (2002). Data Mining ; A Conceptual Overview. Communications of the Association for Information Systems, 8, 267-296. https://doi.org/10.17705/1CAIS.00819

[74] World Health Organization (2016, 28 Ocak). Cancer. http://www.who.int/en/ adresinden 20 Ekim 2019 tarihinde alınmıştır.

[75] Chang C. & Chen, C. (2009). Expert Systems with Applications Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Syst. Appl., 36(2), 4035–4041. https://doi.org/10.1016/j.eswa.2008.03.007

[76] Khan, M. U., Choi, J. P., Shin, H., & Kim, M. (2008). Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare. 30th Annual International IEEE EMBS Conference, Canada, 5148–5151.

[77] Rokach L., & Maimon, O. (2010) Clustering Methods, Data Mining Knowledge Discovery Handbook.

[78] Sharma, N., Bajpai, A., & Litoriya, R. (2012). Comparison the various clustering algorithms of weka tools. International Journal of Emerging Technology and Advanced Engineering, 2(5), 73–80. https://api.semanticscholar.org/CorpusID:8057035

[79] Kumar, S. (2014). K-Mean Evaluation in Weka Tool and Modifying It using Standard Score Method. International Journal on Recent and Innovation Trends in Computing and Communication, 2(9), 2704–2706. https://ijritcc.org/index.php/ijritcc/article/download/3280/3280

[80] Wikipedia Organization (2021, 26 Ağustos). Regresyon analizi. https://tr.wikipedia.org/wiki/Regresyon_analizi adresinden 22 Aralık 2022 tarihinde alınmıştır.

[81] Munn, J. S., Lanting, B. A., MacDonald, S. J., Somerville L., Marsh, J. D., & Bryant D., (2022). Logistic Regression and Machine Learning Models Cannot Discriminate Between Satisfied and Dissatisfied Total Knee Arthroplasty Patients. The Journal of Arthroplasty, 37, 267-273. https://doi.org/10.1016/j.arth.2021.10.017

[82] Rossi, R., Socci, V., Talevi, D., Mensi, S., Niolu, C., Pacitti, F., Di Marco, A., Rossi, A., Siracusano, A., & DiLorenzo, G. (2020). COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. Front. Psychiatry, 11, 790. https://doi.org/10.3389/fpsyt.2020.00790

[83] Alfawaz, H., Yakout, S. M., Wani, K., Aljumah, G. A., Ansari, M. G. A., Khattak, M. N. K., Hussain, S. D., & Al-Daghri, N. M. (2021). Dietary intake and mental health among Saudi adults during COVID-19 lockdown. Int. J. Environ. Res. Public Health, 18 (4), 1653 https://doi.org/10.3390/ijerph18041653

[84] Ting, D. S. J., Krause, S., Said, D. G., & Dua, H. S. (2021). Psychosocial impact of COVID-19 pandemic lockdown on people living with eye diseases in the UK. Eye, 35(7), 2064–2066. https://doi.org/10.1038/s41433-020-01130-4

[85] Jacob, L., Smith, L., Armstrong, N. C., Yakkundi, A., Barnett, Y., & Tully, M. A. (2021). Alcohol use and mental health during COVID-19 lockdown: A cross-sectional study in a sample of UK adults. Drug and alcohol dependence, 1, 219. https://doi.org/10.1016/j.drugalcdep.2020.108488

[86] Fu, W., Yan, S., Zong, Q., Anderson, L. D., Song, X., Lv, Z., & Lv, C. (2021). Mental health of college students during the COVID-19 epidemic in China. J. Affect. Disord., 280, 7–10. https://doi.org/10.1016/j.jad.2020.11.032

[87] Jiang, W., Liu, X., Zhang, J., & Feng, Z. (2020). Mental health status of Chinese residents during the COVID-19 epidemic. BMC Psychiatry, 20 (1), 1–14. https://doi.org/10.1186/s12888-020-02966-6

[88] Liu, Z., Liu, R., Zhang, Y., Zhang, R., Liang, L., Wang, Y., Wei, Y., Zhu, R., & Wang, F. (2021). Latent class analysis of depression and anxiety among medical students during COVID-19 epidemic. BMC Psychiatry, 21(1), 498. https://doi.org/10.1186/s12888- 021-03459-w

[89] Liu, Y., Chen, H., Zhang, N., Wang, X., Fan, Q., Zhang, Y., Huang, L., Hu, B. O., & Li, M. (2021). Anxiety and depression symptoms of medical staff under COVID-19 epidemic in China. J. Affect. Disord., 278, 144–148. https://doi.org/10.1016/j.jad.2020.09.004

[90] Brownlee, J. (2020). Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, Machine Learning Mastery Publishing.

ÖZGEÇMİŞ

Ad-Soyad : Mehmet Emin ASAN

ÖĞRENİM DURUMU:

- **Lisans :** 2005, SUNY College at Oldwestbury, New York, A.B.D., Mathematics, Computer & Information Science, Computer Science.
- **Yükseklisans :** 2008, Politechnic Institute of NYU, New York, A.B.D. Management of Technology, M.S., Telecommunication and Information Management.

MESLEKİ DENEYİM :

- 2002-2010 yılları arasında Exxon Mobil ,New York, A.B.D., operasyonel şube müdürü olarak çalıştı.
- 2010-2011 yılları arasında Sakarya Büyük Şehir Belediyesinde Sözleşmeli Mühendis olarak çalıştı.
- 2014 yılı itibariyle Sakarya Uygulamalı Bilimler Üniversitesi, Hendek Meslek Yüksek Okulu, Bilgisayar Teknolojileri bölümünde Öğretim Görevlisi olarak dersler vermektedir.

TEZDEN TÜRETİLEN ESERLER:

- Asan, M.E., Taşkın, H., Alemdar, M. ve Çapoğlu, R. 2023. Use of logistic regression in the diagnosis of thyroid cancer, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,* Doi: 10.17341/gazimmfd.1253193.