FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ VERİLERDE SINIFLANDIRMA VE SAĞLIK VERİLERİ ÜZERİNE UYGULAMASI

Adil Hani Abdulkareem ABDULKAREEM

Danışman
Dr. Öğr. Uyesi Mustafa Cem KASAPBAŞI

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
İSTANBUL – 2021

KABUL VE ONAY SAYFASI

Adil Hani Abdulkareem ABDULKAREEM tarafından hazırlanan **"ÇOK DEĞİŞKENLİ VERİLERDE SINIFLANDIRMA ve SAĞLIK VERİLERİ ÜZERİNE UYGULAMASI"** adlı tez çalışması 01/03/2021 tarihinde aşağıdaki jüri üyeleri önünde başarı ile savunularak, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği Anabilim Dalı**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Danışman	**Dr. Öğr. Uyesi Mustafa Cem KASAPBAŞI** İstanbul Ticaret Üniversitesi	.....
Jüri Üyesi	**Prof. Dr. Serhat ÖZEKES** Üsküdar Üniversitesi	.....
Jüri Üyesi	**Dr. Öğr. Üyesi Arzu KAKIŞIM** İstanbul Ticaret Üniversitesi	.....

Onay Tarihi: 15.03.2021

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsünün 15.03.2021 tarih ve 2021/308 numaralı Yönetim Kurulu Kararının 1. maddesi gereğince, ders yüklerini ve tez yükümlülüğünü yerine getirdiği belirlenen "Adil Hani Abdulkareem ABDULKAREEM" (TC:99943513968) adlı öğrencinin mezun olmasına oy birliği ile karar verilmiştir.

Prof. Dr. Necip ŞİMŞEK

Enstitü Müdürü

AKADEMİK VE ETİK KURALLARA UYGUNLUK BEYANI

İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı beyan ederim.

15/03/2021

Adil Hani Abdulkareem
ABDULKAREEM

İÇİNDEKİLER

ÖZET

Yüksek Lisans Tezi

ÇOK DEĞİŞKENLİ VERİLERDE SINIFLANDIRMA ve SAĞLIK VERİLERİ ÜZERİNE UYGULAMASI

Adil Hani Abdulkareem ABDULKAREEM

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Uyesi Mustafa Cem KASAPBAŞI
2021, 68 sayfa

Meme kanseri, dünyadaki en tehlikeli ve ikinci en yaygın kanser türlerinden biridir. Gelişmiş cihazlarla ve tıbbi tedavilerle meme kanseri ile mücadele daha kolay hale geldi. Meme kanseri tedavisinde en iyi sonucu elde etmek ve erken tanı için periyodik kontroller yapılmalıdır. Makine öğrenme teknikleri, tedavinin başarısını tahmin etmek veya teşhis etmek için kullanılır. Bu çalışmada meme kanserinin erken tespiti için K-En Yakın Komşu (k-NN), Destek Vektör Makinesi (SVM), Naïve Bayes (NB), Lojistik Regresyon (LR) ve Rastgele Ormanlar (RF) sınıflandırıcısı makine öğrenmesi algoritmaları kullanılmıştır. Kullanılan veri seti, yaş, glikoz, BMI, resistin, insülin, adiponektin, HOMA, MCP1 ve leptin özelliklerinden oluşan UCI kütüphanesinden alınan Coimbra meme kanser veri setidir. Yaş, Resistin, Glikoz ve BMI kullanan K-En Yakın Komşu modeli en yüksek sonuçları vermektedir. Burada özgüllüğün % 90'ı hassasiyetin % 84'ü ve % 87.5'i doğruluk elde edilir. SVM algoritması, çalışmamızda % 83 doğruluk oranıyla ikinci en yüksek doğruluğu elde edildi. Ayrıca, RF algoritması % 79 doğruluk oranına, NB algoritması % 79 doğruluk oranına ve LR algoritmasının % 75 doğruluk oranına elde edildi. Bu bulgular resistin, glikoz, yaş ve BMI'yı birleştiren modellerin meme kanseri tespiti için güçlü bir araç olabileceğine dair umut verici kanıtlar sunmaktadır.

Anahtar Kelimeler: Bio belirteci, Coimbra veriset, KNN, Meme Kanseri, Veri madenciliği.

ABSTRACT

M.Sc. Thesis

CLASSIFICATION OF MULTIVARIATE DATA AND APPLICATION ON HEALTH DATA

Adil Hani Abdulkareem ABDULKAREEM

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Asst. Prof. Dr. Mustafa Cem KASAPBAŞI
2021, 68 pages

Breast cancer is one of the most dangerous and second most common types of cancer in the world. Breast cancer-fighting with developed devices and medical therapies has become easier. To obtain the best result in breast cancer treatment, periodic checks should be carried out to follow the early diagnosis. Data Mining techniques are used to predict the success of treatment or diagnosis. In this study, the K-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forests (RF) classifier algorithms of machine learning were used for early detection of breast cancer. From the UC Irvine Machine Learning Repository (UCI) library Coimbra Breast Cancer data set which consists of age, glucose, body mass index (BMI), resistin, insulin, adiponectin, homeostatic model assessment (HOMA), monocyte chemoattractant protein-1 (MCP1), and leptin attributes were used. K-NN model using Age, Resistin, Glucose, and BMI give the highest results, where 90% of specificity 84% percent of sensitivity, and 87.5% accuracy is achieved. The SVM algorithm achieved the second-highest accuracy in our study with an 83% accuracy rate. Also, the RF algorithm has a 79% accuracy rate, the NB algorithm a 79% accuracy rate, and the LR algorithm has a 75% accuracy rate. These findings provide promising evidence that models combining resistin, glucose, age, and BMI may be a powerful tool for breast cancer detection.

Keywords: Breast cancer, Cancer biomarker, Coimbra dataset, Data mining, KNN.

TEŞEKKÜR

Bu araştırma için beni yönlendiren, karşılaştığım zorlukları bilgi ve tecrübesi ile aşmamda yardımcı olan değerli Danışman Hocam Dr. Öğr. Uyesi Mustafa Cem KASAPBAŞI'a teşekkürlerimi sunarım.

Araştırmanın yürütülmesinde maddi ve manevi yardımlarını gördüğüm anneme SUBHYA ABDULRAZZAQ ve babama HANI ABDULKAREEM çok teşekkür ederim.

Tezimin her aşamasında beni yalnız bırakmayan abime SAAD, ablama Entisar, kardeşime ALİ ve ailelerine, son olarak sevgili arkadaşlarıma sonsuz sevgi ve saygılarımı sunarım.

Türkiye'de beni destekleyen ve ailemin yokluğunu hissettirmeyen canım ablam Ayşegül KARGIN'a ve ailesine sonsuz teşekkür ederim.

Sonunda kalbimde yaşayan, resimleri aklımdan çıkmayan ve bu hayatta beni yola yönlendiren rahmetli ağabeyime AQEEL sonsuz teşekkür ederim

Adil Hani Abdulkareem ABDULKAREEM
İSTANBUL, 2021

ŞEKİLLER

Sayfa

ÇİZELGELER

Sayfa

SİMGELER VE KISALTMALAR

ACS	Amerikan Kanser Derneği
AI	Yapay Zekâ
AUC	Eğrinin Altındaki Alan
BMI	Vücut kitle indeksi
DM	Veri Madenciliği
DNN	Derin Sinir Ağı
ELM	Aşırı Öğrenme Makinesi
FN	Yanlış Negatif
FNN	Bulanık Sinir Ağı
FP	Yanlış Pozitif
GLOBOCAN	Global Cancer Gözlemevi
HOMA	Homeostatik model değerlendirmesi
IARC	Uluslararası Kanser Araştırma Ajansı
KDD	Veritabanlarındaki Bilgi Keşfi
KNN	K-en yakın Komşu Algoritması
LR	Lojistik Regresyon
MCP1	Monosit Kemoatraktan Protein-1
ML	Makine öğrenme
MLP	Çok Katmanlı Perceptron
MRI	Manyetik Rezonans Görüntüleme
NB	Naive Bayes
OOB	Out Of Ba
PSO	Parçacık Sürüsü Optimizasyonu
RBF	Radyal Baz Fonksiyonu
RF	Random Forest algoritması
ROC	Alıcı Çalışma Özellikleri
SVM	Destek Vektör Makinesi
TN	Gerçek Negatif
TP	Gerçek Pozitif
UCI	Irvine'deki California Üniversitesi
WEKA	Bilgi Analizi için Waikato Ortamı
WHO	Dünya Sağlık Örgütüne
YSA	Yapay Sinir Ağı

1. GİRİŞ

Kanser, bilinç hücrelerinin kaybıdır ve kontrolsüz bölünmede başlar. Daha sonra hücreler tümörler (kitleler) oluşturmak için birikir, tümörler normal dokuları yok eder, kan veya lenf dolaşımı yoluyla vücudun diğer bölgelerine gider. Meme kanseri, meme dokularının hücresindeki anormal dönüşümüdür. Bu prosedür, tümör olarak adlandırılan yeni doku kütlesi oluşturabilir, tümörler kötü huylu ve iyi huylu olarak iki türe ayrılırlar. Bu iki tip kanser sınıflandırması, tümörün metastaz veya istila yoluyla yayılabilmesine bağlıdır. Kanserli olmayan tümör lokal olarak büyür ve metastaz veya invazyonla yayılamaz fakat malign tümörleri yayılabilir. Meme kanserinin olmasının nedeni, memede kanserli tümörlerin gelişmesidir (Imaginis, 2019) (Soliman & AboElHamd, 2014). Küresel Kanser Olayları (GLOBOCAN) 2018 kullanan Uluslararası Kanser Araştırma Ajansı (IARC) bilgilerine dayanarak, ölüm oranları ve yayılan kadın meme kanseri, dünyada en çok teşhis edilen ikinci kanser türüdür ve kanserin önde gelen etkisi kadınlar arasında ölümdür (Ferlay vd., 2019). Sadece 2018'de yaklaşık 2.1 milyon yeni tanı konacağı tahmin edilmektedir ek olarak dünya çapındaki kanser vakalarında toplam vakaların ölümünün % 11.6 Meme kanserinden kaynaklanmaktadır (Bray vd., 2018).

Amerika'daki Amerikan Kanser Derneği (ACS) araştırmasına göre, sadece 2019'da invaziv meme kanseri teşhisi konan 268.600 kadına yapılan bir araştırma, bu hastalık tüm meme kanseri vakalarının % 1'inden daha az erkeklerde görülen erkek meme kanseridir. 2019'da Amerika'da meme kanseri nedeniyle 41.760 kadının ve 500 erkeğin öleceği tahmin edilmektedir (Society, 2020).

Erken teşhis ve iyileştirilmiş tedaviler sayesinde meme kanserinin ölüm oranları azalmaktadır (Society, 2020). Meme kanserini erken evrede bulmak ölümü önlemek için çok önemlidir. Sağlık sistemleri ve cihazları, hastalıkların erken evrelerinde teşhis edilmesine yardımcı olabilecek veri madenciliği algoritmalarından yararlanabilir. Tıp alanında yeni teknolojilerin gelmesiyle makine öğrenim algoritmalarının (ML) tıbbi alanlarda avantajı giderek artmaktadır. Böylece büyük miktarda kanser verisi elde edildi ve bu veriler artık

tıbbi araştırma topluluğu için de mevcut. Bu aynı zamanda tıp alanındaki araştırmacılar için karmaşık veri setlerinden doğru veri setini tanımlamak ve tespit etmek için bu veri setlerine dayalı kanseri tespit etme yöntemlerini bulmakta bir zorluk oluşturmaktadır. Ancak, hastalardan alınan örneklerin incelenmesi, doğru tanıyı bulmak için ana sebeptir. Bu tezde Naif Bayes (NB), Random Forest (RF), K-En Yakın Komşular (KNN), Lojistik Regresyon (LR), Destek Vektör Makinesi (SVM) gibi makine öğrenim tekniklerinden kullanılmıştır. Bu yöntemler, kanser araştırmalarında, doğru karar vermeyi ve etkililiği amaçlayan bir öngörücü model geliştirmek için yaygın olarak kullanılmaktadır. Makine öğrenim algoritmalarının büyük hedefi tahmin ve sınıflandırma yürütmek için kullanılabilecek modelleri yapmaktır. Test verilerini doğru bir şekilde sınıflandırmak için iyi sınıflandırma modelleri oluşturmak lazımdır, bu nedenle veri seti iki gruba; eğitim ve test gruplarına ayrılacaktır. Model eğitimi için eğitim seti kullanılır ve daha sonra veri setini tahmin etmek için etiketlenmemiş test kullanılır. Bu araştırmanın temel amacı, meme kanserinin erken evrelerinde tespit edilmesine yardımcı olabilecek doğru bir veri madenciliği algoritması oluşturmaktır. Meme kanseri erken evrelerinde teşhis edilirken ölüm oranları azaltılabilirdir. Bu araştırmanın tedavi kalitesini artırması beklenmektedir. En iyi algoritmaları belirlemek ve ölçmek için uygulanan yöntemlerin performansı ile sınıflandırma doğruluğunu karşılaştırmak lazımdır.

Meme Kanseri Coimbra Veri Kümesini kullandık. Bu Veri Seti, eğitim ve test deneyleri için (UCI) Makine Öğrenim Deposundan alınmıştır (Anon., 2018).

Bölüm 1'de meme kanseri hakkında genel bilgi veriyoruz ve bu hastalığın teşhis edilmesine ilişkin diğer araştırmalardan bahsediyoruz, ayrıca veri madenciliği kavramlarını ve sağlık hizmetlerinde neden önem taşıdığını açıklıyoruz.

1.1. Veri Madenciliği Genel Bilgiler

Makine öğrenimi, bir bilgisayarın verilerden nasıl öğrenebileceğini araştıran yapay zekanın bir uygulamasıdır ve buna ek olarak İstatistik ve Bilgisayar Biliminin kesişimi olarak tanımlanabilir. Makine Öğrenimi (ML) algoritmalarının

farklı türlerini kullanarak karmaşık desenler ve yaklaşık sonuçlar ortaya çıkmasına yardımcı olur (Han vd., 2011) (Mueller & Massaron, 2016).

1.1.1. Tanım

Bazı performans ölçütleri ve görevleri ile ilgili bir eğitim deneyiminden öğrenmek için bir bilgisayar programı performans ölçütleri ile ölçülen görevlerdeki performansın deneyimi ile gelişir (Mitchell, 1997).
“Makineler düşünebilir mi?” 1950'de Alan Turing bilgisayar bilimcisi bu ünlü ifadeyi söyledi, sonra öğrenebilecek bir makine fikrini takip etmeye çalıştı (Turing, 1950). Ayrıca taklit oyunu adı verilen ve genellikle 'Turing Test' olarak bilinen bu oyunun insan sorgulayıcı ile oynandığını ve oyunda iki oyuncunun sorgulayıcı amacının bilgisayar tarafından hangisinin insan olduğuna karar vermektir. Sorgulayıcı tarafından sorulan soruya cevap verir. (Turing, 1950). Bu deneyim, makinenin düşünüp düşünemeyeceğini bulmak için nesnel bir yöntemdir. Bu araştırmadan sonra Alan Turing’in çalışmasının yapay zekanın (AI) başlangıcına etkisi olduğu söylenebilir.

Yapay Zeka (AI) bilgisayar biliminin bir koludur, makinelerin deneyimlerden öğrenmesini mümkün kılar ve insan gibi davranacak eğilimdedir. Yapay zekâ; son yıllarda desen tanıma, makine öğrenimi ve sinir ağlarından yararlanır. Satranç oynayan bilgisayarlar günlük hayatta kendi kendine giden otomobiller ve sesten metne çevirme özellikleri Yapay zekâ sistemlerine örnektir (Ertel, 2017).

Veri Madenciliği (DM) Makine Öğreniminin bir alt alanıdır. Veri Madenciliği (DM) Makine Öğreniminin bir alt alanıdır ayrıca verilerden desenleri çıkarmak için belirli algoritmaların uygulanmasıdır. Keşif algoritmalarından oluşan süreç ve kabul edilebilir hesaplama verimliliği sınırlamaları altında veri analizi uygulayarak, veriler üzerinde belirli bir model listesi üretir (Fayyad vd., 1996). Bu verilerden ve büyük veritabanlarından yararlı, ilgili bilgilerin çıkarılması ve analiz edilmesi ile ilgilidir.

Yapay zekâ ve istatistik kullanan depolar, yararlı bilgiler elde etmek için algoritmalar, teknikler, modeller ve yöntemler içerir (Berson & J Smith, 1997).

Makine öğreniminde, denetimli ve denetimsiz öğrenmenin iki türü vardır. Denetimli öğrenme hem çıktı hem de girdi verilerine sahip olduğumuzda kullanılır ve gözetimsiz öğrenme yalnızca girdi verisine sahip olduğumuz, çıktı verisine sahip olmadığımızda kullanılır. Dolayısıyla, denetlenen algoritmalar modeli oluşturmak için eğitim verileri olarak girdileri ve beklenen çıktıları kullanır ama denetimsiz algoritmalar, veri kümelerinin iç yapısını kendi başlarına bulmak zorundadır (Han vd., 2011) (Witten vd., 2016).

Denetimli öğrenme çoğunlukla regresyon modellerinde ve sınıflandırma modellerinde kullanılır. Bir sınıflandırıcının ana amacı, nicelikleri tahmin etmek ve ayrı gruplara gözlemler koymaktır. Örneğin, ikili bir sınıflandırıcı sadece iki olası kategoriye sahiptir. Naif Bayes, Lojistik Regresyon (LR), K-En Yakın-Komşu, Random Forest (RF) ve Destek Vektör Makineleri ve Karar Ağaçları en çok kullanılan sınıflandırma algoritmalarıdır. Regresyon analizinin amacı, ilişkinin iki veya daha fazla değişkenini dikkate aldığını tahmin etmektir. Polinom Regresyonu, Doğrusal Regresyon, Lojistik Regresyon ve Kuantil Regresyon yaygın olarak kullanılan Regresyon modellerinden bazılarıdır (Han vd., 2011) (Witten vd., 2016).

Veri madenciliği Veritabanlarında Bilgi Keşfi olarak da adlandırılır. Veri seçimi; veri hazırlama, veri temizleme, gözlemlenen sonuçlardan doğru çözümlerin yorumlanması ve veri kümeleri hakkında önceden bilgi verilmesi işlemlerinden oluşan bir yöntemdir (Sahu vd., 2011).

Veritabanlarındaki Bilgi Keşfi (KDD), büyük veri havuzlarının modellenmesi, otomatik ve keşifsel analiz, ek olarak büyük ve karmaşık veri kümelerinden faydalı yeni kalıpların tanımlanması için organize bir süreçtir. Uygulama etki alanının anlaşılması, Keşfin gerçekleştirileceği bir veri kümesinin seçilmesi ve oluşturulması, önişleme ve temizleme, Veri dönüştürme, Uygun Veri Madenciliği görevini seçme, Veri Madenciliği algoritmasını seçme, Veri Madenciliği algoritmasını kullanma, Değerlendirme ve Keşfedilen bilgileri kullanma, KDD'nin temel adımlarıdır (Maimon & Rokach, 2005).

Veritabanlarındaki Bilgi Keşfinin temel adımlarından bazıları özetlenebilir. Veri temizleme, veri kümelerindeki gürültülü ve yanlış bilgileri algılar, istenmeyen bilgileri kaldırır, ancak eksik değerlerin işlenmesi veri temizlemedeki önemli görevlerden biridir. Ayrıca, önemli bir adım doğru veri madenciliği algoritmalarının seçilmesidir. Bu adım kullanılan algoritmaya bağlı olarak uygun parametrelerin belirlenmesini ve doğru paternin bulunmasını içerir. Veri madenciliği adımı, araçlar ve gerekli veri madenciliği yöntemleri kullanılarak verilerle ilgilenir. Veritabanlarında Bilgi Keşfi (KDD) sürecinin ana adımı Veri madenciliği olup, bu adım işin çoğunu içerir (Fayyad vd., 1996).

1.2. Meme Kanseri: Genel Bilgiler

Dünya sağlık örgütüne (WHO) göre, sadece 2018'de 9.6 milyon insan kanser nedeniyle öldü, ayrıca tüm dünyada görülen 18.1 milyon yeni vaka iddia edildi. Kanser genel olarak, vücuttaki belirli doku türlerinde yetişen anormal hücreler olarak tanımlanabilir ve daha sonra vücudun diğer bölgelerine yayılır (Organization, 2018).

Teşhis sayısı açısından meme kanseri ve Akciğer kanseri en önde gelen kanser türüdür ancak kadınlar arasında meme kanseri en yaygın kanser türüdür. Meme Kanseri'nde yaklaşık (2.1 milyon tanı konmuştur, bu da 2018'deki toplam vakaların % 11.6'sını oluşturmaktadır. Ayrıca yaklaşık 627.000 kişi meme kanserinden sağ çıkamamış, ölüm oranları teşhis edilen tüm vakaların yaklaşık % 11.6'sı olmuştur. Mevcut sağkalım oranları geçerli olmaya devam ederse, 2025 yılında (21.5 milyon), 2035 yılında 26.84 milyon, ve 2040 yılında 29.5 milyon yeni kanser vakasının olacağı tahmin edilmektedir (Organization, 2018).

Tümörler iyi huylu ve kötü huylu olarak iki türe ayrılırlar, kansere neden olan kötü huylu hücrelerdir. Bu aile öyküsü dışında genetik risk faktörleri, radyasyona maruz kalma, alkol tüketimi ve daha önce bazı iyi huylu göğüsler teşhisi konması, meme kanserine yol açan risk faktörleridir (Prevention, 2019).

Meme kanserinin evreleri (Memorial Sloan Kettering Cancer Center, 2019).

Evre 0: Kanser hücrelerinin komşu normal dokuyu istila ettiğine dair bir kanıt yoktur.

Evre I: Tümör hücreleri normal çevre meme dokusuna yayılmıştır, ancak lenf düğümlerinde kanser yoktur.

Evre II: Kanser kolun altındaki lenf düğümlerine yayılmıştır ya da memede 20 ila 50 milimetre tümör vardır.

Evre III: Kanser memeye daha da yayılmıştır ve yakındaki dört ila dokuz lenf nodunda bulunur, kanser de cilde yayılmış olabilir.

Evre IV: Memenin ötesindeki uzak lenf düğümlerine yayılmıştır, bu muhtemelen beyin, kemikler, akciğerler veya karaciğer gibi vücut organlarını içerir.

Erken teşhis ve tarama olmak üzere iki stratejiye dayanan kanserin erken saptanması ölüm oranlarının azaltılmasının anahtarıdır. Tıbbi alanlardaki kaynakların sınırlı olduğu ve sağlık hizmetlerinin yetersiz olduğu gözenek ülkelerde sağkalım oranı, meme kanserinin geç aşamalarda teşhis edilmesi nedeniyle % 10-40 arasındadır (Organization, 2018). Meme kanseri erken teşhis edildiğinde tedavi edilmesi daha kolaydır. Meme ultrasonu, diagnostik mamogram ve manyetik rezonans görüntüleme (MRI) gibi meme kanserini bulmak veya teşhis etmek için çeşitli testler kullanılır. Mamogramlar, meme kanserini erken bulmak için en iyi testlerdir (Imaginis, 2019) (Prevention, 2019). Ancak kanserin bazen Mamogramlar tarafından bulunamamasının bazı dezavantajları vardır. Tarama mamografisi, 40-74 yaş arası kadınlarda meme kanseri ölümlerinin sayısını azaltmaya yardımcı olabilir, ancak bugüne kadar yapılan araştırmalarda, 40 yaşın altındaki kadınlarda düzenli tarama mamografisinden yarar gösterilmemiştir (Inflammatory Breast Cancer, 2016).

1.3. Literatür İncelemesi

S. Poorani tarafından yapılan bir çalışmada ark. Meme kanserini tahmin etmek için farklı düğümler ve gizli katmanlara sahip Derin Sinir Ağı (DNN) sınıflandırıcılar kullanıldı. Daha sonra Karar ağacı ve Destek Vektör Makinesi gibi makine öğrenme algoritmaları da aynı verilerle eğitildi. Bu çalışmada araştırmacılar doğruluk oranını hesapladılar. Destek Vektör Makinesi, Karar

ağacı ve DNN'nin sırasıyla % 65.96, % 40.83 ve % 75.94 doğruluk oranları gösterdiği gözlenmektedir (Poorani & Balasubramanie, 2019).

V. Jonathan Silva Araújo ve ark. WEKA veri madenciliği aracını kullanarak melez yapay zeka modeli (bulanık sinir ağı FNN) ve meme kanseri veri kümeleri üzerinde sınıflandırma algoritmaları kullanılarak karşılaştırmalı bir çalışma sundu. Seçilen modeller arasında Çok Katmanlı Perceptron, C 4.5 adı verilen karar ağacı modeli, Naive Bayes, Zero R ve Random Tree vardı. Verilerin % 70'i eğitim % 30'u test için kullanılmıştır. Elde edilen bulanık sinir ağına sahip modeller % 81 Doğruluk, 0.8052 AUC % 82 Hassasiyet ve % 81 Özgüllük ile diğer algoritmalardan daha iyidir (Araújo vd., 2019).

Bir projenin amacı, rutin kan analizinin sonuçlarını farklı ML stilleriyle işlemek ve bu stillerin açıklama için ne kadar etkili olduğunu anlamaktır. Destek Vektör Makinesi, Yapay Sinir Ağı, K-En Yakın Komşu ve standart Aşırı Öğrenme Makinesi olarak kullanılan algoritmalar. Bu çalışmada araştırmacılar Doğruluk oranları oranını ve eğitim sürelerini hesapladılar. Çalışmalarıyla elde edilen standart Extreme Learning Machine'in sınıflandırma doğruluğu ve eğitim süreleri diğer algoritmalardan daha iyidir. Standart Aşırı Öğrenme Makinesi'nin (ELM) % 80 doğruluk oranları ve eğitim süresi (0.0075), Yapay Sinir Ağı (YSA) % 79.4 doğruluk oranları ve eğitim süresi (0.4282), K-En Yakın Komşu (KNN) gösterdiği gözlenmiştir % 77.5 doğruluk oranları ve eğitim süresi (0.15781) ve Destek Vektör Makinesi (SVM) % 73.5 doğruluk oranları ve eğitim süresi (0.1866) göstermiştir (Aslan vd., 2018).

M. Patrício Tarafından yapılan bir çalışma ve ark. Lojistik regresyon, rastgele ormanlar, Destek Vektör Makineleri performansını meme kanseri veri setinde değerlendirdi. Bu çalışmada araştırmacılar duyarlılık, özgüllük ve AUC'yı hesapladılar. Resistin, BMI, Glikoz ve Yaş kullanan SVM modellerinin daha iyi oranlar sergilediği gözlenmiştir. % 85 ila % 90 arasında bir özgüllük elde etmişler, duyarlılık % 82 ila % 88 arasında değişmekte ve AUC için % 95 güven aralığı [0.87, 0.91] idi (Patrício vd., 2018).

E. Purwaningsih Tarafından yapılan bir çalışma. Sinir Ağının (PSO) ve Destek Vektör Makinesinin (SVM) meme kanseri veri setindeki performansını karşılaştırır. Bu çalışmada, araştırmacılar bu algoritmaları kullanarak doğruluk

ve AUC oranını hesapladılar. Bu çalışmada, PSO tabanlı Sinir Ağı modelinin Partikül Sürüsü Optimizasyonu bazlı SVM'den % 84.55 daha yüksek bir doğruluk değerine sahip olduğu gözlenmiştir (Purwaningsih, 2019).

Gültepe ve ark. WEKA veri madenciliği araçlarını kullanarak biyomedikal alanda kullanılan veri madenciliği algoritmaları üzerinde karşılaştırmalı bir çalışma uyguladı. Araştırmalarında J48, Çok Katmanlı Perceptron (MLP), K-En Yakın Komşu (K-NN) ve Destek Vektör Makinesi (SVM) algoritmasını içeren farklı veri madenciliği yöntemleri uyguladılar. Bu çalışmada, % 76.92 ile J48 algoritması en yüksek doğruluk oranını kaydetmiştir (Gültepe & Kartbayev, 2019).

2. VERİ MADENCİLİĞİ YÖNTEMLERİ VE MALZEMELERİ

Bu bölümde, sınıflandırma yöntemleri hakkında genel kavramları, bir algoritmanın doğruluğunun nasıl hesaplanacağını ve aşırı uyumu tanımlayacağız.

2.1. Sınıflandırma Algoritmaları

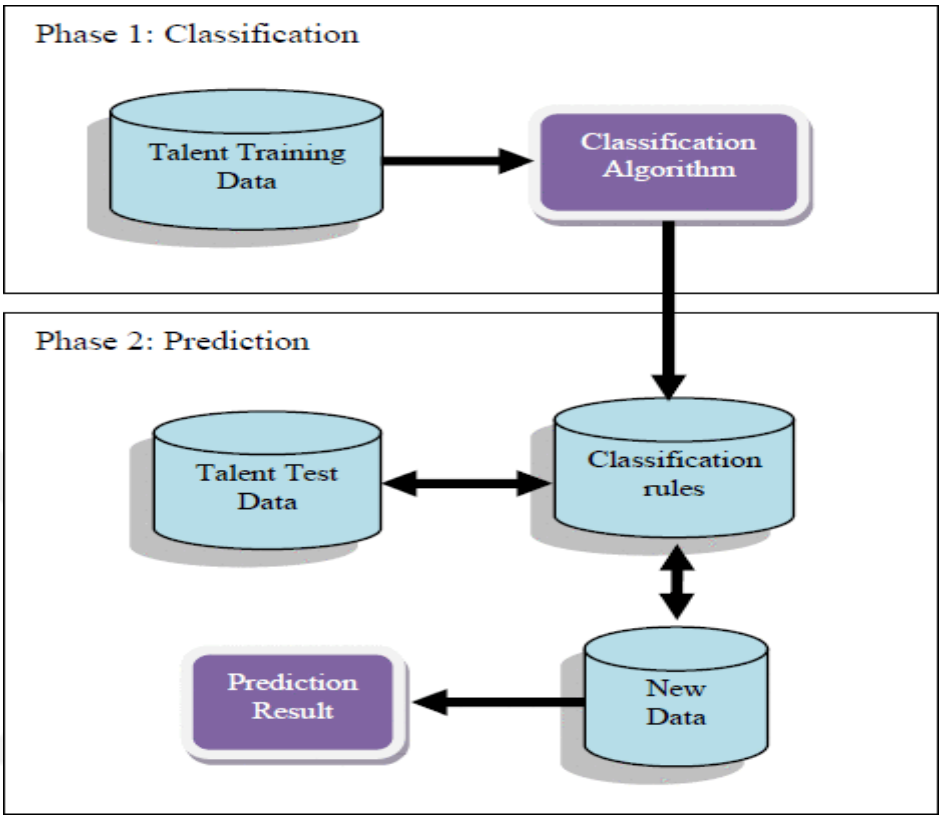
Veritabanları, büyük boyutlarından dolayı eksik ve belirsiz verilere karşı son derece savunmasızdır, bu nedenle ön işleme ve ardından sınıflandırma çok önemli aşamalardır (Singh & Thakral , 2018). Sınıflandırma algoritmalarının amaçlarından biri, bir dizi değişken kullanarak eğitim seti üzerinde veri modelleri oluşturmak için kullanılırdır. Sınıflandırma, büyük veri kümesindeki gizli fikirleri bulmak için kullanılır ve tıp alanında en yaygın olarak kullanılan veri madenciliği yöntemlerinden biridir (Han vd., 2011).

Sınıflandırma algoritmaları iki aşamadan oluşur, öğrenme aşaması ve test aşamasıdır. Öğrenme aşaması, eğitim örnekleri kullanılarak önceden belirlenmiş bir veri sınıfları grubunu tanımlayan bir sınıflandırma modeli oluşturulurdur. Test aşaması, modelin belirli veriler için sınıf etiketlerini tahmin etmek ve sistemin çıktılarını sınıflandırmak için kullanılır (Han vd., 2011) (Şekil 2.1).

Sınıflandırma, verileri farklı örneklere göre genelleme sürecidir, başka bir deyişle verileri farklı sınıflara ayırmak için kullanılan model bulma işlemidir diyebiliriz (Kumar & Verma, 2012).

Sınıflandırma algoritması, girdi özelliklerini tartan bir işlevdir ve çıktı bir sınıfı gerçek değerlere ve diğerini yanlış değerlere ayırmak için bir tür denetimli öğrenmedir (Iaizzo, 2019) (Aggarwal, 2015). Eğitim veri setinin yapısını öğrenmek için insandan türetilmiş örnek veri seti kullanıldığından denetimli öğrenme algoritmaları kullanacağız (Han vd., 2011) (Annasaheb & Verma, 2016). Bu adımdan sonra, modelin doğruluğunu hesaplamak için test verilerini kullanarak bir tahmin modeli oluştururdur (Artsın, 2019) (Levy vd., 2019) (Suthaharan, 2015). Bu araştırmada, insanlar sınıflandırma modelleri

kullanılarak rutin kan analizine dayanarak meme kanseri ile "enfekte" veya "enfekte olmamış" olarak sınıflandırılabilirdir (Tekieh & Raahemi, 2015).



Şekil 2.1. Sınıflandırma ve Tahmin (Jantan vd., 2009)

2.2. Naive Bayes (NB)

Basitliği ve etkinliği ile karakterize edilen veri sınıflandırması için algoritmamsı yöntemlerden biridir. Öğrenilen özellik ağırlıklarının uygulandığı yerler (Jiang vd., 2016).

2.2.1. Bayes teoremi

Bayes teoremi, verilen verileri kullanarak diğer olasılıkları bildiğimizde bir olasılık bulmanın bir yolu veya matematiksel bir formülüdür (Leman, 2018).

$$\mathrm{P(H|X)} = \frac{P(X|H)P(H)}{P(X)} \tag{2.1}$$

Burada H hipotezdir ve X kanıttır, bu nedenle X'in meydana geldiği göz önüne alındığında, H'nin gerçekleşme olasılığını bulabiliriz.

(2.1) denklemin göz önüne alındığında:

- $P(H \mid X)$ Posterior olasılıktır ve kanıtın halihazırda meydana geldiğini bilerek ortaya çıkan hipotez olasılığı anlamına gelir.
- $P\ (X \mid H)$ Hipotezin gerçekten meydana geldiğini bilerek kanıtların ortaya çıkma olasılığıdır.
- $P\ (H)$ Hipotezin ortaya çıkma olasılığıdır.
- $P(X)$ Kanıtın ortaya çıkma olasılığıdır (Villejoubert & Mandel , 2002).

2.2.2 Naive bayes

Naif Bayes, Bayes ağınından en basit şeklidir. Sınıf değişkeninin değeri göz önüne alındığında tüm özniteliklerin bağımsız olduğu ve koşullu bağımsızlık denir. Genellikle Duyarlılık Analizi, Metin sınıflandırması ve Spam Filtreleme uygulaması için kullanılır (Zhang, 2004).

2.2.3. Gauss naïve bayes sınıflandırması

Gaussian Naive Bayes, sırasıyla bir sınıf verilen veri kümesindeki sınıfların önceden olasılık hesaplamasını içeren olasılıklı bir Yaklaşıma sahip bir algoritmadır. Sayısal veri dağılımının normal olduğunu ve öznitelik değerlerinin sayısal olduğunu varsayarsak, Denklem (2.2)'de standart olasılık yoğunluk fonksiyonu kullanılır (Griffis vd., 2015) (Jahromi & Taheri, 2017) (Mavani vd., 2019).

$$P(x_k|C_i) = f(x_k, \mu_{c_i}, \sigma_{c_i}) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{c_i}} e^{-\frac{(x_k - \mu_{c_i})^2}{2\sigma_{c_i}^2}} \quad (2.2)$$

Burada σ standart sapmadır, μ popülasyon ortalama ve σ^2 varyans.

2.3. Random Forest Algoritması (RF)

Karar ağacı, her bir yaprak düğümünün bir kararı temsil ettiği ve her dal düğümünün birkaç alternatif arasında bir seçimi temsil ettiği bir ağaçtır. Karar verme amacıyla, karar ağacı genellikle bilgi elde etmek için kullanılır. Karar ağacı, her bir düğümü karar ağacı öğrenme algoritmasına göre özyinelemeli olarak ayıran bir kök düğüm ile başlar (Peng vd., 2009). Farklı veri kümeleri üzerinde sınıflandırıcıların değişkenlik göstermesinden dolayı, ara sıra tek sınıflandırıcıdan alınan sonuçlar verimli olmayabilir. Bu durumda başarı oranını ve doğruluğunu artırmak için sınıflandırıcılar birleştirilir. Oylama yöntemi olasılık değerlerini ya da sayısal tahminleri kullanarak sınıflandırıcıları birleştirme bunlardan biri yöntemidir. Her sınıf için sayısal tahminleri ya da olasılık değerler toplanır. En yüksek toplama olan kazanırdır (Akar & Güngör, 2012).

Random Forest sınıflama yöntemi 2001 yılında Adele Cutler ve Leo Breiman tarafından geliştirilmiştir. İçerisinde oylama metodunu barındıran bir sınıflama yöntemidir (Breiman, 2001). Karar ağaçları, veri setinden bootstrap tekniği ile

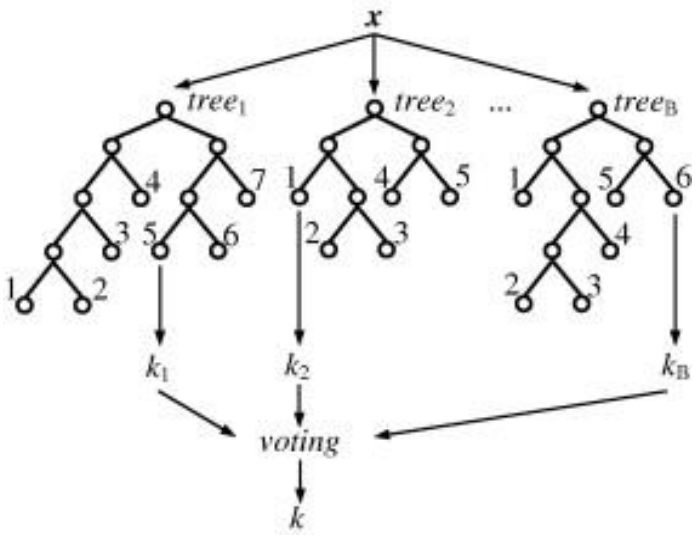
çekilen örneklerden oluşturulur. Birbirinden bağımsızdır. Birbirinden farklı olarak kurulan regresyon ve sınıflama karar ormanları topluluğunu oluşturur.

RF yönteminde ağaçlar, seçilen bootstrap örneklemleri m adet tahminci ile oluşturulur. Toplam tahmincinin m adet tahminci sayısından oldukça büyük olmasına dikkat edilir (James vd., 2013).

2.3.1. RF tanımı

RF, bir ağaç öngörücüleri toplamı veya karar ağaçları modellerinin bir kombinasyonudur. RF yöntemi nedeniyle, kontrolün aşırı uyumunu ve tahmin doğruluğunu iyileştirmek için kullanılan ortalama denetlenir. Bu nedenle orijinal eğitim setinin her bir alt örneğini (ağacı) girdi olarak eğitti. Bu, birbirinden tamamen farklı bir dizi karar ağacı döndürür. RF yönteminin tahminlerinin kesinliğinin sebepleri ağaçlar arasındaki düşük ve yanlılığı düşük sonuçlar vermesi korelasyondur. Büyük ağaçların oluşturulması için düşük yanlılık miktarı oldukça sonucu elde edilir (Breiman, 2001).

RF Sınıflandırma makinesi, mükemmel bir geçerlilik sunup kısa sürede sonuç verir. Dengesiz bir dağılım veya kayıp verili sergileyen veri setlerini kullanarak sonuçlar verir. Topluma ağaçlar eklenmeli, test setine ait hata tahmini için yanlılığı düşük sonuçlar vermeye gösterir. RF'nin genel mimarisi, (k1, k2, kB) sınıf etiketleridir ve B ağaçların sayısıdır (Englund & Verikas, 2012) (Şekil 2.2).



Şekil 2.2. RF'nin genel mimarisi (Englund & Verikas, 2012)

2.3.2. RF algoritmasının adımları

Kullanılan temel parametreler Yaptığımız çalışmada tahminci parametresini ormandaki ağaç sayısı olarak kullandık. Kalite fonksiyonunu ölçmek için kriter parametresi kullanılır.

RF Algoritmasının Adımları

1: Veri kümesini açılır, bu veri kümesini, test veri seti (OOB out of bag) ve eğitim veri seti olarak ikiye ayrılır.

2: Her düğümün bölünmesinde toplam p tane tahminci değişkenden m tanesi rasgele seçilen ve m<p koşulu sağlanmalıdır.

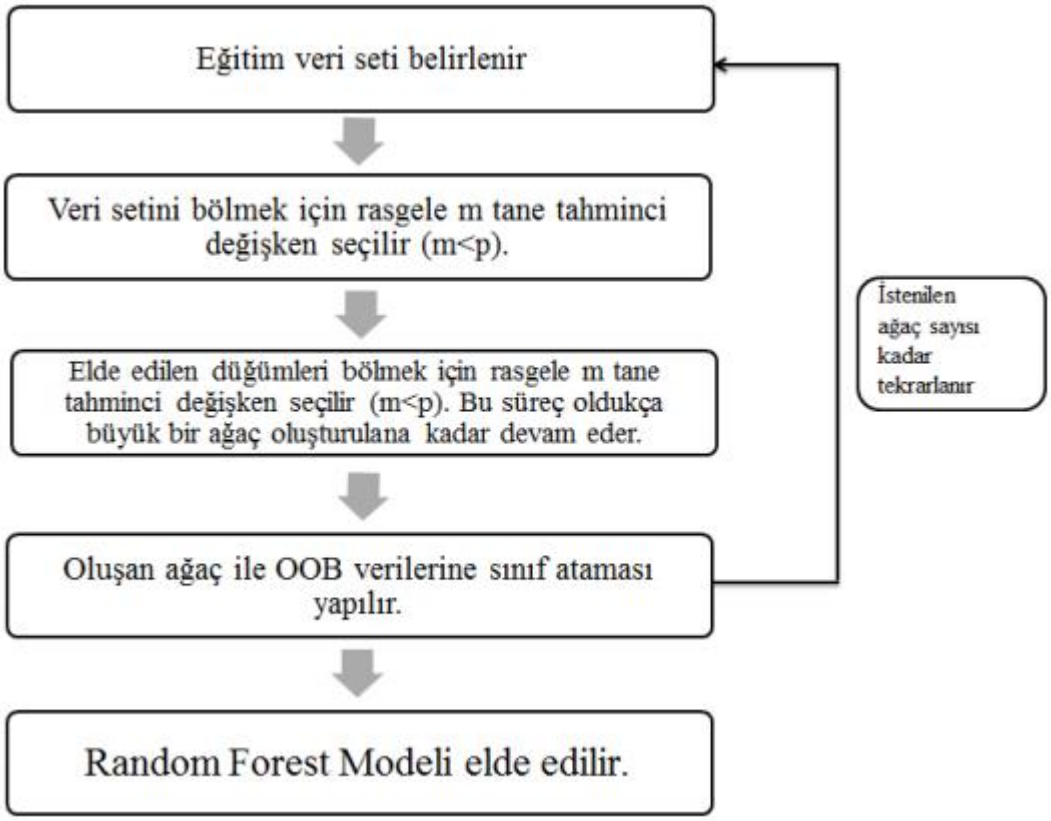
3: Test veri seti ağacın en tepesinden bırakılır, ama daha önce her yaprak düğüme bir sınıf atanır ve sonunda bu veri setinde yer alan her gözleme atanan sınıf kaydedilir.

4: Önceki tüm aşamalar (1., 2. ve 3.) B defa tekrar edilir.

5: Kullanılmayan gözlemler test veri setinde bir değerlendirme yapılır.

6: İncelenen bir gözlemin kaç defa sınıflandırıldı hangi kategorilerdeki sayılır.

7: Her gözleme üzerinden oy çoğunluğu ile sınıf ataması yapılır (Yılmaz, 2014) (Şekil 2.3).



Şekil 2.3 RF Adımları (Yılmaz, 2014)

2.3.3 RF Genelleme hatası ve parametreleri ayarlama

Ağaç oluşturulurken bazı gözlemler yer almaz. Özellikle bir bootstrap örneklemi seçildiği zaman. Bu gözlemler, OOB ile genelleme hatasına yönelik bir iç tahmin yapılır. OOB hata oranını almak için. Her ağaç bir sınıf değeri tahmin edip tahminler kaydedilir. Ağaçlardaki hata oranı tahminlerinin ortalaması alınarak genelleme hatası hesaplanabilir. Tüm gözlemlerin ortalamasını alarak, genel hata oranı hesaplanabilir.

Birçok sınıflandırma problemi göre, $m = \sqrt{p}$ Denklemi, her düğümde rastgele seçilecek değişkenlerin sayısını hesaplamak için kullanılır. Burada p tahminci değişkenleri sayısını göstermesi, m parametresi $m = p/3$ olarak elde edilir (Zhang & Ma, 2012) (James vd., 2013).

2.3.4. Gini indeksi

Gini indeksi bir düğüme atanmış örneklemin eşitsizlik ya da karışıklık ölçer. Bir örnek, iki sınıflı bir sınıflandırma sorununda p; $1 - p$ düğümünde yer alan negatif gözlemlerin oranını ve k düğümünde yer alan pozitif gözlemlerin oranını göstersin. Yani bu durumda Gini indeksi, k düğümünde yer alan aşağıdaki bu şekilde hesaplanır (Denklem 2.3):

$$\mathbf{G_k = 2_p(1 - p)} \tag{2.3}$$

Burda G Gini, $1 - p$ negatif gözlemleri oranı ve k pozitif gözlemleri oranı.
Ne kadar Gini değeri küçülürse, düğüm o kadar kesindir. Her düğümde üzerinden bölünme gerçekleştiğinde, yeni Gini değeri, bölünen düğümün Gini değerinden daha küçük olur (Zhang & Ma, 2012).

2.3.5. Kayıp değerleri ve örnekler arası uzaklık

Veri analizlerindeki en çok karşılaşılan zorluklardan biri, veri setinin tutarlı olup olmadığını bir şekilde fark edememektir. Sapan değerleri ya da bilinen sınıflarda alt grup oluşumu var mıdır? Çok sınıflı durumlarda bazı gözlemler birbirinden ayrıyken, bazıları birbirleriyle örtüşür mü? RF bu soruların anlamak için veri setine bir bakış sağlar. Bunu, iki gözlem arasındaki uzaklığı hesaplayarak yapar.

Kayıp değer veri setinde sorunlardan biri. RF, kayıp değerleri geliştirdiği bir algoritma ile onların veri setinde kalmasına olanak izin verir. Bu algoritmanın temelini uzaklık ölçüsü oluşturmaktadır (Zhang & Ma, 2012) (James vd., 2013).

2.3.6. RF yönteminin özellikleri

Yerleştirilen veriler açısından ağaçlar birbirinden bağımsızdır. RF yönteminde gözlem sayısının veri setindeki tahminci sayısından büyük olması gerekli değildir. OOB veri setini kullanarak iç hata oranı hesaplar. Bu yöntem model performansını değerlendirmek için kullanılır. Doğruluk yüksektir. Orman oluşturma sürecinde, Kayıp veri tahmini ve yansız genelleştirilmiş hata tahmininde etkin bir yöntemdir (Sleath & Brown, 2015).

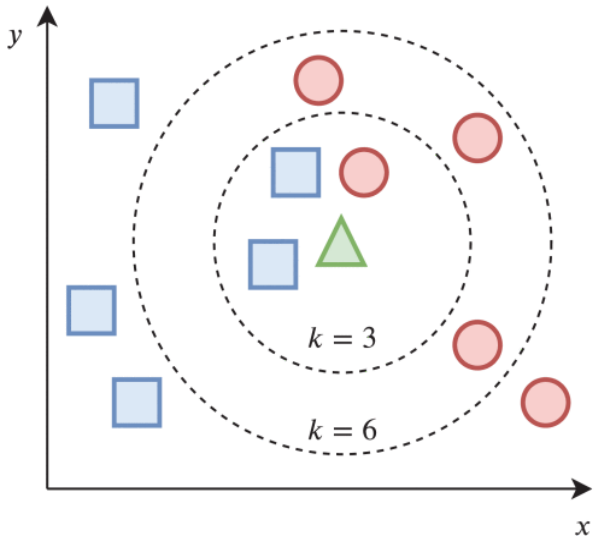
2.4. K-en yakın Komşu Algoritması (KNN)

K-En Yakın Komşular (KNN), sınıflandırma için önemli bir Makine Öğrenimi (ML) algoritmasıdır ve uygulanması diğer yöntemlere göre daha kolaydır (Peng vd., 2018). KNN algoritması, hesaplamalı genomik, tıbbi görüntüleme, veri madenciliği, örüntü tanıma, veri sıkıştırma, yüz tanıma, tahmine dayalı analiz ve metin sınıflandırma gibi çeşitli alanlarda kullanılırdır (Vieira vd., 2019). KNN kuralı, sınıf akıl yürütmesinin en eski yöntemlerinden biridir ve karar verme fikri çok basittir (Xing & Bei, 2020). KNN algoritmasında iki önemli sorun vardır: İlk sorun, algoritmanın performansında büyük önem taşıyan k'nin doğru seçimidir ve ikinci sorun test örneği ve komşuları arasındaki boyutu hesaplamaktır (Zhang, 2016). KNN algoritması çok boyutlu örnekleri mesafelerine göre sınıflandırır nerede sınıflandırma sonucu k en yakın eğitim örneklerinin sınıfından türetilir.

Ayrıca, KNN algoritması veri setini eğitim seti ve test seti olmak üzere iki parçaya ayırır (Vieira vd., 2019) İki sınıflı, kırmızı daire ve mavi kare içeren bir eğitim seti kullanarak yeşil üçgenle temsil edilen iki boyutlu bir örneği sınıflandırmak için KNN algoritması örneği. K=3 için, en yakın komşuların çoğu mavi karelerdir, bu da test örneğinin mavi bir kare olarak sınıflandırılması gerektiğini belirler ve k=6 için komşuların çoğunluğu kırmızı dairelerdir, dolayısıyla atanan sınıf da kırmızı bir daire olur, Şekil 2.4'de gösterildiği gibi.

2.4.1. KNN algorithm phases

1. Mesafe hesaplama, test örnek ve eğitim örnekleri set arasındaki mesafeler belirlenir, nerede mesafeyi hesaplamak için farklı metrikler, örneğin Manhattan mesafesi ve Öklid mesafesi uygulanabilir (Vieira vd., 2019) (Pandit & Gupta, 2011).
2. İlk aşamada hesaplanan mesafeler sıralanır. En küçük k değerleri çıkarılır. Bu aşama basit bir sokma sıralaması kullanılarak uygulandı ve ilk k elementleri belirlendikten sonra sokma sıralaması kesilebilir (Vieira vd., 2019).
3. Karşılık gelen k eğitim örnekleri arasında en yaygın sınıf belirlenir ve test örneğine atanır (Vieira vd., 2019).



Şekil 2.4 KNN algoritması çalışmasını (Vieira vd., 2019)

2.4.2. KNN algoritması mesafe ölçümü

K-en yakın komşu (KNN) sınıflandırıcı, bu sınıflandırıcının çekirdeğinin temel olarak eğitim paradigmaları ile test edilen paradigmalar arasındaki mesafenin ölçülmesine bağlı olduğu en basit ve en yaygın sınıflandırıcılardan biridir. Mesafe ölçümü, bir noktaya en yakın komşuları bulmak için en önemli anahtardır. Öklid uzaklığı, Minkowski uzaklığı, Manhattan uzaklığı ve Gini vb. gibi çeşitli uzaklığı ölçümleri vardır. Öklid uzaklığı, Manhattan ve Minkowski uzaklığıdan daha genel olarak kullanılır (Abu Alfeilat vd., 2019).

Öklid boyutu verilen şekilde hesaplanabilir Denk. (2.4):

$$D_{\text{Euclidean}}(X, Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{2.4}$$

Manhattan boyutu verilen şekilde hesaplanabilir Denk. (2.5):

$$D_{\text{Manhattan}}(X, Y) = \sum_{i=1}^{n}|X_i - Y_i| \tag{2.5}$$

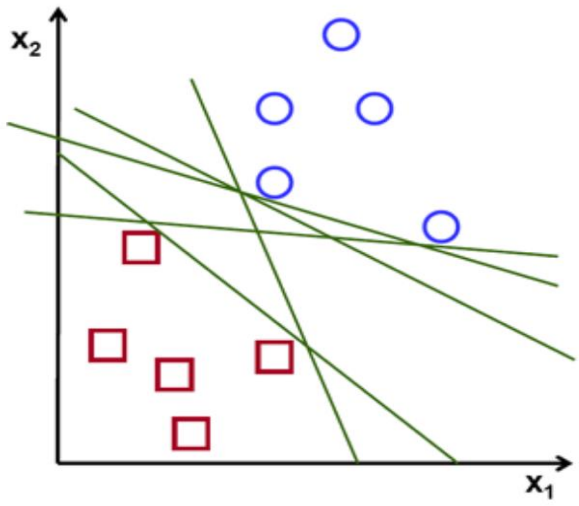
Minkowski boyutu verilen şekilde hesaplanabilir Denk. (2.6):

$$D_{\text{Minkowski}}(X, Y) = \left(\sum_{i=1}^{n}|(X_i - Y_i)|^p\right)^{1/p} \tag{2.6}$$

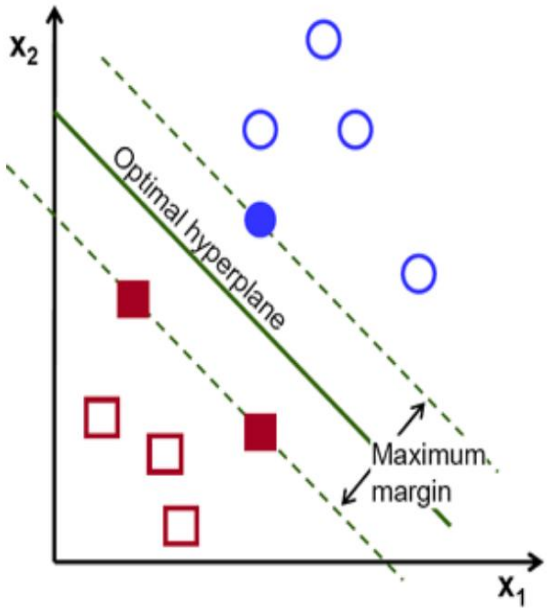
Sabit bir $p \in (0, \infty)$.

2.5. Destek Vektör Makinesi (SVM)

Destek vektör makinesi (SVM), en iyi hiper düzlemi oluşturan, veri miktarı ile çok iyi performans gösteren iki gruplu sınıflandırma için sınıflandırma algoritmalarını kullanan denetimli bir makine öğrenme modelidir. Destek Vektör Makineleri (SVM), makine sınıflaması ve makine öğrenimi alanında yeni bir öneme sahiptir. Sınıflandırma, girdi alanındaki doğrusal olmayan veya doğrusal bir ayırma ile yapılır (Vapnik, 2000) (Murty & Vishwanathan, 2002). SVM uygulamaları iyi sonuçlar vermesi, yüksek performans isteyen uygulamalarda ve basit fikirler üzerine kurulmuş olması kullanabilmesinden dolayı avantajlıdır. SVM'ler El yazısı, yüz tanınması, ses tanınması, uzaysal veri analizi ve kanser hücrelerinin tanınması gibi birçok alanda da kullanılır (Karadağ, 2013). SVM'nin amacı, bir veri kümesini iki sınıfa ayıran bir karar yüzeyi veya bir hiper düzlem bulmaktır. Şekil 2.5, 2.6'te gösterildiği gibi. Destek vektörleri hiper düzlemi belirleyen veri noktalarıdır. Veri noktalarını iki kategoriye ayıran birçok hiperplan vardır. Veri noktalarını iki kategoriye ayıran birçok hiper uçak olduğundan, ana amaç, onu iki sınıf arasındaki mesafe marjına sahip n-boyutlu bir alanda bulmaktır (Srivastava & Bhambhu, 2010) (Kotsiantis, 2007).



Şekil 2.5 Olası hiperplanlar (Kotsiantis, 2007)



Şekil 2.6 Veri kümesini iki sınıfa ayıran bir hiper düzlem (Srivastava & Bhambhu, 2010)

Şekil 2.5, 2.6 SVM modeli. İki sınıf (kareler karşı daireler) sınıflandırıldı. Veri kümesini iki sınıfa ayıran bir hiper düzlem.

2.5.1. Doğrusal destek vektör makineleri

Doğrusal SVM, veri noktalarını iki ayırma noktasından birine ayırarak sınıflandırma ve tahmin etme eğilimindedir. Tahmin edilen veri noktaları birbirinden uzak duran iki paralel düzlemden en yakınına atanacaktır. Doğrusal SVM'ler, daha fazla hesaplama süresi geliştiren doğrusal veya karesel programları çözme eğilimindedir (Ranganayaki & Deepa, 2019). Doğrusal SVM'de karar düzlemi, iki özelliğe sahip veri noktaları için -1 ve +1 olarak etiketlenmiş iki sınıf arasında düz bir çizgi çizebileceğimiz sınıfları mükemmel bir şekilde ayırır, de olduğu gibi:

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) | (x_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}\} \quad (2.7)$$

$\boldsymbol{n}$ özellik sayısıdır.

$\boldsymbol{yi}$ etiketleri $-\mathbf{1}$ veya $+\mathbf{1}$'dir.

Denklem 2.7'i kullanarak bir hiper düzlemin denklemini yazabiliriz. P1 ve P2 iki paralel düzlemdir

$$P_1: w \cdot x_i + b = +1 \text{ Eğer } y_i = +1$$
$$P_2: w \cdot x_i + b = -1 \text{ Eğer } y_i = -1 \quad (2.8)$$

w hiper düzlemin normal vektörüdür.

x_i giriş vektörüdür.

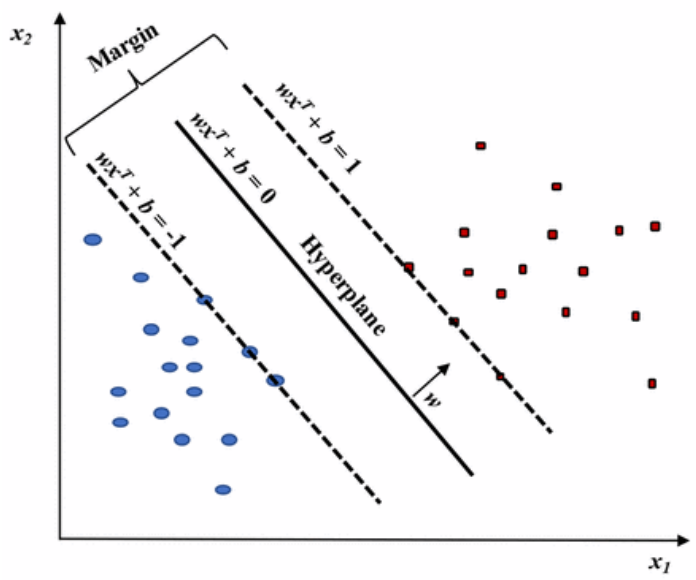
b skalerdir.

Denklem 2.8'yi kullanarak P1 ve P2 arasındaki mesafeyi hesaplayabiliriz.

$$w \cdot x_1 + b = +1$$
$$w \cdot x_2 + b = -1$$
$$= w \cdot (x_1 - x_2) + b = 2 \quad (2.9)$$
$$= \left(\frac{w}{\| w \|} \cdot (x_1 - x_2) + b\right) = \frac{2}{\| w \|} = \frac{2}{\sqrt{w \cdot w}}$$

P1 ve P2 arasındaki mesafe $\frac{2}{||w||} = \frac{2}{\sqrt{w.w}}$ 'dir. Denklem 2.9 gösterdiği gibi.

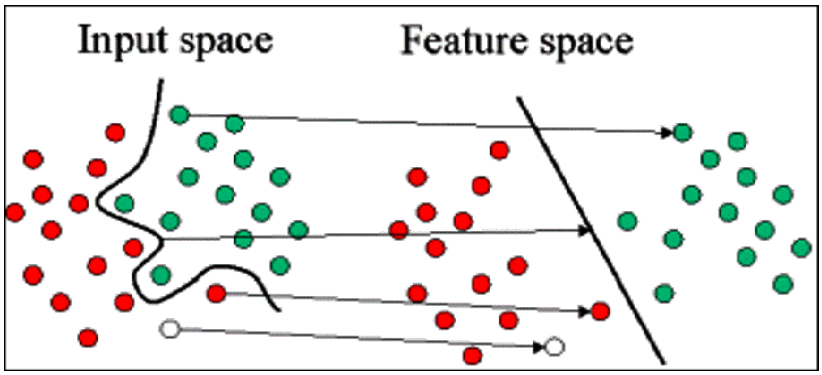
$P0$ Medyan düzlemler olmalı, burada $P0 = w.x_0 + b = 0$. 2.9'ye göre. $P0$ ve $P1$ arasındaki mesafe d^+ 'dir. Burada d^+, pozitif girişten hiper düzleme en kısa mesafedir ve $\frac{1}{||w||}$ hesaplanabilir. Ayrıca, P0 ve P2 arasındaki mesafe d^- 'dir. Burda d^-, negatif girişten hiper düzleme en kısa mesafedir ve $\frac{1}{||w||}$ hesaplanabilir. 2 özelliğe sahip veri noktaları için Şekil 2.7'de gösterildiği gibi (Huang vd., 2018).



Şekil 2.7 Doğrusal SVM modeli. İki sınıf (kırmızıya karşı mavi) sınıflandırıldı (Huang vd., 2018)

2.5.2 Doğrusal olmayan destek vektör makineleri

SVM, doğrusal sınıflandırmalara hizmet etmek için tasarlanmıştır, ancak 20. yüzyılın sonlarında, çekirdeklerin yardımıyla doğrusal olmayan sınıflandırma için de kullanılmak üzere tasarlanmıştır. Çekirdeğin en yaygın türleri sigmoid çekirdek işlevi, RBF çekirdek işlevi, polinom çekirdek işlevi ve birkaç tane daha vardır. Ana hedef, orijinal egzersiz verilerini haritaya yerleştirmektir, böylece uzayda veriler doğrusal olarak ayrılabilir. Şekil 2.8'de gösterildiği gibi. Bu yeni alanda veriler doğrusal olarak ayrılabilir. Yani tek yapmamız gereken çekirdeği kandırmak, yani iki sınıfı ayıran en iyi hiper düzlemi keşfetmek. Polinom Çekirdek İşlevi, bir özellik alanındaki eğitim örneklerinin orijinal değişkenlerin polinomları ile benzerliğini temsil eder (Han vd., 2011) (Ghosh vd., 2019) (Qizhong, 2007).

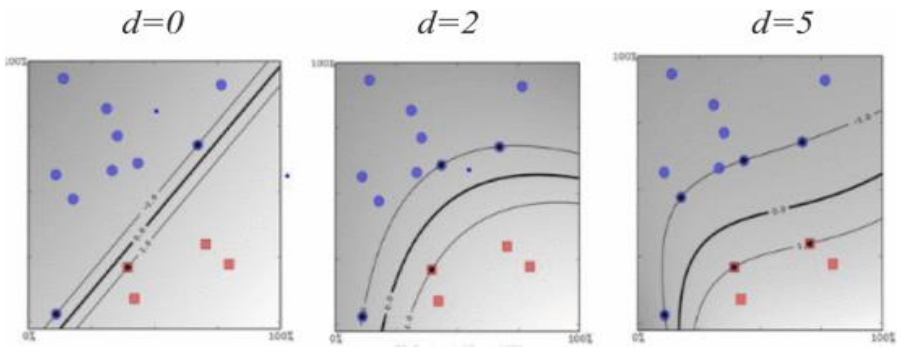


Şekil 2.8 Doğrusal olmayan destek vektör makineleri modeli (Ghosh vd., 2019)

$\emptyset(x)$, $\emptyset(y)$, eğitim tuplesinin nokta ürün şeklidir ve K, Denklem 2.10'a dayanan çekirdek işlevidir.

$$k(x,y) = \emptyset(x)\emptyset(y) \tag{2.10}$$

Daha yüksek boyutlu uzayda doğrusal SVM ararken, veri kümesinin iç çarpımını hesaplamaya gerek yoktur. Denklem 2.9'a göre, nokta ürün çekirdek işlevine eşittir. Böylece $k(x,y)$ kullanabiliriz. Şekil 2.9'da çekirdek fonksiyonlarının farklı derecelerde tepkisi gösterilmektedir.



Şekil 2.9 Çekirdeğin farklı derecelerde yanıtı (Ghosh vd., 2019)

Sigmoid çekirdek, böyle bir durumda çekirdeklerin performansının kişinin ihtiyaç duyduğu çapraz doğrulama seviyesine bağlı olduğu sinir ağlarından kaynaklanır Denklem (2.11) 'da gösterilmiştir.

$$K(x, y) = \tanh(\propto x^T y + c) \tag{2.11}$$

Radyal Baz Fonksiyonu (RBF) Çekirdek, değeri Öklid mesafesine bağlı olan ve SVM'de eğitim verilerinin (2.12), (2.11) 'de verilen kaynaktan sınıflandırılması için kullanılan en çekirdek olan gerçek değerli bir fonksiyondur (Ghosh vd., 2019).

$$\Phi(x, c) = \Phi(\| x - c \|) \tag{2.12}$$

$$K(x, x') = \exp(-\gamma \| x - x' \|^2) \tag{2.13}$$

2.6. Lojistik Regresyon (LR)

LR, bağımsız ve bağımlı değişkenler arasındaki ilişkiyi belirlemek için kullanılır. Ek olarak, bağımlı değişken yalnızca iki değer aldığında kullanılır. LR analizinde amaç, bağımsız ve bağımlı değişkenler arasındaki ilişkileri, en iyi uyuma en az değişken ile sahip olacak şekilde tanımlamaktır. Lojistik eğri 0 ile 1 arasındaki değerlerle sınırlı göre, LR tekniği, iki değere sahip sonuçları sınıflandırırken iyidir (Çokluk, 2010).

Lojistik regresyonda parametreler, analitik olarak elde edilemediğinden dolayı maksimum olabilirlik tekniğinden yararlanılarak tahmin yapılmaktadır (Albayrak, 2006).

Lojistik fonksiyonun Denklemisi (2.14)'deki gösterir:

$$g(z) = \frac{1}{1+e}, R \rightarrow [0,1] \tag{2.14}$$

Burada $g(z)$ değeri, 1 ile 0 arasında almaktadır. Eğer aldığı değer 0,5 değerinden küçük ise 0'e yaklaştığı, ama 0,5 değerinden büyük ise 1'a yaklaşır (Balaban & Kartal, 2015).

Bu çalışmada, Coimbra veri setinin sınıflandırılması için iki terimli lojistik regresyon kullandık.

Lojistik regresyonun matematiksel kavramı, tahmin değişkenleri ile sonuç değişkeni arasındaki ilişkiyi, olasılıkların doğal logaritması açısından göstermektir.

Lojistik regresyon modeli şu şekilde yazılabilir (Denklem 2.15):

$$logit(Z) = In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X \quad (2.15)$$

Burada Z, "0" ve "1" olarak kategorize edilen ikili bir bağımlı değişkendir.

X bağımsız bir değişkendir.

π, Z sonucunun gerçekleşme olasılığıdır.

$\frac{\pi}{1-\pi}$ Başarı şansıdır.

β_0, kesişme olarak bilinir.

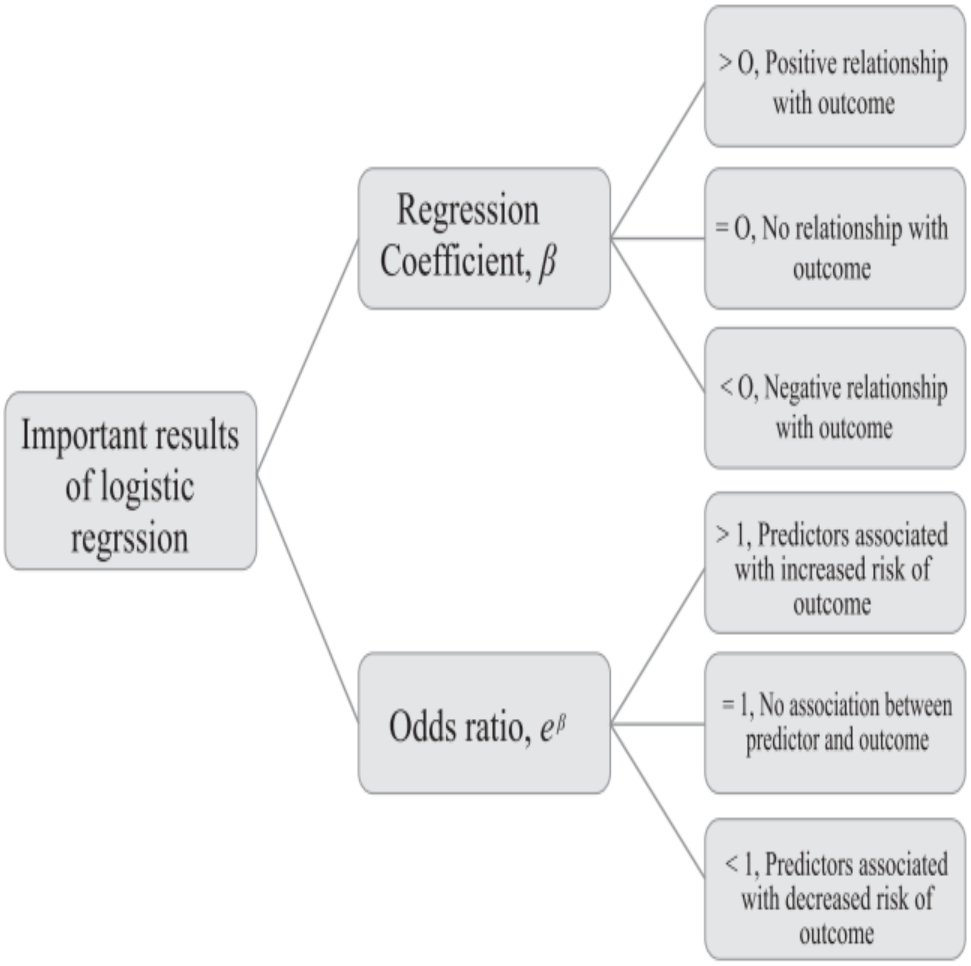
β_1 Eğim olarak bilinir.

Belirli bir X tahmini değeri için Z sonucunu tahmin edebiliriz.

Denklem (2.16) 'ün gösterdiği gibi, Lojistik regresyon birden fazla yordayıcı için genişletilebilir.

$$logit(Z) = In\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X + \cdots + \beta_p X_p \quad (2.16)$$

Regresyon parametresi βs ağırlıklı en küçük kareler yöntemi veya maksimum olasılıkla tahmin edilebilir. B1 ... βp değeri, Z ve X'lerin logitleri arasındaki uygunluğa işaret eder (Abedin vd., 2016) (Şekil 2.10).



Şekil 2.10 Lojistik regresyon tahminleri (Abedin vd., 2016)

2.7. Sınıflandırma Yöntemleri: Genel Bakış

Uyumsuz veri kümesi, makine öğrenme algoritmalarının sınıflandırma performansı üzerinde ciddi bir olumsuz etkiye neden olabilir. Bu nedenle, sınıflandırıcıların doğruluk oranının ve performansının ölçülmesi, görünmez veriler üretildikten sonra önemli faktörlerdir. Sınıflandırıcının kaçınılmaz hatalardan kaçınmasına ve algoritmanın daha iyi çalışmasına yardımcı olan bazı yöntemler ve püf noktaları olduğu yerlerde. Yaygın sınıflandırma algoritması sorunları, doğruluk hesaplaması, Normalleştirme ve aşırı sığdırmanın hesaplanmasını içerir.

2.7.1. Doğruluk

Doğruluk, bir ölçümün gerçeğe ne kadar yakın olduğudur ve dahili modelde doğru cevapsız veya tahmini verilerin yüzdesini bulur. Doğruluk aynı zamanda bir algoritmanın hastayı ve sağlıklı vakaları doğru bir şekilde nasıl ayırt edebildiğini ölçmeye bağlıdır (Baratloo vd., 2015) (Zhu vd., 2010).

İkili sınıflandırma için doğruluk önlemleri, Hasta (hastalık için pozitif), Sağlıklı (hastalık için negatif):

- Gerçek pozitif (TP) = hasta olarak tanımlanan pozitif olarak sınıflandırılan vaka sayısı.
- Gerçek negatif (TN) = Sağlıklı olarak tanımlanan negatif olarak doğru sınıflandırılan vaka sayısı.
- Yanlış pozitif (FP) = hasta olarak tanımlanan yanlış sınıflandırılmış pozitif vaka sayısı.
- Yanlış negatif (FN) = Sağlıklı olarak tanımlanan yanlış sınıflandırılmış negatiflerin sayısı.

Testin doğruluğunu hesaplamak için, matematiksel olarak şu şekilde ifade edilebildiği durumlarda değerlendirilen tüm durumlarda gerçek negatif ve gerçek pozitif oranını bulmalıyız:

$$\text{Doğruluk} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{2.17}$$

Bu değerler, Çizelge 2.1'te gösterildiği gibi, karışıklık matrisi adı verilen 2x2'lik bir matris kullanılarak kategorize edilebilir (Dangare & Apte, 2012).

Çizelge 2.1 karışıklık matrisi

	Öngörülen Sınıf		
Doğru Sınıf		Sınıf=1	Sınıf=2
	Sınıf=1	TP	FN
	Sınıf=2	FP	TN

Çizelge 2.1'e göre, Özgüllük, Duyarlılık, Hassasiyet, Geri Çağırma, Doğruluk (Denklem 2.17'te gösterildiği gibi) ve AUC-ROC Eğrisi hesaplanabilir.

Özgüllük, sağlıklı vakalarda gerçek negatif oranını hesapladığımız sağlıklı vakaları doğru bir şekilde belirleme yeteneğidir ve bu matematiksel olarak şu şekilde ifade edilebilir (Denklem 2.18):

$$\text{Özgüllük} = \frac{TN}{TN+FP} \tag{2.18}$$

Duyarlılık, hasta vakalarında gerçek pozitif oranını hesapladığımız hasta vakalarını doğru bir şekilde belirleme yeteneğidir, bu matematiksel olarak şu şekilde ifade edilebilir (Denklem 2.19):

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{2.19}$$

Hassasiyet, doğru bir şekilde tanımlanmış olan pozitif verilerin yüzdesidir ve ölçülen değerlerin, matematiksel olarak şu şekilde ifade edilebildiğimiz birbirine ne kadar yakın olduğunu gösterir (Denklem 2.20):

$$\text{Hassasiyet} = \frac{TP}{TP+FP} \tag{2.20}$$

Geri Çağırma, matematiksel olarak şu şekilde ifade edilebildiğimiz yerlerde pozitif verilerin yüzde kaçının doğru tahmin edildiğidir (Denklem 2.21):

$$\text{Geri Çağırma} = \frac{TP}{TP+FN} \tag{2.21}$$

2.7.2. Aşırı uyum gösterme

Makine öğrenimi sürecinde, verileri modeli eğitmek için kullanılan eğitim verilerine ve model performansını değerlendirmek için kullanılan test verilerine böleriz. Aşırı Uyum sorunu, model eğitim verilerini mükemmel bir şekilde öğrendiğinde ortaya çıkar. Hata oranının az olduğu yerler. Ama performansı düşük ve test verilerindeki hata oranı yüksek. Aşırı Uyum Gösterme, sınırlı bir veri noktası kümesinde oluşan bir modelleme hatasıdır. Genel olarak, çalışılan verilerdeki özellikleri açıklamak için çok karmaşık bir model oluşturma konusunda çalışırdır. İncelenen veriler genellikle belirli bir hata derecesi içerirdir. Bu nedenle, modeli hatalı verilere uygun hale getirmeye çalışmak, modeli hatalara bulaştırabilir ve tahmin gücünü azaltabilir. Yeterli antrenman verilerinin korunması, aşırı sığmayı önlemenin en iyi yoludur. K-kat çapraz adı verilen çok yaygın bir yöntem vardır. K-katlı çapraz, kullanılan eğitim veri setini bölümlendirmektir. Bu veriler daha sonra birkaç k kez alt grubuna bölünür. İlk grupa, test seti denir, ve geri kalan gruplara eğitim seti denir. İlk grupa, test seti denir, ve geri kalan gruplara eğitim seti denir. Nerede bir modele uymak için kullanılır. Eğitim veri setine dayalı olarak modeldeki yüksek varyans ve eğitim verilerinin göz ardı edilmesi, aşırı uyumun bir işaretidir (Pham & Triantaphyllou, 2008) (Al-Muhaideb & El Bachir Menai, 2013) (Barga vd., 2015) (Mutasa vd., 2020) (Bilbao & Bilbao, 2017).

2.7.3 ROC eğrisi altındaki alan (AUC)

Bu araştırmada, önemli bir ölçü türü de kullandık, ROC eğrisinin altındaki alan (AUC). AUC ölçümü, öğrenme algoritmalarını değerlendirmenin önemli türlerinden biridir. Karşılaştırma, aşağıdaki gibi denklemlerle tanımlanan ROC Eğrisi (AUC) altındaki alan hesaplanarak yapılır (Ling vd., 2003) (Denklem 2.22).

$$AUC = \frac{U_1}{n_1 + n_2} \ldots. \quad (2.22)$$

$(U1)$ Örnek 1 kullanılarak hesaplanan değer.

$(n1)\ ve\ (n2)$, $(1\ ve\ 2)$´ye örneklem büyüklüğünü gösterir.

3. DENEYSEL SONUÇLAR

Bu bölümde, UC Irvine Machine Learning Repository'den alınan meme kanseri veri setine NB, KNN ve SVM sınıflandırma algoritmaları uygulanmaktadır (Anon., 2018). Bu çalışmada kullanılan veri kümesi Meme Kanseri Coimbra Veri Kümesi'dir (Anon., 2018). Temel amaç, bu üç yöntemin performansını, biyobelirteçlere dayalı performans modelleme ölçütleri olarak Özgüllük, Duyarlılık, Doğruluk ve AUC'ya göre yapılan diğer çalışmalarla karşılaştırmamızı sağlayan bir sınıflandırma modeli oluşturmaktır.

3.1. Veri Kümesinin Açıklaması

Bu tezde kullanılan veri seti UCI Makine Öğrenimi Deposunda kamuya açıktır. Ve Joana Crisóstomo, José Pereira, Miguel Patrício, Paulo Matafome, Raquel Seiça, Francisco Caramelo, hepsi Coimbra Üniversitesi Tıp Fakültesi'nden ayrıca Coimbra Üniversite Hastane Merkezi'nden Manuel Gomes tarafından oluşturuldu. Meme Kanseri Coimbra Veri Kümesi iki sınıf ve 116 örnek içerir. Birinci sınıf, meme kanseri olan 64 hasta ve ikinci sınıf, 52 sağlıklı kontrolden oluşmaktadır (Anon., 2018).

Meme Kanseri Coimbra Veri Kümesi sınıflandırma yöntemlerinde, tanı sütunlarının değerlerini tahmin etmek için 10 öngörücü kullanılır. Burada Veri Kümesi öznitelikleri ve açıklamaları Çizelge 3.1'de gösterilmiştir.

Bu araştırmada, Meme Kanseri Coimbra Veri Seti iki sete ayrılmıştır, eğitim verileri ve test verileri, burada veri setinin% 80'i modeli eğitmek içindir (eğitim seti), % 20 veri seti modeli test etmek içindir (test seti). Başka bir deyişle, verileri eğitmek için 92 kayıt ve verileri test etmek için 24 kayıt kullanılacaktır.

Çizelge 3.1 Veri Kümesi öznitelikleri

Attribute #	Attribute Name	Value
A1	Age	Numerical Value
A2	BMI (Body Mass Index)	(kg/m2),
A3	Glucose	(mg/dL)
A4	Insulin	(μU/mL)
A5	HOMA (Homeostatic Model Assessment)	Numerical Value
A6	Leptin	(ng/mL)
A7	Adiponectin	(μg/mL)
A8	Resistin	(ng/mL)
A9	MCP-1	(pg/dL)
Class	Class	{1-Healthy, 2- Patient}

3.2. Sıcaklık haritası

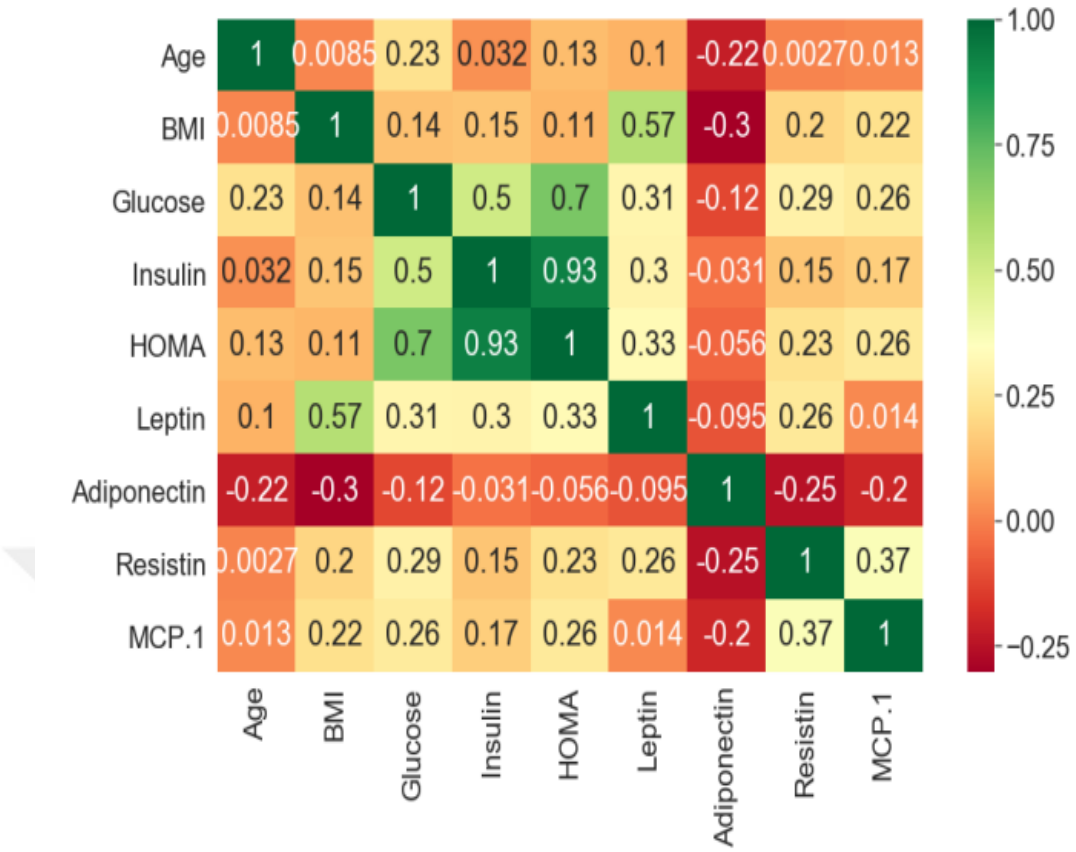
Karmaşık veri kümesini basitleştirmek için, tüm özellikler arasındaki korelasyonları hesapladığı bir ısı haritası kullandık. Değerler Isı haritasında renkli olarak gösterilir. Bu ön analiz, tüm yordayıcı değişkenler arasındaki korelasyonları ve modellemede her bir meme kanseri öngörücüsünün önemi hakkında bize iyi bir istatistiksel bilgi verecektir (Samek vd., 2017).

Şekil 3.1'de gösterildiği gibi korelasyon puanları, mükemmel korelasyon olan 1 ile mükemmel negatif korelasyon olan -1 arasında gider.

Şekil (3.1) 'e göre, iki belirleyicinin, insülin ve HOMA'nın oldukça ilişkili özellikler olduğunu gözlemleyebiliriz (0.93). Bu iki özelliği bir seferde ele alırsak, kötü sınıflandırma performansına neden olabilir. Bu nedenle, özelliklerden birinin veri setinden çıkarılması gerekir. Çoklu bağlantı probleminden kaçınmak için.

HOMA ve glikoz tahmin değişkenleri ikinci en büyük korelasyon değeridir (0.70). HOMA tahmincisi aşağı veya yukarı hareket ettiğinde, glikoz tahmincisi aynı yönde aynı büyüklükte hareket eder.

Öngörücü önlemlerin önemine dayanan bir öngörü analizi, yaş, glikoz, BMİ ve resistinin meme kanseri varlığında çok önemli bir rol oynamış olabileceğini gösterir.



Şekil 3.1 Sıcaklık haritası (Kumar vd., 2020)

3.3. Yöntemlerin Uygulanması

Bu bölümde, daha önce bahsedilen sınıflandırma yöntemleri Meme Kanseri Coimbra Veri Kümesine uygulanırdır. Kod uygulaması ve tüm deneyler R-projesi (R 3.6.3) kullanılarak gerçekleştirilir. Bu tez çalışmasında NB, KNN, SVM ve LR sınıflandırma algoritmaları sonuçları karşılaştırılmıştır. Eğitim veri seti bu çalışmanın girdisi olacaktır. Görünmeyen verileri sınıflandırabilen ve tahmin edebilen en iyi eğitimli ve en doğru modeli üretmeyi umuyoruz. Aşağıdaki bölümlerde, yukarıdaki bahsedilen sınıflandırma yöntemlerinin performansı gösterilecektir.

3.3.1. NB algoritmasının uygulanması

NB algoritması, belirli verileri kullanarak insan vücudunun iyi huylu bir hücresi veya kötü huylu bir hücresi olma olasılığını hesaplar. Bu çalışmada NB sınıflandırıcısını R-projesinde Gauss dağılımı yöntemini kullanarak çalışıyoruz. Naive Bayes algoritmasının değerleri doğruluk % 79, Hassasiyet % 75, Özgüllük % 83 ve AUC % 79 olduğunu gösterir. (Çizelge 3.2)

Çizelge 3.2 NB Öznitelikleri ve değerleri

Yöntemler	Hassasiyet	Özgüllük	AUC	Doğruluk
NB A1-A4 (selected)	0.77	0.75	0.79	0.79

Daha önce bahsedildiği gibi, sınıflandırma algoritmalarını eğitmek için 92 örnek ve test etmek için 24 örnek kullanmıştır.

İki olası sınıf vardır: “1” ve “2”. Bu durumda “1” sağlıklı kontroller ve “2” hasta anlamına gelir.

Bu 24 durumdan Çizelge 3.3'e göre, NB algoritması Sağlıklı kontroludan 10 kez ve hastadan 9 kez öngörmüştür. NB karışıklık matrisinin bir sonucu olarak, 2 kişi hasta olarak teşhis edilen, ancak gerçekte hasta değildir. Ayrıca 3 kişi sağlıklı

kontrol olarak teşhis edilen, ama gerçekte hastalar edildi. 24 vakadan, algoritması 5 vaka yanlış ve 19 vaka doğru olarak tahmin etmiştir.

Çizelge 3.3 NB algoritması karışıklık matrisi

	Positive 1	**Negative 2**
Positive 1	10	2
Negative 2	3	9

3.3.2. KNN algoritmasının uygulanması

Veri kümesinde KNN algoritması kullanırken doğru k değerine karar vermek çok önemlidir. KNN algoritması temel olarak tüm veri noktaları arasındaki mesafeyi hesaplar, burada en iyi doğruluğu bulmak için doğru K değerini bulmakla lazımdır. Bu çalışmada k = 7 diğer k değerlerinden daha yüksek doğruluk verir ve bu nedenle minimum hata oranı verirdir. Çizelge 3.4'e göre, k = 7 değerleri doğruluk % 87.5, Hassasiyet % 84.62, Özgüllük % 91 ve AUC % 87'dır.

Çizelge 3.4 KNN Öznitelikleri ve değerleri

Yöntemler	**Hassasiyet**	**Özgüllük**	**AUC**	**Doğruluk**
KNN A1-A4 (selected) k=7	0.8462	0.9091	0.87	0.875

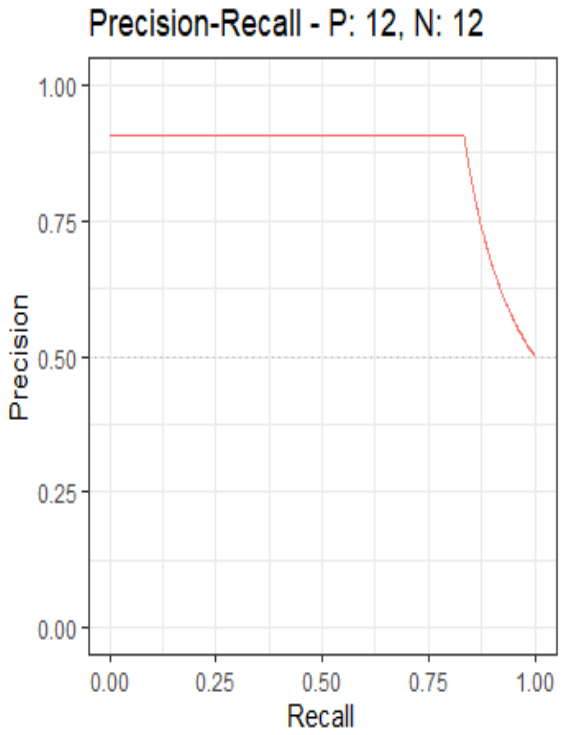
Çizelge 3.5'te gösterildiği gibi Meme Kanseri Coimbra Veri Kümesi üzerindeki KNN sınıflandırma modelinin performansını görmek için bir karışıklık matrisi oluşturulmuştur.

Çizegle 3.5 KNN algoritması karışıklık matrisi

	Positive 1	**Negative 2**
Positive 1	11	1
Negative 2	2	10

İki olası sınıf vardır: “2” ve “1”. Bu durumda “2” hasta ve “1” sağlıklı kontroluler anlamsına gelir. Bu 24 durumdan Çizegle 3.5'e göre, algoritma “1” 11 kez ve “2” 10 kez öngörmüştür. Gerçekte, 11 Sağlıklı kontrol ve 10 hasta tespit edildi. KNN karışıklık matrisinin bir sonucu olarak, 1 durum hasta olarak teşhis edildi ancak gerçekte sağlıklı kontrol edildi. Ayrıca 2 durum sağlıklı kontrol olarak teşhis edildi, ama gerçekte hastalar edildi. 24 vakadan, algoritması 21 vaka doğru olarak tahmin etmiştir. Ayrıca 3 vaka yanlıştır.

K değerini ve Hassasiyet-Geri Çağırma'ya Şekil 3.2'de göstermektedir. Kısaca, k = 7 olduğunda KNN algoritması en iyi performansa sahiptir.



Şekil 3.2 KNN modeliyle Hassasiyet-Geri Çağırma

3.3.3. RF algoritmasının uygulanması

Tahminci parametresini ormandaki ağaç sayısı olarak kullandık. Bölünmüş kalite fonksiyonunu ölçmek için kriter parametresi kullanılır. Eğitilen modellerin performansını değerlendirmek için Hassasiyet, Özgüllük, AUC, Doğruluk ve karmaşa matrisi performansleri ölçütü kullanılır. RF algoritması karışıklık matrisi (Çizelge 36'da) göstermektedir.

Çizelge 3.6 RF algoritması karışıklık matrisi

	Positive 1	**Negative 2**
Positive 1	9	3
Negative 2	2	10

RF, Çizelge 3.7'de gösterildiği gibi % 79 doğruluk oranına, % 82 Hassasiyet, % 83 Özgüllüğe ve % 77 AUC'ye sahiptir.

Çizelge 3.7 RF Öznitelikleri ve değerleri

Yöntemler	Hassasiyet	Özgüllük	AUC	Doğruluk
RF A1-A4 (selected)	0.82	0.83	0.77	0.79

3.3.4. SVM algoritmasının uygulanması

SVM, doğrusal olmayan sınıflandırma problemlerini çözmek için kullanılır. SVM, marj optimizasyonu ve çekirdek(kernel) gösterimi gibi iki önemli özelliğe sahiptir. Çekirdek (kernel) işlevi, doğrusal olmayan verileri doğrusal verilere dönüştürmek için kullanılan bir numaradır. Daha sonra, iki sınıf arasında maksimum ayırma marjı ile hiper-planı bulma sorununu çözebiliriz.

SVM algoritmasında Coimbra veri setini iki farklı sınıfa ayırmak için. SVM, yüksek boyutlu bir uzayda en büyük marjı olan, $w \cdot x + b = 0$ hiper düzlemi oluşturur. B önyargıdır, w hiper düzleme normaldir ve iki sınıf arasındaki kenar boşluğu en yakın veri noktaları arasındaki en uzun mesafedir. İyi huylu negatif sınıf için ($w \cdot x + b = -1$) ve kötü huylu pozitif sınıf için ($w \cdot x + b = 1$) olacaktır. Bu algoritmanın temel amacı, iki sınıfı ayırmak ve doğru sınıflandırıcıyı bulmaktır. Doğru sınıflandırıcı, görünmeyen örneklerde işe yarardır. İki sınıfı ayırmak için bir çekirdek(kernel) işlevi kullanmıştır.

Doğrusal olmayan sınıflandırmayı doğrusal sınıflandırmaya dönüştürmek için üç işlev vardır, bunlar radyal temel işlev, polinom işlevi ve sigmoiddir.

Veri setimiz doğrusal olmadığı için. Bu çalışmamızda, Gaussian RBF denklemini izleyen radyal temel fonksiyonu (RBF) kullandık.

RBF çekirdek karışma matrisi, kanser hastası olmayan 2 hastanın kanser hastası olarak tahmin edildiğini ve malign hücreye sahip hastaların 2'sinin kanser dışı hasta olarak tahmin edildiğini göstermektedir. 10 kişi Kanser dışı hasta olarak doğru tahmin edilen ve 10 kişi Kanser hastası olarak doğru tahmin edildi. (Çizelge 3.8)

Çizelge 3.8 SVM algoritması karışıklık matrisi

	Positive 1	**Negative 2**
Positive 1	10	2
Negative 2	2	10

SVM, Çizelge 3.9'de gösterildiği gibi %83,3 doğruluk oranına, %83 Hassasiyet, %83 Özgüllüğe ve %82 AUC'ye sahiptir.

Çizelge 3.9 SVM Öznitelikleri ve değerleri

Yöntemler	**Hassasiyet**	**Özgüllük**	**AUC**	**Doğruluk**
SVM A1-A4 (selected)	0.83	0.83	0.82	0.83.3

3.3.5. LR algoritmasının uygulanması

Eğitilen modellerin performansını değerlendirmek için Hassasiyet, Özgüllük, AUC, Doğruluk ve karmaşa matrisi olmak üzere performansler ölçütü kullanılır. Çizelge 3.10, LR modeli için karışıklık matrisini göstermektedir.

RF algoritmasının karışma matrisi göre, kanser hastası olmayan 2 hastanın kanser hastası olarak tahmin edildi. Malign hücreye sahip hastaların 3'sinin kanser dışı hasta olarak tahmin edildiğini göstermektedir. 10 kişi Kanser dışı hasta olarak doğru tahmin edildi. 9 kişi Kanser hastası olarak doğru tahmin edildi. (Çizelge 3.10)

Çizelge 3.10 LR Öznitelikleri ve değerleri

	Positive 1	**Negative 2**
Positive 1	10	2
Negative 2	4	8

LR, Çizelge 3.11'te gösterildiği gibi % 75 doğruluk oranına, % 71 Hassasiyet, %66 Özgüllüğe ve % 74 AUC'ye sahiptir.

Çizelge 3.11 LR Öznitelikleri ve değerleri

Yöntemler	Hassasiyet	Özgüllük	AUC	Doğruluk
LR A1-A4 (selected)	0.71	0.66	0.74	0.75

3.4. Sınıflandırma Yöntemlerinin Sonuçları

Bu tezin temel amacı, Coimbra veri setine uygulanan sınıflandırma algoritmalarının performansını karşılaştırmaktır.

KNN algoritması, SVM, LR, NB ve RF algoritmalarına kıyasla en iyi doğruluk oranına sahiptir. KNN algoritması % 87.5 doğruluk oranına ulaştı. SVM algoritması % 83.3 doğruluk oranına sahiptir ve bu çalışmada ikinci sırada yer almaktadır. Ayrıca hem RF algoritması hem de NB algoritması % 79 doğruluk oranına sahiptir. Sonunda LR algoritması % 75 doğruluk oranına sahiptir. LR algoritması, bu çalışmada uygulanan diğer algoritmalarla karşılaştırıldığında en düşük doğruluk oranına sahiptir.

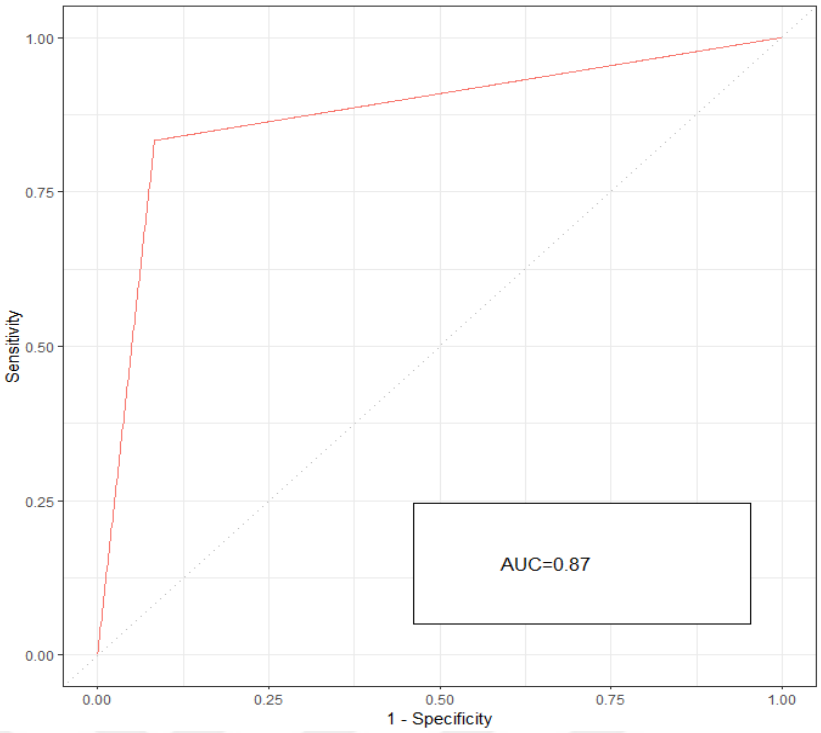
Çizelge 3.12, Coimbra Meme Kanseri veri setinde farklı sınıflandırma yöntemlerinin uygulanmasından elde edilen yukarıdaki sonuçları göstermektedir.

Çizelge 3.12 Sınıflandırma yöntemlerinin sonuçları

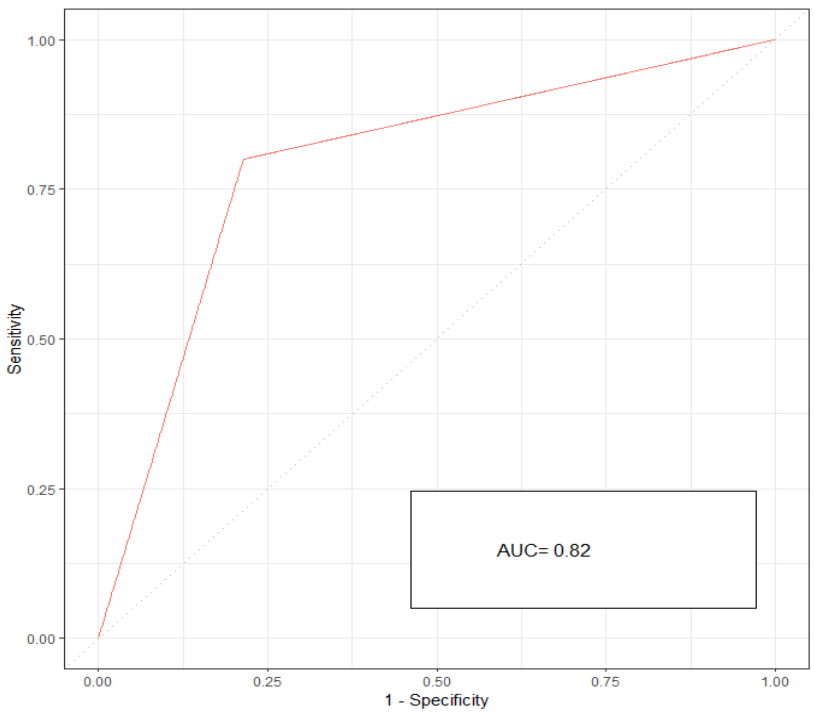
Algoritma	Doğruluk
KNN	% 87.5
SVM	% 83.3
NB	% 79
RF	% 79
LR	% 75

3.5. AUC Ölçüm Sonuçları

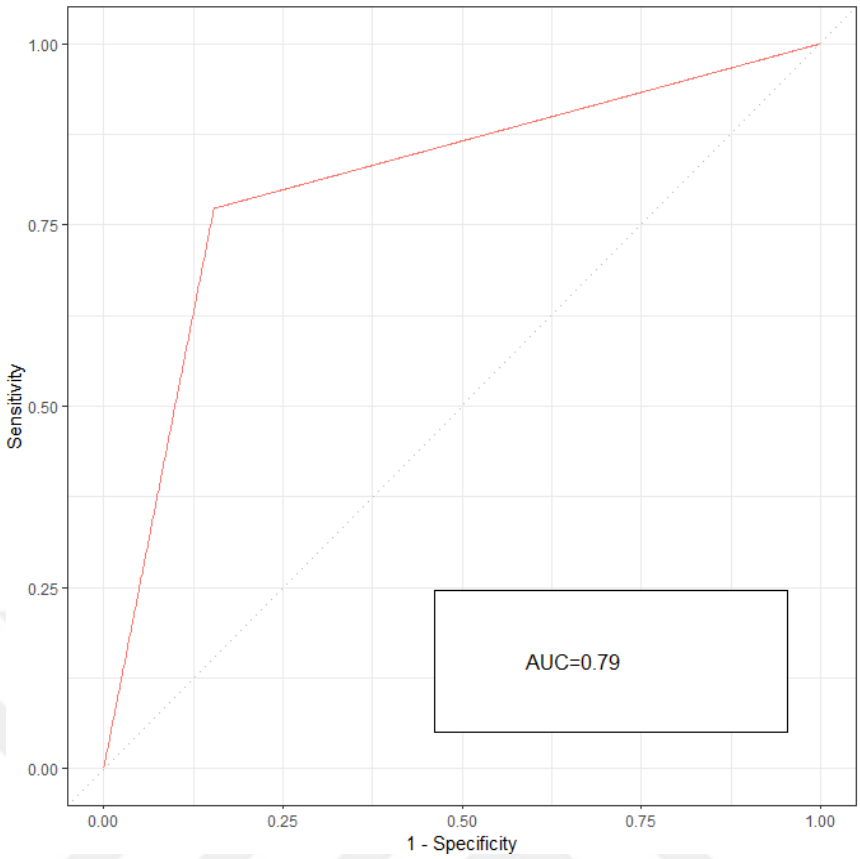
Bu çalışmada, KNN algoritması, %87 en iyi AUC ölçüsü sonuçları sahiptir (Şekil 3.3). İkinci sırada, SVM algoritması, %82 AUC ölçüsü sonuçlar sahiptir (Şekil 3.4). Ayrıca, NB algoritması %79 AUC ölçüsü sahiptir (Şekil 3.5). Ve RF algoritması %77 AUC ölçüsü sahiptir (Şekil 3.6). Sonunda LR algoritması %75 AUC ölçüsü sahiptir (Şekil 3.7).



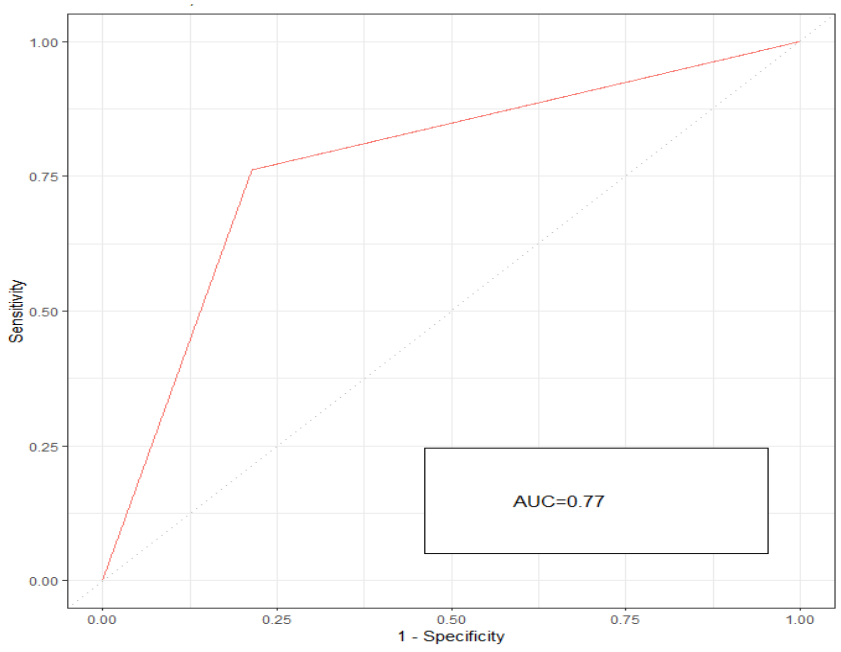
Şekil 3.3 KNN için AUC ölçümü



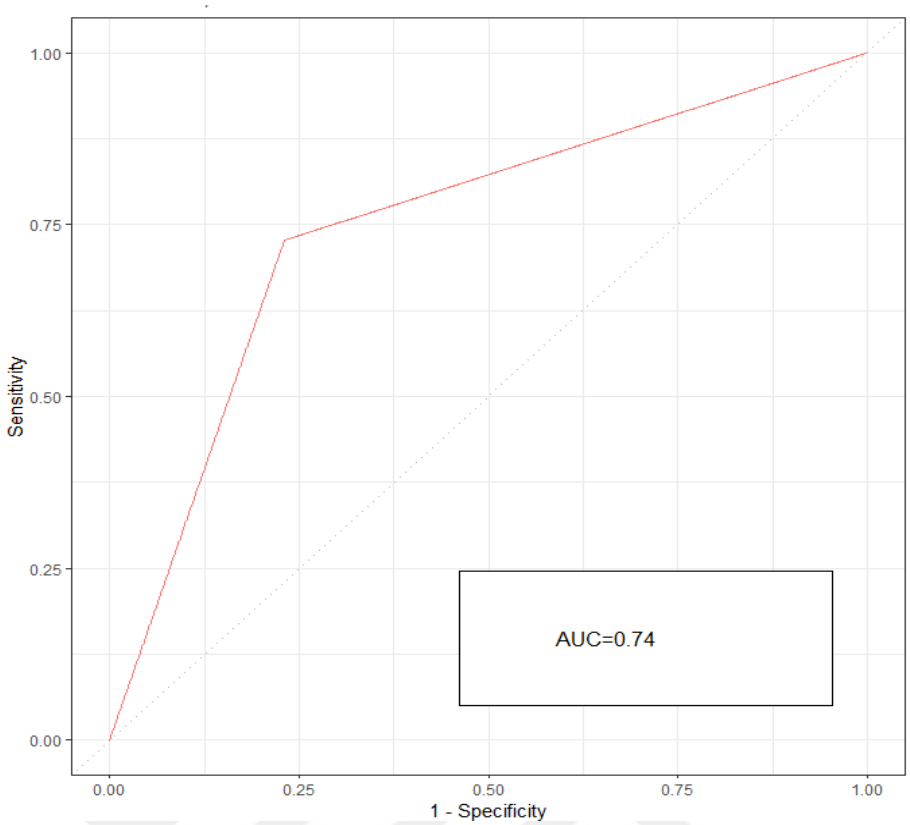
Şekil 3.4 SVM için AUC ölçümü



Şekil 3.5 NB için AUC ölçümü



Şekil 3.6 RF için AUC ölçümü



Şekil 3.7 LR için AUC ölçümü

4. SONUÇ VE ÖNERİLER

Meme kanseri, kanser türlerinde ikinci en yüksek ölüm oranına sahiptir ve bu hastalık çoğunlukla kadınları etkiler. Akciğer kanserinin dünyadaki en yüksek ölüm oranına sahip olduğu bir hastadır. Pek çok araştırmacı araştırmalarında, göğüs kanserinin tahmini için yapay zekâ ve makine öğrenimi tekniklerini kullandı.

Makine öğrenimi ve yapay zekâ tekniklerinin, verileri girdi olarak alır ve aldığı verilerden öğrenir. Bu teknikler, aynı özelliklere sahip herhangi bir yeni veri üzerinde tahminlerde bulunabilecektir. Bu tezde veri madenciliği ve makine öğrenmesi kavramları, yapay zekâ ile ilişkilerinden bahsedilmiştir. Ayrıca meme kanseri hakkında ortak bilgiler ve bu hastalığın dünya çapındaki riskiyle ilgili anket sonuçları verildi.

4.1. Sonuç

Bu araştırmada, çeşitli makine öğrenme teknikleri, bunlar KNN, SVM, RF, NB ve LR kullandık. Bu yöntemlerin her biri, ilk dört biyobelirteç olan Yaş, Glikoz, Resistin ve BMI'ya göre gerçekleştirildi. R-projesi (R 3.6.3) kullanılarak meme kanseri veri seti kullanılarak uygulanmıştır. Bu çalışmada kullanılan veriler, UCI makine öğrenimi havuzundan alınan Coimbra meme kanseri veri setidir. Verinin 10 özniteliği ve 116 Örneği vardır. Veriler, test seti (verilerin%20'si) ve eğitim seti (verilerin%80'i) olmak üzere iki gruba ayrılmıştır. Eğitim seti, sınıflandırma algoritmalarını eğitmek için kullanılır. Bu adımdan sonra, test seti eğitilen algoritmaları test etmek için kullanılır. NB sınıflandırıcısını R-projesinde Gauss dağılımı yöntemini kullanarak çalıştık. SVM uygulamasında, Radial Basis Kernel yöntemini seçildik. KNN uygulamasında, Öklid mesafesi kullandık. Bir karışıklık matrisi yardımıyla, yanlış ve doğru tahminlerin sayısı geldi. Ayrıca, her yöntemin doğruluk oranı, Duyarlılık oranı ve Özgünlük oranı hesaplanmış ve karşılaştırılmıştır. En yakın yedi komşuya sahip KNN sınıflandırıcı, diğer sınıflandırıcılara kıyasla en yüksek sonuç oranını verdi. İyi sonuçlara sahip olmanın ile, KNN, SVM, LR, RF ve Naïve Bayes metodolojilerinin geliştirilmesi

nispeten daha kolaydır. Bu sonuçlar iyi huylu veya kötü huylu vakaların varlığının değerlendirilmesinde kullanılırsa, böyle bir yöntem doktorların meme kanserinin erken teşhisi için daha iyi teşhis koymasına yardımcı olacaktır.

4.2. Öneriler

Algoritmaların performansını daha iyi görmek için, veri kümesini artan boyutta genişletmeyi öneriyoruz.

Farklı meme kanseri veri setlerinde aynı tür algoritmaları kullanmayı öneriyoruz.

KAYNAKLAR

Abedin, T., Chowdhury, Z., Afzal, A. R., Yeasmin, F., Turin, T. C., 2016. Review Article Application of Binary Logistic Regression in Clinical Research. Journal of National Heart Foundation of Bangladesh, 5, 8-11.

Abu Alfeilat, H.A., Hassanat, A.B.A., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., Eyal Salman, H.S., Prasath, V.B.S., 2019. Effects of distance measure choice on K-nearest neighbor classifier performance: A review. Big Data 7:4, 221–248. Erişim Tarihi: 16.12.2019. https://doi.org/10.1089/big.2018.0175

Aggarwal, C. C., 2015. Data Mining: The Textbook. Springer, New York.

Akar, Ö., Güngör, O., 2012. Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması. Jeodezi ve Jeoinformasyon Dergisi, Trabzon(2), 139-146.

Albayrak, A. S., 2006. Uygulamalı Çok Değişkenli İstatistik Teknikler. Asil Yayın Dağıtım, Ankara.

Al-Muhaideb, S., El Bachir Menai, M., 2013. Hybrid Metaheuristics for Medical Data Classification. In: Hybrid Metaheuristics. Berlin, Springer, 187-217.

Annasaheb, A. B., Verma, V. K., 2016. Data Mining Classification Techniques: A Recent Survey. International Journal of Emerging Technologies in Engineering Research (IJETER), 4(8), 51-54.

Anon., 2018. Breast Cancer Coimbra Data Set. Erişim Tarihi: 2018. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Araújo, V. J., Guimarães, A. J., Souza, P. V., Rezende, T. S., Araújo, V. S., 2019. Using Resistin, Glucose, Age and BMI and Pruning Fuzzy Neural Network for the Construction of Expert Systems in the Prediction of Breast Cancer. Machine Learning and Knowledge Extraction, 466-482.

Artsın, M., 2019. Veri madenciliği ve bilgi keşfi. Açıköğretim Uygulamaları ve Araştırmaları Dergisi, 5 (3), 174-180.

Aslan, M. F., Celik, Y., Sabanci, K., Durdu, A., 2018. Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. International Journal of Intelligent Systems and Applications in Engineering, 289-293.

Balaban, M. E., Kartal, E., 2015. Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları. Çağlayan Kitabev, İstanbul.

Baratloo, A., Hosseini, M., Negida, A., El Ashal, G., 2015. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. Pervasive and Mobile Computing PMC, 48–49.

Barga, R., Fontama, V., Tok, W.H., 2015. Introduction to Statistical and Machine Learning Algorithms. In: Predictive Analytics with Microsoft Azure Machine Learning. Apress, Berkeley, CA, 133-148.

Berson, A., J Smith, S., 1997. Data Warehousing, Data Mining, and Olap., McGraw-Hill, New York.

Bilbao, I., Bilbao, J., 2017. Overfitting problem and the over-training in the era of data: Particularly for Artificial Neural Networks. Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), 173-177.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. A Cancer Journal for Clinicians, 68(6), 394–424.

Breiman, L., 2001. Random Forests. Machine Learning. 45, 5–32. Erişim Tarihi: 03.07.2001. https://ezproxy.ticaret.edu.tr:2885/10.1023/A:1010933404324

Çokluk, Ö., 2010. Lojistik regresyon Analizi: Kavram ve Uygulama. 10(3), 1357-1407.

Dangare, C. S., Apte, S. S., 2012. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications, 47(10), 44-48.

Englund, C., Verikas, A., 2012. A novel approach to estimate proximity in a random forest: An exploratory study. Expert Systems with Applications, 39, 13046–13050.

Ertel, W., 2017. First-order Predicate Logic. Introduction to Artificial Intelligence, Springer, Cham, 1-21.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. Association for the Advancement of Artificial Intelligence, 17(3), 37-54.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., Bray, F., 2019. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. International Journal of Cancer, 1941-1953.

Ghosh, S., Dasgupta, A., Swetapadma, A., 2019. A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification. International Conference on Intelligent Sustainable Systems (ICISS), 24-28.

Griffis, J. C., Allendorfer, J. B., Szaflarski, J. P., 2015. Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. Journal of Neuroscience Methods, 257, 97-108.

Gültepe, Y., Kartbayev, T., 2019. A Study of Data Mining Methods for Breast Cancer Prediction. Proceedings Book, 303-305.

Han, J., Kamber, M., Pei, J., 2011. In: Data Mining: Concepts and Techniques, 3rd ed. California, Morgan Kaufmann.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., Xu, W., 2018. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 41-51.

Iaizzo, P. A., 2019. The Ability to Predict Seizure Onset. In: Engineering in Medicine. 365-378.

Imaginis, 2019. American Cancer Society. Erişim Tarihi: 15.03.2019. http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-2

Inflammatory Breast Cancer., 2016. National Cancer Institute. Erişim Tarihi: 06.01.2016. https://www.cancer.gov/types/breast/mammograms-fact-sheet#q5

Jahromi, A.H., Taheri, M., 2017. A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. Artificial Intelligence and Signal Processing Conference (AISP), 209-212.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Tree-Based Methods. In: S. T. i. Statistics, ed. In: An Introduction to Statistical Learning. Springer, New York.

Jantan, H., Hamdan, A. R., Othman, Z. A., 2009. Potential Data Mining Classification Techniques for Academic Talent Forecasting. 2009 Ninth International Conference on Intelligent Systems Design and Applications, 1173-1178.

Jiang, Q., Wang, W., Han, X., Zhang, S., Wang, X., Wang, C., 2016. Deep feature weighting in Naive Bayes for Chinese text classification. 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), 160-164. Beijing.

Karadağ, K., 2013. Ecog Tabanlı Parmak Hareketlerinin KNN ve DVM Yöntemleri ile Sınıflandırılması Dicle Üniversitesi, Elektrik Elektronik Fakültesi, Yüksek Lisans Tezi, Diyarbakır.

Kotsiantis, S. B., 2007. Supervised Machine Learning: A Review of Classification Techniques. 31, 249-268.

Kumar, K., Singh, V. V., Ramaswamy, R., 2020. Different Perspective of Machine LearningTechnique to Better Predict Breast Cancer Survival. Erişim Tarihi: 05.07.2020. https://doi.org/10.1101/2020.07.03.186890

Kumar, R., Verma, R., 2012. Classification Algorithms for Data Mining: A Survey. International Journal of Innovations in Engineering and Technology (IJIET), 1, 7-14.

Leman, D., 2018. Expert System Diagnose Tuberculosis Using Bayes Theorem Method and Shafer Dempster Method. 6th International Conference on Cyber and IT Service Management (CITSM), 1-4.

Levy, M., Raviv, D., Baker, J., 2019. Data Center Predictions using MATLAB Machine Learning Toolbox. IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 0458-0464.

Ling, C. X., Huang, J., Zhang, H., 2003. AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. International Joint Conferences on Artificial Intelligence, 519–524.

Maimon, O., Rokach, L., 2005. Introduction to Knowledge Discovery in Databases. Data Mining and Knowledge Discovery Handbook. 1-17.

Mavani, U., Lobo, V. B., Pednekar, A., Pereira, N. C., Mishra, R., Ansari, N., 2019. Naïve Bayes Classification on Student Placement Data: A Comparative Study of Data Mining Tools. Information and Communication Technology for Sustainable Development. Springer, 363-372.

Memorial Sloan Kettering Cancer Center, 2019. Stages of Breast Cancer. Erişim Tarihi: 10.05.2019. https://www.mskcc.org/cancer-care/types/breast/diagnosis/stages-breast

Mitchell, T. M., 1997. Does Machine Learning Really Work. AI Magazine, 18(3), 11-20. Erişim Tarihi: 15.09.1997. https://doi.org/10.1609/aimag.v18i3.1303

Mueller, J., Massaron, L., 2016. Machine Learning For Dummies. Erişim Tarihi: 31.05.2016. https://books.google.com.tr/books?id=--ASEAAAQBAJ

Murty, M. N., Vishwanathan, S. V. N., 2002. SSVM : A Simple SVM Algorithm. Bangalore, 2393-2398.

Mutasa, S., Sun, S., Ha, R., 2020. Understanding artificial intelligence based radiology studies: What is overfitting?. Clinical Imaging, 96-99.

Organization, W. H., 2018. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. Erişim Tarihi: 12.09.2018. https://www.who.int/cancer/PRGlobocanFinal.pdf

Pandit, S., Gupta, S., 2011. A comparative study on distance measuring approaches for clustering. International Journal of Research in Computer Science, 2(1), 29-31.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., Caramelo, F., 2018. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BioMed Central Cancer. 18, 29. Erişim Tarihi: 04.01.2018. https://doi.org/10.1186/s12885-017-3877-1

Peng, W., Chen, J., Zhou, H., 2009. An Implementation of Decision Tree Learning Algorithm. 1-9.

Peng, Y., Kou, G., Chen, Z., 2018. A Descriptive Framework for the Field of Data Mining and Knowledge Discovery. International Journal of Information Technology & Decision Making, 7, 639-682.

Pham, H. N. A., Triantaphyllou, E., 2008. The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining. Soft Computing for Knowledge Discovery and Data Mining. Springer, 391-431.

Poorani, S., Balasubramanie, P., 2019. Deep Neural Network Classifier in Breast Cancer Prediction. International Journal of Engineering and Advanced Technology, 2106-2109.

Prevention, C. f. D. C. a., 2019. Basic Information About Breast Cancer. Erişim Tarihi: 26.07.2019. https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm

Purwaningsih, E., 2019. Application of the Support Vector Machine and Neural Network Model Based on Particle Swarm Optimization for Breast Cancer Prediction. SinkrOn, 66-73.

Qizhong, Z., 2007. Gene Selection and Classification Using Non-linear Kernel Support Vector Machines Based on Gene Expression Data. IEEE/ICME International Conference on Complex Medical Engineering, Beij, 1606-1611.

Ranganayaki, V., Deepa, S. N., 2019. Linear and non-linear proximal support vector machine classifiers for wind speed prediction. Cluster Comput 22, 379–390.

Sahu, H., Shrma, S., Gondhalakar, S., 2011. A Brief Overview on Data Mining Survey. International Journal of Computer Technology and Electronics Engineering (IJCTEE), 1(3), 114-121.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., less, K. M., 2016. Evaluating the visualization of what a Deep Neural Network has learned. IEEE transactions on neural networks and learning systems, 28, 2660-2673.

Singh, S. N., Thakral, S., 2018. Using Data Mining Tools for Breast Cancer. 4th International Conference on Computing Communication and Automation (ICCCA), 1-4.

Sleath, E., Brown, S., 2015. The treatment of missing data. %1 içindeResearch Methods for Forensic Psychologists. Journal of Interpersonal Violence, 30(15), 2726–2750.

Society, A. C., 2020. Breast Cancer Facts & Figures 2019-2020. Atlanta, American Cancer Society. Erişim Tarihi: 05.01.2020. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

Soliman, O. S., AboElHamd, E., 2014. Classification of Breast Cancer using Differential Evolution and LeastSquares Support Vector Machine. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3(2), 155–161.

Srivastava, D., Bhambhu, L., 2010. Data classification using support vector machine. Journal of Theoretical and Applied Information Technology, 1-7.

Suthaharan, S., 2015. Machine Learning Models and Algorithms for Big Data Classification. Springer, 36.

Tekieh, M. H., Raahemi, B., 2015. Importance of data mining in healthcare: A survey. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 1057-1062.

Turing, A., 1950. Computing Machinery and Intelligence. Mind, LIX (236), 433–460. Erişim Tarihi: 01.10.1950. https://doi.org/10.1093/mind/LIX.236.433

Vapnik, V. N., 2000. The Nature of Statistical Learning Theory. Springer, New York.

Vieira, J., Duarte, R. P., Neto, H. C., 2019. kNN-STUFF: kNN STreaming Unit for Fpgas. Institute of Electrical and Electronics Engineers IEEE, 7, 170864-170877.

Villejoubert, G., Mandel, D. R., 2002. The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. Memory & Cognition, 30(2), 171-178.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., 2017. Data Mining Practical Machine Learning Tools and Techniques. Elsevier Science & Technology, San Francisco.

Xing, W., Bei, Y., 2020. Medical Health Big Data Classification Based on KNN Classification Algorithm. Institute of Electrical and Electronics Engineers IEEE, 8, 28808-28819.

Yılmaz, H., 2014. Random Forests yönteminde kayıp veri probleminin incelenmesi ve sağlık alanında bir uygulama. Eskişehir Osmangazi Üniversitesi, Yüksek Lisans Tezi, Eskişehir.

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning Methods and Applications. Erişim Tarihi: 30.07.2012. https://link.springer.com/book/10.1007%2F978-1-4419-9326-7#reviews

Zhang, H., 2004. The Optimality of Naive Bayes. American Association for Artificial Intelligence, 1-6. New Brunswick.

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 1-7.

Zhu, W., Zeng, N., Wang, N., 2010. Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS®Implementations. Health Care and Life Sciences, 1-9.

EKLER

```
rm(list = ls())

library(ggplot2)

library(e1071)

library(dplyr)

library(pROC)

library(doBy)

library(ResourceSelection)

library(xlsx)

library(randomForest)

library(pscl)


data2<- read.csv("D:/Belgelerim/Desktop/R-data/data2.csv")

head(data2)

table(data2$Classification)

Agedata=summaryBy(Age ~ Classification, data=data2, FUN=c(median,IQR))

Agedata=cbind(Agedata[1, 2:3], Agedata[2, 2:3], wilcox.test(Age ~ Classification, data=data2)[3])

colnames(Agedata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

BMIdata=summaryBy(BMI ~ Classification, data=data2, FUN=c(mean,sd))

BMIdata=cbind(BMIdata[1, 2:3], BMIdata[2, 2:3], wilcox.test(BMI ~ Classification, data=data2)[3])

colnames(BMIdata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Glucosedata=summaryBy(Glucose ~ Classification, data=data2, FUN=c(mean,sd))
```

```
Glucosedata=cbind(Glucosedata[1, 2:3], Glucosedata[2, 2:3], wilcox.test(Glucose ~ Classification, data=data2)[3])

colnames(Glucosedata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Insulindata=summaryBy(Insulin ~ Classification, data=data2, FUN=c(mean,sd))

Insulindata=cbind(Insulindata[1, 2:3], Insulindata[2, 2:3], wilcox.test(Insulin ~ Classification, data=data2)[3])

colnames(Insulindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

HOMAdata=summaryBy(HOMA ~ Classification, data=data2, FUN=c(mean,sd))

HOMAdata=cbind(HOMAdata[1, 2:3], HOMAdata[2, 2:3], wilcox.test(HOMA ~ Classification, data=data2)[3])


colnames(HOMAdata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Leptindata=summaryBy(Leptin ~ Classification, data=data2, FUN=c(mean,sd))

Leptindata=cbind(Leptindata[1, 2:3], Leptindata[2, 2:3], wilcox.test(Leptin ~ Classification, data=data2)[3])

colnames(Leptindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

Adiponectindata=summaryBy(Adiponectin ~ Classification, data=data2, FUN=c(mean,sd))

Adiponectindata=cbind(Adiponectindata[1, 2:3], Adiponectindata[2, 2:3], wilcox.test(Adiponectin ~ Classification, data=data2)[3])

colnames(Adiponectindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")


Resistindata=summaryBy(Resistin ~ Classification, data=data2, FUN=c(mean,sd))

Resistindata=cbind(Resistindata[1, 2:3], Resistindata[2, 2:3], wilcox.test(Resistin ~ Classification, data=data2)[3])
```

```
colnames(Resistindata)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

MCP.1data=summaryBy(MCP.1 ~ Classification, data=data2, FUN=c(mean,sd))

MCP.1data=cbind(MCP.1data[1, 2:3], MCP.1data[2, 2:3], wilcox.test(MCP.1 ~ Classification, data=data2)[3])

colnames(MCP.1data)=c("Control.median", "Control.IQR", "Patient.median", "Patient.IQR", "p-value")

summarytable=rbind(Agedata, BMIdata, Glucosedata, Insulindata, HOMAdata, Leptindata, Adiponectindata, Resistindata, MCP.1data)


rownames(summarytable)=c("Age", "BMI", "Glucose", "Insulin", "HOMA", "Leptin", "Adiponectin", "Resistin", "MCP.1")

summarytable[, 1:4]=round(summarytable[, 1:4], digits=1)

summarytable[, 5]=round(summarytable[, 5], digits=4)

summarytable

# convert data to numeric

data_numeric<- function(x){ (as.numeric( as.character(x)))}

data2.numeric <-  as.data.frame(lapply(data2, data_numeric))

data_norm<- function(x){ ((x-min(x))/(max(x)-min(x)) )}

data2.norm <-  as.data.frame(lapply(data2.numeric, data_norm))

levels(data2.norm$Classification)=c(1, 2)

glmmodel=glm(Classification~.,data=data2.norm,family='binomial')

summary(glmmodel)


glmtable=cbind(coef(summary(glmmodel)), exp(coef(summary(glmmodel))[, 1]), exp(confint(glmmodel)))

colnames(glmtable)[5]="Odds ratios"

round(glmtable, digits=3)
```

```
pR2(glmmodel)[4]
hoslem.test(as.numeric(data2$Classification)-1, fitted(glmmodel), g=10)

splitter <- function(datatosplit, trainsizeinpercentage) {

 CAdatatosplit=datatosplit[datatosplit$Classification==2,]
 COdatatosplit=datatosplit[datatosplit$Classification==1,]

 n=nrow(CAdatatosplit)
 ntrain = round(n*trainsizeinpercentage)
 tindex = sample(n,ntrain)

 CAtrainset = CAdatatosplit[tindex,]
 CAtestset = CAdatatosplit[-tindex,]
 n=nrow(COdatatosplit)
 ntrain = round(n*trainsizeinpercentage)
 tindex = sample(n,ntrain)

 COtrainset = COdatatosplit[tindex,]
 COtestset = COdatatosplit[-tindex,]

 trainset=rbind(CAtrainset, COtrainset)
 testset=rbind(COtestset, CAtestset)

 return(list(trainset, testset))
}
```

```
evaluateclassifier <- function(pred, testset) {
 real.diagnosis=testset$Classification
 roccurve=roc(real.diagnosis, pred)
 AUCclassification=auc(roccurve)

threshold=roccurve$threshold[which.max(roccurve$sensitivities+roccurve$specificities)]

 binary.classification=as.numeric(pred>threshold)+1

 CLAtable=table(real.diagnosis, binary.classification)
 a=CLAtable[1, 1]
 b=CLAtable[1, 2]
 c=CLAtable[2, 1]
 d=CLAtable[2, 2]

 accuracy=(d+a)/sum(CLAtable)

specificity=roccurve$specificities[which.max(roccurve$sensitivities+roccurve$specificities)]

sensitivity=roccurve$sensitivities[which.max(roccurve$sensitivities+roccurve$specificities)]

 return(list(AUCclassification, accuracy, specificity, sensitivity))
}

LRclass <- function(trainset, testset) {
```

```
 LRfit = glm(Classification~.,data=trainset,family='binomial')

 cancer.probability = predict.glm(LRfit,testset[c(-ncol(testset))], type = 'response')

 return(cancer.probability)
}
RFclass <- function(trainset, testset) {
 RFfit = randomForest(Classification ~ ., data=trainset)

 cancer.probability = predict(RFfit,testset[c(-ncol(testset))], type = "vote")

 cancer.probability=cancer.probability[, 2]

 return(cancer.probability)
}
SVMclass <- function(trainset, testset) {
 SVMfit=svm(Classification~.,data=trainset, kernel="radial", probability = TRUE)

 cancer.probability = predict(SVMfit,testset[c(-ncol(testset))], probability=TRUE)

 cancer.probability=attr(cancer.probability, "prob")[,1]
 return(cancer.probability)
}
```

```
KNNclass <- function(trainset, testset) {
 KNNfit =knn(Classification~.,data=trainset, k=7, prob = TRUE)

 cancer.probability = predict(KNNfit,testset[c(-ncol(testset))], prob = TRUE)
 cancer.probability=attr(cancer.probability, "probp")[,3]

 return(cancer.probability)
}

NBclass <- function(trainset, testset) {
 NBfit =naiveBayes(Classification~.,data=trainset)

 cancer.probability = predict(NBfit,testset[c(-ncol(testset))], type = "class")
 cancer.probability=attr(cancer.probability, "cla")[,4]

 return(cancer.probability)
}

RFalg <- randomForest(Classification ~ ., data = data2)
importance.table=varImpPlot(RFalg, type=2)
importance.table
rownames(importance.table)[order(importance.table, decreasing = TRUE)]

real.diagnosis=data2$Classification
pred=data2$Age
results=evaluateclassifier(pred, data2)
computed_AUC=results[[1]]
```

```
computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Age=c(computed_AUC, computed_sensitivity, computed_specificity)

pred=data2$BMI

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_BMI=c(computed_AUC, computed_sensitivity, computed_specificity)


pred=data2$Glucose

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Glucose=c(computed_AUC, computed_sensitivity,
computed_specificity)


pred=data2$Insulin

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Insulin=c(computed_AUC, computed_sensitivity, computed_specificity)


pred=data2$HOMA

results=evaluateclassifier(pred, data2)
```

```
computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_HOMA=c(computed_AUC, computed_sensitivity, computed_specificity)


pred=data2$Leptin

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Leptin=c(computed_AUC, computed_sensitivity, computed_specificity)


pred=data2$Adiponectin

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Adiponectin=c(computed_AUC,computed_sensitivity,
computed_specificity)


pred=data2$Resistin

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_Resistin=c(computed_AUC, computed_sensitivity,
computed_specificity)
```

```
pred=data2$MCP.1

results=evaluateclassifier(pred, data2)

computed_AUC=results[[1]]

computed_sensitivity=results[[4]]

computed_specificity=results[[3]]

classifier_MCP.1=c(computed_AUC, computed_sensitivity, computed_specificity)


univariate_diagnostic=rbind(classifier_Age, classifier_BMI, classifier_Glucose, classifier_Insulin,

                            classifier_HOMA, classifier_Leptin, classifier_Adiponectin, classifier_Resistin, classifier_MCP.1)


univariate_diagnostic=as.data.frame(univariate_diagnostic)

colnames(univariate_diagnostic)=c('AUC', 'sensitivity', 'specificity')

univariate_diagnostic


data2_subset <- dplyr::select(data2, Glucose, BMI, Age, Resistin, Classification)


number.of.iterations=400


KNN_AUC=matrix(NA, number.of.iterations, 1)

KNN_accuracy=matrix(NA, number.of.iterations, 1)

KNN_specificity=matrix(NA, number.of.iterations, 1)

KNN_sensitivity=matrix(NA, number.of.iterations, 1)


SVM_AUC=matrix(NA, number.of.iterations, 1)
```

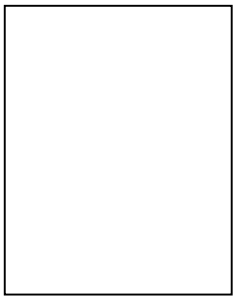
```
SVM_accuracy=matrix(NA, number.of.iterations, 1)
SVM_specificity=matrix(NA, number.of.iterations, 1)
SVM_sensitivity=matrix(NA, number.of.iterations, 1)

RF_AUC=matrix(NA, number.of.iterations, 1)
RF_accuracy=matrix(NA, number.of.iterations, 1)
RF_specificity=matrix(NA, number.of.iterations, 1)
RF_sensitivity=matrix(NA, number.of.iterations, 1)
NB_AUC=matrix(NA, number.of.iterations, 1)
NB_accuracy=matrix(NA, number.of.iterations, 1)
NB_specificity=matrix(NA, number.of.iterations, 1)
NB_sensitivity=matrix(NA, number.of.iterations, 1)
LR_AUC=matrix(NA, number.of.iterations, 1)
LR_accuracy=matrix(NA, number.of.iterations, 1)
LR_specificity=matrix(NA, number.of.iterations, 1)
LR_sensitivity=matrix(NA, number.of.iterations, 1)
for (iterationindex in 1:number.of.iterations){
 set.seed(iterationindex)
 splitteddata=splitter(data2, 0.8)
 trainset=splitteddata[[1]]
 testset=splitteddata[[2]]
 KNN.pred=SVMclass(trainset, testset)
 SVM.pred=SVMclass(trainset, testset)
 RF.pred=RFclass(trainset, testset)
 NB.pred=SVMclass(trainset, testset)
 LR.pred=LRclass(trainset, testset)
 KNN.evaluation=evaluateclass(SVM.pred, testset)
```

```
SVM.evaluation=evaluateclass(SVM.pred, testset)
RF.evaluation=evaluateclass(RF.pred, testset)
NB.evaluation=evaluateclass(SVM.pred, testset)
LR.evaluation=evaluateclass(LR.pred, testset)
KNN_AUC[iterationindex]=SVM.evaluation[[1]]
KNN_accuracy[iterationindex]=SVM.evaluation[[2]]
KNN_specificity[iterationindex]=SVM.evaluation[[3]]
KNN_sensitivity[iterationindex]=SVM.evaluation[[4]]
SVM_AUC[iterationindex]=SVM.evaluation[[1]]
SVM_accuracy[iterationindex]=SVM.evaluation[[2]]
SVM_specificity[iterationindex]=SVM.evaluation[[3]]
SVM_sensitivity[iterationindex]=SVM.evaluation[[4]]
 RF_AUC[iterationindex]=RF.evaluation[[1]]
RF_accuracy[iterationindex]=RF.evaluation[[2]]
RF_specificity[iterationindex]=RF.evaluation[[3]]
RF_sensitivity[iterationindex]=RF.evaluation[[4]]
SVM_AUC[iterationindex]=SVM.evaluation[[1]]
SVM_accuracy[iterationindex]=SVM.evaluation[[2]]
SVM_specificity[iterationindex]=SVM.evaluation[[3]]
SVM_sensitivity[iterationindex]=SVM.evaluation[[4]]
LR_AUC[iterationindex]=LR.evaluation[[1]]
LR_accuracy[iterationindex]=LR.evaluation[[2]]
LR_specificity[iterationindex]=LR.evaluation[[3]]
LR_sensitivity[iterationindex]=LR.evaluation[[4]]

}
```

ÖZGEÇMİŞ



Adı Soyadı : Adil Hani ABDULKAREEM

Doğum Yeri ve Yılı :

Medeni Hali :

Yabancı Dili : Türkçe, Arapça, İngilizce

E-posta :

Eğitim Durumu

Lisans : Bağdat Üniversitisi
Bilgisayar Bilimleri Bölümü, 2012

Yüksek Lisans : İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği, 2021

Yayınları

Abdulkareem, A, Kasapbaşı, M. (2020). ENHANCING DETECTION METHOD OF BREAST CANCER USING COIMBRA DATASET. İstanbul Ticaret Üniversitesi Teknoloji ve Uygulamalı Bilimler Dergisi, 3 (1), 51-59. Retrieved from https://dergipark.org.tr/en/pub/icujtas/issue/57160/824195