T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SVM, NB, KNN, ADABOOST ve RANDOM FOREST SINIFLANDIRMA ALGORİTMALARI KULLANILARAK MEME KANSERİNİN TAHMİNİ

YÜKSEK LİSANS TEZİ

Ayça ACET

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Abdullah Erhan AKKAYA

HAZİRAN 2022

T.C.
İNÖNÜ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

SVM, NB, KNN, ADABOOST ve RANDOM FOREST SINIFLANDIRMA ALGORİTMALARI KULLANILARAK MEME KANSERİNİN TAHMİNİ

YÜKSEK LİSANS TEZİ

Ayça ACET
(36193619006)

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğretim Üyesi Abdullah Erhan AKKAYA

HAZİRAN 2022

TEŞEKKÜR VE ÖNSÖZ

Bu tez çalışmasının her aşamasında yardım, öneri, bilgi, tecrübe ve desteklerini esirgemeden beni her konuda yönlendiren danışman hocam Sayın Dr. Öğretim Üyesi Abdullah Erhan AKKAYA'ya; öğrenim hayatım boyunca desteğini bir an olsun esirgemeyen aileme; ayrıca bu tezin her aşamasında ve sürecinde önerilerini aldığım değerli hocalarıma,

teşekkür ederim.

ONUR SÖZÜ

Doktora veya yüksek lisans tezi olarak sunduğum “SVM, NB, KNN, AdaBoost ve Random Forest Sınıflandırma Algoritmaları Kullanılarak Meme Kanserinin Tahmini” başlıklı bu çalışmanın bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığına ve yararlandığım bütün kaynakların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterilenlerden oluştuğunu belirtir, bunu onurumla doğrularım.

Ayça ACET

İÇİNDEKİLER

TEŞEKKÜR VE ÖNSÖZ ..... **i**
ONUR SÖZÜ ..... **ii**
İÇİNDEKİLER ..... **iii**
ÇİZELGELER DİZİNİ ..... **iv**
ŞEKİLLER DİZİNİ ..... **v**
SEMBOLLER VE KISALTMALAR ..... **vi**
ÖZET ..... **vii**
ABSTRACT ..... **viii**
1. GİRİŞ **1**
1.1 Veri Madenciliği 3
1.2 Temel Terimler ve Kavramlar 4
1.2.1 Veri ön işleme 4
1.2.2 Özellik ölçeklendirme 5
1.2.3 Sınıflandırma 5
1.2 Literatür Taraması 5
2. MATERYALLER VE YÖNTEMLER **8**
2.1 Sınıflandırma Algoritmaları Adımları 8
2.1.1 Eğitim aşaması 8
2.1.2 Test aşaması 9
2.2 AdaBoost (AB) 9
2.3 k-Nearest Neighbor (kNN) 10
2.4 Naïve Bayes (NB) 11
2.5 Random Forests (RF) 12
2.6 Support Vector Machine (SVM) 14
2.7 R Dili 16
2.8 Sınıflandırıcı Performans Değerlendirme Kriterleri 16
2.8.1 Hata matrisi (Confusion Matrix) 17
2.8.2 Doğruluk (Accuracy) 18
2.8.3 Duyarlılık (Sensitivity or Recall) 18
2.8.4 Kesinlik (Precision) 18
2.8.5 F1-skoru (F1-Score) 18
3. UYGULAMALAR **20**
3.1 Meme Kanseri Veri Seti 20
3.2 Meme Kanseri Veri Setine Ait Korelasyon Matrisi 21
3.3 Sınıflandırma Yöntemlerinin Veri Setine Uygulanması 24
3.3.1 AB yönteminin uygulanması 24
3.3.2 kNN yönteminin uygulanması 25
3.3.3 NB yönteminin uygulanması 26
3.3.4 RF yönteminin uygulanması 27
3.3.5 SVM yönteminin uygulanması 28
3.4 Uygulama Sonuçlarının Karşılaştırılması 29
4. SONUÇ ve GELECEK ÇALIŞMALAR **31**
4.1 Sonuç 31
4.2 Gelecek Çalışmalar 32
KAYNAKLAR **33**
ÖZGEÇMİŞ **36**

ÇİZELGELER DİZİNİ

Çizelge 2.1: İki sınıflı bir veri setine ait hata matrisi. 18
Çizelge 3.1: Veri seti WDBC'nin ortalama, standart sapma ve maksimum değerleri. 21
Çizelge 3.2: AdaBoost algoritmasına ait performans kriterleri. 24
Çizelge 3.3: kNN algoritmasına ait performans kriterleri. 26
Çizelge 3.4: Naïve Bayes algoritmasına ait performans kriterleri. 27
Çizelge 3.5: Random Forest algoritmasına ait performans kriterleri. 28
Çizelge 3.6: SVM algoritmasına ait performans kriterleri. 29
Çizelge 3.7: Sınıflandırma algoritmalarına ait performans kriterleri. 30

ŞEKİLLER DİZİNİ

Şekil 2.1: RF algoritmasına ait modelin şematik gösterimi. 14
Şekil 2.2: Doğrusal SVM karar düzlemi. 16
Şekil 3.1: WDBC veri setinde ortalama değerlere ait korelasyon matrisi. 22
Şekil 3.2: WDBC Veri setinde standart sapma değerlerine ait korelasyon matrisi. 23
Şekil 3.3: WDBC Veri setinde maksimum değerlere ait korelasyon matrisi. 23
Şekil 3.4: AdaBoost sınıflandırma algoritmasına ait hata matrisi. 24
Şekil 3.5: Farklı k değerlerine ait doğruluk performansı. 25
Şekil 3.6: kNN sınıflandırma algoritmasına ait hata matrisi. 26
Şekil 3.7: Naïve Bayes sınıflandırma algoritmasına ait hata matrisi. 27
Şekil 3.8: Random Forest sınıflandırma algoritmasına ait hata matrisi. 28
Şekil 3.9: SVM sınıflandırma algoritmasına ait hata matrisi. 29

SEMBOLLER VE KISALTMALAR

AB **:** AdaBoost

kNN **:** k-Nearest Neighbor

NB **:** Naïve Bayes

RF **:** Random Forest

SVM **:** Support Vector Machine

ÖZET

Yüksek Lisans Tezi

SVM, NB, KNN, ADABOOST ve RANDOM FOREST SINIFLANDIRMA ALGORİTMALARI KULLANILARAK MEME KANSERİNİN TAHMİNİ

AYÇA ACET

İnönü Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

36+VIII sayfa

2022

Danışman: Dr. Öğretim Üyesi Abdullah Erhan AKKAYA

Günümüzde kanser en yaygın hastalıklardan biridir. Yaygın olan kanser türünden biri de meme kanseridir. Meme kanseri erkek bireylere oranla kadın bireylerde daha fazla gözükmektedir. Meme kanserine yakalanma nedenleri; genetik yatkınlık, stres, bireylerin kullandığı kötü alışkanlıklar (tütün ürünleri, alkol vb.) gibi etkenler bireyin kansere yakalanma riskini arttırmaktadır. Meme hücrelerinin kontrol dışı olarak çoğalması, büyümesi ve diğer dokulara yayılması meme kanseri oluşmasının nedenidir. Kanser hücreleri normal hücre bölünmesi, çoğalması ve büyümesinden farklı bir davranış sergileyerek kontrol dışı çoğalma ve büyümesiyle başka dokulara yayılmasıdır. Kanser hücreleri sağlıklı hücre davranışının aksine genellikle çok hızlı ve sürekli olarak çoğalırlar. Kanser hücreleri iyi huylu tümör ve kötü huylu tümör olarak ikiye ayrılmaktadır. İyi huylu tümör (benign) diğer dokulara yayılmadan sadece kendi bulunduğu alanda büyüyen bir yapıya sahip. Kötü huylu tümör (malignant) ise hem bulunduğu alanda büyümüş hem de diğer dokulara yayılmıştır. Meme kanserini yenmenin en önemli etkenlerinden biri de erken teşhistir. Hatta her kanser türünde erken teşhisin hayat kurtardığı belirtilmektedir. Erken teşhis ile hasta bireye verilecek olan zarar azalır ve bununla birlikte iyileşme sürecinde başarı oranı artmaktadır.

Meme kanserinin sınıflandırabilmesi için makine öğrenmesi yöntemleri kullanılmaktadır. Bu yöntemlerde meme kanseri hastası bireyin meme hücresi bilgileri alınır ve bu bilgiler girdi olarak makine öğrenmesi algoritmalarına verilmektedir. Girdi verileri ve makine öğrenmesi metodlarını kullanarak bir çıktı olarak bir veri oluşturur. Yani makine öğrenimi algoritmaları, girdi olarak meme kanseri hücre modelini ve çıktı olarak da kanser türünün iyi huylu ya da kötü huylu olarak etiketlenerek kullanılır. Modelin iyi olarak adlandırılabilmesi için meme kanserinin doğru olarak sınıflandırılması gerekmektedir. Makine öğrenimi için kullanılan algoritmalar; AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) algoritmaları. Sınıflandırma yapabilmek için bu algoritmaların uygulanması ve sonuçlara göre doğruluk oranlarının karşılaştırılması sağlanmıştır.

Anahtar Kelimeler: Meme kanseri, SVM, AdaBoost, KNN, Random Forest, Naïve Bayes

ABSTRACT

Master Thesis

PREDICTION OF BREAST CANCER USING SVM, NB, KNN, ADABOOST AND RANDOM FOREST CLASSIFICATION ALGORITHMS

AYÇA ACET

Inonu University
Graduate School of Nature and Applied Sciences
Department of Computer Engineering

36+VIII sayfa

2022

Supervisor: Asst. Prof. Dr. Abdullah Erhan AKKAYA

Cancer is one of the most common diseases today. One of the most common types of cancer is breast cancer. Breast cancer is more common in women than in men. Causes of breast cancer; Factors such as genetic predisposition, stress, bad habits used by individuals (tobacco products, alcohol, etc.) increase the risk of developing cancer. Uncontrolled proliferation, growth and spread of breast cells to other tissues is the cause of breast cancer. Cancer cells show a behavior different from normal cell division, proliferation and growth, and spread to other tissues by uncontrolled proliferation and growth. Cancer cells usually multiply very rapidly and continuously, in contrast to the behavior of healthy cells. Cancer cells are divided into benign tumors and malignant tumors. Benign tumor (benign) has a structure that grows only in its own area without spreading to other tissues. A malignant tumor has grown both in the area where it is located and has spread to other tissues. One of the most important factors in beating breast cancer is early detection. In fact, it is stated that early diagnosis saves lives in all types of cancer. With early diagnosis, the damage to the sick individual is reduced and the success rate increases in the recovery process. Machine learning methods are used to classify breast cancer. In these methods, breast cell information of an individual with breast cancer is taken and this information is given to machine learning algorithms as input. It creates a data as an output using input data and machine learning methods. In other words, machine learning algorithms are used by labeling the breast cancer cell model as input and the cancer type as benign or malignant as output. In order for the model to be called good, breast cancer must be classified correctly.

Algorithms used for machine learning; AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) algorithms. In order to make classification, these algorithms were applied and the accuracy rates were compared according to the results. By making use of these algorithms, a good model is used to correctly classify breast cancer cells. With this correct classification, it is possible to start treatment early and reduce the risk of death of the patient, thanks to early diagnosis.

Keywords: Breast Cancer, SVM, AdaBoost, KNN, Random Forest, Naïve Bayes

1. GİRİŞ

Kas ve sinir hücreleri dışında insan vücudundaki bütün sağlıklı hücreler bölünebilme yeteneğine sahiptir. Sağlıklı bir hücre, kendi hücresel yaşam süresi boyunca ne zaman ve ne kadar bölüneceğini bilmektedir. Genlerdeki mutasyonlar ve anormal değişiklikler nedeniyle oluşan kontrolsüz bölünmeye bağlı olarak sağlıklı hücreler kanser hücrelerine dönüşmektedir. Kanser hücreleri genellikle çok hızlı ve sürekli olarak çoğalırlar. Herhangi bir organizmada başka dokulara sızma ve yayılma özelliği gösteren bir tümörün varlığı kanserin belirtisidir (Danacı ve diğ., 2010). Eğer tümör sadece kendi bulunduğu alanda büyüyen bir yapıdaysa iyi huylu (benign) tümör olarak adlandırılır. Bunun tersine bulunduğu alanda büyümekle kalmayıp diğer dokulara da yayılan bir yapıdaysa kötü huylu (malignant) tümör olarak adlandırılır. Kanser iyi huylu aşamada daha az risk taşıyarak yaşamı tehdit etmezken, kötü huylu aşamada yaşamı tehdit etmekte ve ölümle sonuçlanabilmektedir.

Meme kanseri, meme hücrelerinin kontrol dışı olarak çoğalması, büyümesi ve diğer dokulara yayılması olarak tanımlanmaktadır. Meme kanserinde üremeye bağlı risk faktörü olarak menarş yaşının küçük olması, menopoz yaşının artması, ileri yaşta çocuk sahibi olma veya hiç çocuk sahibi olmama ve annenin bebeği daha az emzirmesi verilmektedir (Collaborative Group on Hormonal Factors in Breast Cancer, 2002; 2012). Üreme dışı risk faktörleri ise obezite (The Endogenous Hormones and Breast Cancer Collaborative Group, 2002), menopoz sonrası aşırı kilolu kadınlarda meme kanseri riskinin iki katına çıkması, ailede meme kanseri öyküsü, hormonlar, radyasyon tedavisi ve teşhis edilen tüm meme kanseri vakalarının %4'üne katkıda bulunduğu tahmin edilen, giderek artan alkol tüketimi yer almaktadır (Rumgay ve diğ., 2021).

Meme kanseri ciddi bir küresel sağlık problemidir. Sadece 2020 yılı içerisinde kaydedilen tahmini 2,26 milyon yeni vaka ile meme kanseri dünyada en sık teşhis edilen kanser türüdür. Az gelişmiş ülkelerde meme kanserinin yarısından fazlası 50 yaşın altındaki kadınlarda görülürken, İngiltere gibi gelişmiş ülkelerde meme kanserinin üçte birinden fazlası 70 yaşın üzerindeki kadınlarda görülmektedir (Wilkinson ve Gathani, 2022).

Erkeklerde ve kadınlarda bütün ölümlerin %21'lik kısmından kanser sorumludur. Kanser her ne kadar kalp hastalıklarından sonra önde gelen ikinci ölüm nedeni olsa da yaş aralığına bakıldığında 40-79 yaş aralığındaki kadınlar ve 60-79 yaş aralığındaki erkekler arasında önde gelen ölüm nedenidir. Beyin ve sinir sisteminde meydana gelen kansere bağlı ölümler 40 yaş altı erkeklerde ve 20 yaş altı kadınlarda sıklıkla görülmektedir. Meme kanserine bağlı ölümler ise 20-59 yaş arasında gerçekleşmektedir. Sadece Amerika'da 2022 yılı içerisinde 43,250 kadının meme kanserine bağlı olarak hayatını kaybedeceği öngörülmektedir (Siegel ve diğ., 2022).

Bütün alanlarda olduğu gibi tıp alanında da yapay zekâ algoritmaları giderek yaygınlaşmaktadır. Meme kanserinin iyi huylu veya kötü olarak sınıflandırılabilmesi için yapay zekâ yaklaşımlarından makine öğrenmesi yöntemleri yaygın olarak kullanılan birçok metodu içermektedir. Temel olarak bir hastaya ait meme hücresi bilgileri alınır ve girdi olarak bir makine öğrenmesi algoritmasına verilir. Hücre bilgilerine dayalı olarak kötü huylu veya iyi huylu olarak sınıflandırma yapılmaktadır. Makine öğrenmesi tekniklerinden AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) algoritmaları kullanılmıştır. Bütün bu tekniklerin amacı girdi olarak verilen hücresel bilgiye göre sınıflandırma, kanser türünü tahmin etme ve gelecekte tekrar kullanmak üzere bir model oluşturmadır. Oluşturulan modelin iyi olarak adlandırılabilmesi için meme kanserinin doğru olarak sınıflandırılması gerekir. Doğru sınıflandırmaya bağlı olarak konulacak erken tanı sayesinde tedaviye de erken başlanması ve devamında kanserin tamamen ortadan kalkması mümkün olmaktadır.

AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) algoritmaları Irvine California Üniversitesi (UCI) veri madenciliği havuzundan alınan Wisconsin Breast Cancer Dataset (UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, t.y.) veri setine uygulanmıştır. Veri seti toplamda 569 hastanın meme hücrelerine ait özelliklerini içermektedir. Bu veri setinin tercih edilme sebeplerinin başında sık kullanılan bir veri seti olması ve kayıpsız bir yapıya sahip olması gelmektedir.

Bu tez çalışması dört bölümden oluşmaktadır. Bölüm 1'de meme kanseri hakkında genel bilgilerle probleme giriş yapıldıktan sonra veri madenciliğinin temel kavramları, veri ön işleme, özellik ölçeklendirme, denetimli öğrenme ve sınıflandırma kavramları ve veri madenciliğinin Tıp alanındaki öneminden bahsedilmiştir. Bölüm 2'de AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector

Machine (SVM) sınıflandırma algoritmalarından geliştirme ortamı ve sınıflandırıcı performans değerlendirme kriterlerinden bahsedilmiştir. Bölüm 3'te Wisconsin Breast Cancer Database (WDBC) veri setinin detaylı açıklaması, veri setinin hazırlanması, veri setini ait korelasyon matrisi ve sınıflandırma yöntemlerinin nasıl uygulandığından bahsedilmiş; hata matrisleri üzerinden uygulama sonuçları tartışılmıştır. Seçilen sınıflandırma yöntemleri R dili kullanılarak uygulanmıştır. Bölüm 4 sonuç ve gelecek çalışmalara ayrılmıştır.

1.1 Veri Madenciliği

Veri madenciliği (data mining), yapılandırılmış ya da yapılandırılmamış veriyi anlamak ya da öngörülebilir sonuçlar elde edebilmek için geniş veri kümeleri üzerinde modelleri, korelasyonları ve eğilimleri çıkarma işlemidir. Asıl olarak veri madenciliği; seçilen veri setlerinin analize uygun bir biçimde hazırlanması ile anlamlı çıkabilecek beklenilen ya da beklenmeyen bilgiye ulaşma sürecidir denebilir. Veri madenciliğinin genel amacı belirli bir veri kümesinden alakalı önemli veriler çıkarmak ve bu verilerin daha sonra kullanılmak üzere yapılandırılmasını sağlamaktır. Veri madenciliği, istatistik ve yapay zekâ kullanarak geniş veri tabanlarından ve veri ambarlarından ilgili ve faydalı olacak bilgilerin analiz edilmesi ve çıkarılmasıyla ilgilenen makine öğrenmesinin alt alanıdır.

Veri Madenciliği yöntemlerinin iki tür tekniği vardır; denetimli öğrenme ve denetimsiz öğrenmedir. Veri Madenciliğinde net bir şekilde tanımlanmış veya kesin bir amaç ifade edildiğinde denetimli terimi kullanılır. Elde edilmesi hedeflenen sonuç için özel bir tanımlamada bulunulmamışsa veya belirsizlik söz konusu ise denetimsiz terimi kullanılmaktadır (Chapman ve diğ., 2000). Veri madenciliği üç alandan faydalanmaktadır; yapay zekâ, istatistik (veriler arasındaki sayısal ilişkiler) ve makine öğrenmesi (verilerden öğrenerek tahmin yürütme). Veri madenciliği süreçleri:

1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Veri madenciliği algoritmasını uygulama
6. Sonuçları sunum ve değerlendirme

olmak üzere altı adımdan oluşmaktadır.

Veri Temizleme: Veri tabanında yer alan hatalı ve tutarsız verilere gürültü denir. Verilerdeki gürültüyü temizlemek için yapılan çalışmaya veri temizleme denmektedir. Veri temizlemede gereksiz veriler silinebilir veya kayıp veriler yerine uygun bir tahmini değer yazılabilir.

Veri Bütünleştirme: Farklı veri tabanlarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi işlemidir.

Veri İndirgeme: Veri madenciliği uygulamalarında çözümlemeden elde edilecek sonuç değişmeyecek ise veri sayısı ya da değişkenlerin sayısı azaltılabilir. Veri indirgeme yöntemleri; veri sıkıştırma, örnekleme, genelleme veya boyut indirgeme gibidir.

Veri Dönüştürme: Verinin kullanılacak modele göre içeriğini koruyarak biçimin dönüştürülmesi işlemidir. Dönüştürme işlemi kullanılacak modele uygun biçimde yapılmalıdır. Çünkü verinin gösterilmesinde kullanılacak model ve algoritma önemli bir rol oynamaktadır. Değişkenlerin ortalama ve varyansları birbirlerinden büyük oranda farklı olduğu zaman veri üzerinde normalizasyon işlemi yapılmasıdır.

Veri Madenciliği Algoritmasını Uygulama: İşlenecek veri hazır hale getirildikten sonra konuyla ilgili veri madenciliğinin algoritmaları uygulanır.

Sonuçları Sunum ve Değerlendirme: Veri madenciliği algoritmaları veriler üzerinde uygulandıktan sonra sonuçlar düzenlenerek ilgili yerlere sunulur.

1.2 Temel Terimler ve Kavramlar

1.2.1 Veri ön işleme

Veri ön işleme, verilerin hazırlanmasını ve veri madenciliğine uygun bir biçime dönüştürülmesini sağlayan görevlerden biridir. Veri boyutunu küçültme, veriler arasındaki ilişkileri tespit etme, kayıp değerlere sahip veri satırlarını veri setinden kaldırma, veri normalizasyonu ve veriler için özellik çıkarma veri ön işleme adımlarındandır. Veri ön işleme temel olarak veri temizleme, veri entegrasyonu, dönüştürme ve azaltma gibi çeşitli teknikleri içerisinde barındırmaktadır (Alasadi ve Bhaya, 2017).

1.2.2 Özellik ölçeklendirme

Makine öğreniminin doğruluğunu arttırmak için aykırı değer tespiti, özellik ölçekleme, özellik azaltma yoluyla veri temizleme teknikleri kullanılmaktadır (Ahmadi ve diğ., 2021). Özellik ölçekleme yönteminde veri kümelerinden gelen veriler, aykırı değerlere bakılmadan, aynı veri aralığına sıkıştırılır (Nkikabahizi ve diğ., 2022). Özellik ölçekleme bağımsız değişkenlerin veya verilerin özelliklerinin aralığını normalleştirmede kullanılan bir tekniktir. Veri normalleştirme olarak bilinir ve genellikle veri ön işleme aşamasında kullanılır.

1.2.3 Sınıflandırma

Genel olarak bir veri kümesindeki verilerin ortak özelliklerinden faydalanarak veriyi sınırlı sınıflar arasında dağıtma işlemi sınıflandırma olarak adlandırılmaktadır. Bu çalışmada bahsedilen sınıflandırma işlemleri öğrenme algoritmaları aracılığı ile gerçekleştirilmiştir. Sınıflandırma işleminde öğrenme algoritmaları veriye uygulanarak sınıflandırıcılar oluşturulur, bu sınıflandırıcılar yeni verilere uygulanır ve yeni gelen herhangi bir sınıfa ait olmayan veri bu sayede bir sınıfa atanmış olur.

1.2 Literatür Taraması

Bu bölümde meme kanseri teşhisi için kullanılan sınıflandırma algoritmaları olan AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) sınıflandırma algoritmaları kullanılarak literatürde bulunan çalışmalar incelenmiştir.

2006 yılında, meme kanseri teşhisi için bir prensip olan minör kalsifikasyon otomatik olarak segmentasyonunu sağlamak için kullanılan mamogramların segmentasyonu için radikal olarak yeni iki fazlı bir teknik kullanıldı. Birlikte oluşum matrisi ekstraksiyon işlevinde kullanılır ve mamogramı iyi huylu veya kötü huylu olarak tanımlamak için destek vektör makinesi (SVM) kullanılmıştır. Önerilen bu yöntemde %92,8'lik bir ortalama duyarlılık hesaplanmıştır (Selvi ve Malmathanraj, 2006).

Zhang ve arkadaşları 2014 yılında K-ortalama ve SVM sınıflandırıcılarını birleştiren bir hibrit model önerdi. Bu modelin amacı, seçme ve çıkarma metodlarını kullanarak Wisconsin Meme Kanseri teşhisi (WDBC) veri setinde tümörün sınıflandırmasını yapmaktır. Özellik seçme ve çıkarma yöntemini kullanarak Wisconsin

Meme Kanseri teşhisi (WDBC) veri setinden tümör özelliklerini teşhis etmekti. İyi huylu ve kötü huylu tümör modellerini belirlemek için bir K-ortalama sınıflandırıcı kullanıldı. Oluşturulan modeller hesaplanır ve Support Vector Machine (SVM) modelinin eğitimi için yeni modeller olarak kabul edilir. Tümörlerin sınıflandırma tahmini için SVM yürütülür. Hibrit modellerinin kullanılması, doğruluğu %97'ye çıkardı (Seddik ve Shawky, 2015).

2015 yılında yapılan bir çalışmada kompaktlık, aralık ve varyans gibi özelliklerin çıkarılması üzerinde çalışıldı. Performansı analiz etmek için Support Vector Machine (SVM) sınıflandırıcı algoritması kullanıldı. En yüksek varyansı %95 ve kompaktlık %86 değerleri vermektedir. Meme kanseri öngörüsü için elde edilen bu sonuçlara göre Support Vector Machine (SVM) uygun bir yöntem olarak değerlendirilebilir (Gc ve diğ., 2015).

Bir diğer çalışmada, Wisconsin meme kanseri veri setinde Support Vector Machine (SVM) ve k-en yakın komşu (kNN) uygulanmış ve bu algoritmalar kullanılarak prediktif bir model önerilmiştir. Veri seti 699 gözlem ve 11 özellik içermektedir. Yazarlar algoritmaların performansını karşılaştırdılar ve sonuç; destek vektör makinesini K-en yakın komşu (kNN) algoritmasından daha yüksek doğrulukla daha iyi bir algoritma olarak gösterilmektedir (Pawlovsky ve Nagahashi, 2014).

Morgana Darshini Ganggayah, meme kanseri hastalarının sağ kalımını tahmin etmek için Makine Öğrenimini kullanmaktadır. Önerilen model Random Forest'tan alınan bir modeldir. Random Forest'tan alınan bu modelin %82.7 doğrulukla en iyi sonuçları aldığını ortaya koymaktadır (Ganggayah ve diğ., 2019).

Hazra et al. Naïve Bayes (NB) ve Support Vector Machine (SVM) gibi farklı kanser sınıflandırma yaklaşımları üzerinde karşılaştırmalı bir çalışma, sınıflandırıcının her birinin zaman karmaşıklığını ölçerek Wisconsin Meme Kanseri teşhisi (WDBC) veri setinde uygulamıştır. Naïve Bayes algoritmasının en iyi doğruluğa sahip olduğunu ve %97.4 doğruluk oranıyla en düşük zaman karmaşıklığına sahip olduğunu bildirmişlerdir (Hazra ve diğ., 2016).

N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor çalışmalarında Meme Kanseri hücre tespit etmek için bir grup makine öğrenmesi algoritması üzerinde bir karşılaştırma yapılmış ve AdaBoost'un 98.77'de bir doğruluk elde ettiği sonucuna varılmıştır (Mashudi ve diğ., 2021).

Md. Milan İslam, SVM ve kNN olan iki denetlenen öğrenme sınıflandırıcıyı işledi. Bu sınıflandırıcı algoritmalar doğruluk, duyarlılık, özgüllük, yanlış keşif oranı, yanlış

ihmal oranı ve Mathew korelasyon katsayısı açısından tahmin edildiler. 10 kat çapraz doğrulamaya sahip bir sistem kullanıldı ve Wisconsin meme kanseri teşhisi veri kümesinde K-en yakın komşu (kNN) ile Support Vector Machine (SVM) ile %98.57 ve %97.14 doğruluğunu elde ettiler (Singh ve Raj, 2021).

Zheng et al. ölüm oranını en aza indirmek için meme kanserini daha önceki bir aşamada tespit etmek amacıyla makine öğrenmesi destekli etkili bir adaboost algoritması (DLA-EABA) sunmuşlardır. Ayrıca, tümörün sınıflandırılması, aynı zamanda doğruluğu artırmak için aynı anda kusur tahmin paradigması uygulanmıştır. Bu nedenle, %96,5 doğruluk kazanmıştır. Fakat diğer yaklaşımlarla karşılaştırılırken maliyetin açısından pahalıdır (Zheng ve diğ., 2020).

Başka bir çalışmada (Christobel ve Sivaprakasam, 2011) meme kanseri tanısı için WDBC veri kümesini kullandı. Karar ağacı (DT), K-en yakın komşu (kNN), destek vektör makinesi (SVM) ve Naïve Bayes (NB) sınıflandırıcılarını kullanırlar ve sınıflandırmalarının hassasiyetini karşılaştırırlar. SVM'nin %96,99'unun ortalama doğruluğu, yöntemler arasındaki en yüksek doğruluktur.

2. MATERYALLER VE YÖNTEMLER

Bu bölümde sınıflandırma algoritmalarının temel özelliklerinden başlayarak bu tez çalışması kapsamında uygulanan AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF) ve Support Vector Machine (SVM) makine öğrenmesi sınıflandırıcılarından bahsedilmiştir. Tez çalışmasının gerçekleştirildiği yazılım ve donanım ortamından bahsedildikten sonra sınıflandırıcı performans değerlendirme kriterleri ile bölüm sonlandırılmıştır.

2.1 Sınıflandırma Algoritmaları Adımları

Sınıflandırma işlemi makine öğrenmesinin kullandığı yollardan biridir. Bilinen değerlere ait öznitelikler kullanılarak bir sınıflandırıcı oluşturulur. Oluşturulan bu sınıflandırıcı kullanılarak sınıfı belirli olmayan bir veri geldiğinde sınıflandırma işlemi yapılabilir. Sınıflandırma işlemi yapılırken veri setindeki değişkenler arası ilişki belirleyici bir role sahiptir. Bu nedenle sınıflandırma işlemi değişkenler arasındaki ilişkiyi öğrenerek net bir ayrım sunduğu için veri madenciliği ve makine öğrenmesi alanında önemli bir yere sahiptir. Sınıflandırma algoritmaları insan kaynaklı verileri kullanarak, sınıfları önceden bilinen veri kümeleri ile bir model oluşturmaktadır. Bu nedenle sınıflandırma algoritmaları denetimli öğrenmenin bir türüdür. Sınıflandırma algoritmaları genel olarak eğitim ve test aşaması olmak üzere iki aşamaya sahiptir.

2.1.1 Eğitim aşaması

Eğitim aşamasında giriş verisi kullanılarak bir sınıflandırma modeli oluşturulmaktadır. Bu aşama aynı zamanda öğrenmenin gerçekleştiği aşamadır. Eğitim aşamasında kullanılan eğitim veri seti, öğrenme süreci esnasında kullanılan örneklerden oluşan bir veri setidir. Sınıflandırma işlemi gerçekleştirilirken denetimli bir öğrenme algoritması tahmine dayalı model oluşturacak değişkenlerin optimal kombinasyonlarını belirlemek için eğitim veri setine bakar (Larose ve Larose, 2014). Amaç yeni gelen ve bilinmeyen verilere uyumlu bir model üretmektir.

2.1.2 Test aşaması

Test aşaması sınıflandırıcının öğrenme adımında karşılaşmadığı verileri tahmin etmede kullanılan modele ait sınıflandırma aşaması olarak adlandırılır. Test aşamasında kullanılan test veri seti, eğitim veri setinden bağımsız fakat eğitim veri seti ile aynı olasılık dağılımına sahip bir veri setidir. Test veri seti bir sınıflandırıcıların performansını değerlendirmek için kullanılan bir dizi örnekten oluşmaktadır. Sınıflandırıcı çıktıları etiketleme yoluyla sınıflandırılmaktadır.

2.2 AdaBoost (AB)

AdaBoost algoritması Freund ve Schapire (Freund ve Schapire, 1997) tarafından önerilen ve örnek ağırlıklarının dağılımını değiştirerek zayıf öğrenen sınıflandırıcıların doğruluğunu artırabilen topluluk öğrenme yöntemidir. AdaBoost algoritması zayıf olarak nitelendirilmiş hatalı tahminler yapan bir sınıflandırıcıyı yüksek sınıflandırma doğruluğuna sahip yeni bir sınıflandırıcıya dönüştürebilir (Ying ve diğ., 2013). Aynı eğitim seti için birden fazla sınıflandırıcı eğitilerek zayıf sınıflandırıcılar birbiriyle Entegre edilmektedir. AdaBoost algoritması temel olarak veri örneklerinin dağılımını değiştirir. Değiştirilmiş ağırlıklara sahip yeni veri seti, yeni bir zayıf sınıflandırıcı elde etmek için tekrar eğitilir. İlk tekrarda tüm numunelerin ağırlıkları aynıdır. Her tekrarda yanlış sınıflandırılan örneklerin ağırlığı artar; sınıflandırılan örneklerin ağırlığı azalır ve tüm ağırlıklar normalize edilir. Son adımda ise tahmin sınıflarının doğruluk değerlerine göre oylama yapılarak zayıf sınıflandırıcılardan iyi olanlar seçilir ve birbirleri ile entegre edilerek daha iyi bir sınıflandırıcı oluşturmak için bir araya getirilirler (Xu ve Zhang, 2014). Çoklu sınıflandırma problemleri için AdaBoost algoritmasının adımları aşağıda verilmiştir (Hastie ve diğ., 2009).

Adım 1: Veri setindeki tüm gözlemlere eşit ağırlıklar atanır.

Adım 2: Oluşturulan model rastgele örneklere göre şekillendirilir ve orijinal veriler için sınıflar tahmin edilir.

Adım 3: Toplam hata hesaplanır.

Adım 4: Toplam hata kullanılarak temel öğrenicinin başarımı hesaplanır.

Adım 5: Ağırlıklar güncellenir.

Adım 6: İterasyondaki ağırlıklar güncellenir.

Adım 7: Her bir iterasyonda tahmin edilen değerin ağırlıklı toplamı sonucu elde değerin işareti bakılarak sınıflandırma yapılır.

Eğitim iterasyonundan sonra yanlış sonuçları sınıflandıran örneklere daha büyük ağırlıklar atanmaktadır. Modelin nihai çıktıları, tüm temel öğrenicilerin ağırlıklarının kombinasyonu yoluyla oluşturulur.

2.3 k-Nearest Neighbor (kNN)

k-Nearest Neighbor, k-En Yakın Komşu algoritması 1950 yılında keşfedilmiş olmasına rağmen tanınır hale gelmesi 1960'lı yıllara dek sürmüştür (Sharma ve Suryawanshi, 2016). kNN yöntemi basit ve etkili bir yöntemdir. Etiketlenmemiş bir veri seti verildiğinde, kNN eğitim veri setindeki en yakın "k" değeri bulur ve daha sonra bu değerlere uygun sınıf etiketini atar. kNN, en yakın komşu noktaların çoğunluğunu dikkate alarak, komşularından bir veri noktasına etiketleme yapar. İşlemeyi başlatmak için algoritmanın test grubunu görmesi ve ardından depolanan en yakın eğitim gruplarını bularak verileri sınıflandırmak için genelleme yapması gerekir.

kNN; eski, basit ve gürültülü eğitim verilerine karşı dirençli olması sebebiyle en popüler makine öğrenme algoritmalarından biridir. Fakat bunun yanında dezavantajı da mevcuttur. Örneğin, uzaklık hesabı yaparken bütün durumları sakladığından, büyük veriler için kullanıldığında çok sayıda bellek alanına gereksinim duymaktadır.

kNN algoritmasının adımları aşağıda şekildedir:

- **Adım 1:** İlk olarak k parametresi belirlenir. Belirlenen bu parametre verilen bir noktaya en yakın komşuların sayısıdır. k değerinin 2 olması, en yakın 2 komşuya göre sınıflandırma yapılacağı anlamına gelmektedir.
- **Adım 2:** Örnek veri setine katılacak olan yeni verinin, mevcut verilere göre uzaklığı hesaplanır.
- **Adım 3:** İlgili uzaklıklardan en yakın k komşu ele alınır. Öznitelik değerlerine göre k komşu veya komşuların sınıfına atanır.
- **Adım 4:** Seçilen sınıf, tahmin edilmesi beklenen gözlem değerinin sınıfı olarak kabul edilir. Sonuç olarak veri etiketlenmiş olur.

kNN algoritmasında iki önemli problem vardır. İlk problem, k değerinin seçimini doğru yapmaktır. k değeri her bir ikili değer için kaç koşul seçileceğini belirlemektedir. Bu nedenle k değerinin seçimi algoritma performansı üzerinde büyük öneme sahiptir. Büyük k değerleri, küçük veri örüntülerinin göz ardı edilmesine neden olur. İkinci bir problem ise

test verisi ile bu verinin komşuları arasındaki uzaklığı hesaplamaktır. Belirli bir noktanın en yakın komşularını bulmak için mutlaka uzaklık ölçümü yapılmalıdır. En popüler uzaklık metrikleri Öklid, Minkowski ve Manhattan uzaklıklarıdır. Ancak bu uzaklıklar sadece sürekli değişkenler için geçerlidir. Öklid mesafesi, Minkowski ve Manhattan mesafelerinden daha yaygın olarak kullanıldığından bu tez çalışması kapsamında kNN algoritması Öklid uzaklığı kullanılarak çalıştırılmıştır.

Minkowski uzaklığında *p* değeri üssü ifade etmektedir. *p* değeri 1'e eşit olduğunda Denklem 2.31'deki formül bize Manhattan uzaklığını vermektedir. *p* değeri 2'ye eşit olduğunda Denklem 2.31'deki formül bize Öklid uzaklığını vermektedir.

$$L_p(X,Y) = \sqrt[p]{\sum_{i=1}^{n}(|x_i - y_i|)^p} \tag{2.31}$$

Manhattan uzaklığı, 19. yüzyılda Hermann Minkowski adındaki bir Alman bilim insanı tarafından öne sürülmüştür. Manhattan uzaklığı literatürde L_1 norm, Şehir bloğu uzaklığı (city block distance), Doğrusal (Rectilinear) mesafe veya Taxicab normu olarak da bilinmektedir. Vektörlerin noktaları arasındaki mutlak farkların toplamını temsil etmektedir, Denklem 2.32.

$$L_1(X,Y) = \sum_{i=1}^{n}|x_i - y_i| \tag{2.32}$$

Öklid uzaklığı L_2 norm veya Ruler distance olarak da bilinmektedir. Öklid uzaklığı vektörlerdeki noktalar arasındaki farkların karelerinin toplamının kökünü temsil etmektedir, Denklem 2.33.

$$L_2(X,Y) = \sqrt[2]{\sum_{i=1}^{n}(|x_i - y_i|)^2} \tag{2.33}$$

Bir veri setindeki girdi değerleri, yükseklik ve genişlik gibi birbiriyle yakından ilişkili ise Öklid uzaklığı kullanılması; girdi değerleri yaş, boy, cinsiyet gibi birbiriyle ilişkisiz verilerden oluşuyorsa Manhattan mesafesi kullanılması önerilmektedir.

2.4 Naïve Bayes (NB)

Naïve Bayes, 18. y.y.da matematikçi Thomas Bayes tarafından geliştirilmiştir. Temelinde verilen veriler kullanılarak koşullu olasılıklar hesaplanmaktadır. Naive Bayes, pratik öğrenme problemlerinde nadiren geçerli olan bir varsayıma dayanır; bir tahmin türetmek için kullanılan öznitelikler, tahmin edilen değer verildiğinde birbirinden

bağımsızdır. Bu teknik kullanılarak belirli bir sınıfa ait bir örneğin olasılığı tahmin edilir. Tüm öznitelikler, belirli bir sınıftaki öznitelik değeri ile diğer öznitelikler arasında hiçbir bağımlılık olmadığı anlamına gelen Bayes teoremine göre bağımsız varsayılmaktadır.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.41}$$

Denklem 2.41'e göre Bayes teoremi, P(A|B) B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı, P(A) A olayının gerçekleşme olasılığı, P(B) B olayının gerçekleşme olasılığı, P(B|A) B olayı gerçekleştiğinde A olayının gerçekleşme olasılığı olarak tanımlanmaktadır.

Naive Bayes, verilen bir verinin olasılığını hesaplayarak sınıfını bulmaktadır. Burada dikkat edilmesi gereken nokta değişkenlerin birbirinden bağımsız olmasıdır. Dolayısıyla bu yöntemin güçlü özelliği, farklı özelliğe sahip değişkenler arasında karşılıklı olarak bağımsız olmasıdır. Naive Bayes sınıflandırıcı, çok sayıda veri noktasıyla iyi bir performans gösterebilir. Navies Bayes yönteminin belirli bir verinin olasılığını tahmin etmek için az miktarda eğitim verisi ile çalışabilmesi bu yöntemin avantajlarındandır. Naive Bayes'ın hızlı, etkili ve doğru bir sınıflandırma algoritmasıdır. Naïve Bayes yöntemi dört adımdan oluşmaktadır (Bhat ve diğ., 2022; Gopalsamy ve Radha, 2022; Kachhia ve Rathod, 2022; Langarizadeh ve Moghbeli, 2016):

- **Adım 1:** Verilen sınıfa ait etiketler için öncül olasılık hesaplanır.
- **Adım 2:** Her bir sınıf için her öznitelik ile olabilirlik olasılığı (likelihood probability) hesaplanır.
- **Adım 3:** Bayes formülü yardımıyla ardıl olasılık hesaplanır.
- **Adım 4:** En yüksek olasılığa sahip olan sınıf tespit edilir.

2.5 Random Forests (RF)

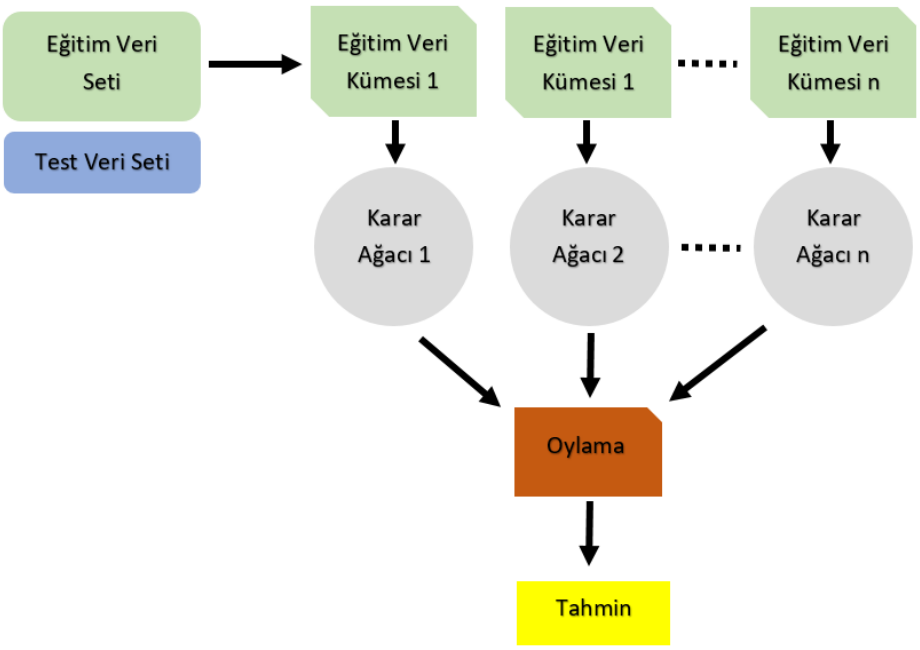
Random Forests algoritması Leo Breiman ve Adele Cutler tarafından geliştirilmiştir. Rastgele orman, denetimli öğrenme tekniğine dayanmaktadır. Hem regresyon (Pal, 2005) hem de sınıflandırma (Grömping, 2009) problemleri için kullanılabilir. Problem karmaşıklığını gidermek ve model performansını iyileştirmek adına tek bir sınıflandırıcı ağaç kullanmak yerine birçok ağaç tahmin edicilerin kombinasyonunun kullanılması mantığına dayanmaktadır (Akar ve Güngör, 2012). Birçok ağaç tahmin edicinin kombinasyonu sonucu bazı ağaçlar doğru tahminde bulunurken

diğerleri yanlış tahmin üretebilmektedir. Bu her ne kadar dezavantaj gibi görünse de bir araya geldiklerinde tüm ağaçlar doğru çıktıyı tahmin edebilmektedir. RF girdi olarak verilen veri kümesinin alt kümelerini inceleyen karar ağaçlarına sahiptir. Giriş veri kümesinin tahmin doğruluğunu iyileştirmek için ortalamayı almaktadır (Nizam ve Akın, 2014). Şekil 2.1'de RF algoritmasını şematik gösterimi verilmiştir. RF algoritmasında, oluşturulan her bir ağaç bağımsız olarak örneklenen rastgele bir vektörün değerlerine bağlıdır. Tüm ağaçlar aynı dağılıma sahiptir. RF, tüm değişkenler arasında en iyi bölünmeyi kullanarak her düğümü bölmek yerine, her düğümü, o düğümde rasgele seçilen bir tahmin edici alt kümesi arasından en iyisini kullanarak bölmektedir. RF algoritmasının adımları aşağıda verilmiştir.

- **Adım 1:** Giriş veri setinden rastgele örnekler seçilir.
- **Adım 2:** RF algoritması, seçilen her örnek için tahmin sonucunu verecek bir karar ağacı oluşturur.
- **Adım 3:** Tahmin edilen her bir sonuç için sınıflandırma probleminde mod kullanılır.
- **Adım 4:** En çok oy alan tahmin çıkış olacaktır.

Eğitilmiş k adet karar ağacının RF modeli içerisindeki tanımı Denklem (2.51)'de verilmiştir (Chen ve diğ., 2016). Denklemde $H(X, \theta_j)$ meta karar ağacı sınıflandırıcıdır. Meta karar ağacı sınıflandırıcıların, sıradan karar ağaçlarından farkı sınıf değerini doğrudan tahmin etmek yerine hangi temel seviye sınıflandırıcının tahminde kullanılması gerektiğini belirtmesidir (Todorovski ve Džeroski, 2003).

$$H(X, \theta_j) = \sum_{i=0}^{k} h_i(x, \theta_j), \quad (j = 1,2,3, \dots, m) \tag{2.51}$$



Şekil 2.1: RF algoritmasına ait modelin şematik gösterimi.

2.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) algoritması ilk olarak 1963 yılında Vapnik tarafından doğrusal bir sınıflandırıcı oluşturmak için önerilmiştir (Vapnik, 1963). SVM denetimli bir öğrenme modelidir. Girdi olarak verilen örnekler bir sınıf etiketine sahiptir. SVM, kendisine girdi olarak verilen verileri iki sınıfa ayırmak için n-boyutlu bir hiperdüzlem kullanmaktadır (Noble, 2006). Hiperdüzlemler aynı zamanda karar sınırları olarak bilinmektedir. Karar sınırı, her bir sınıfın en yakın veri noktalarından mümkün olduğunca uzak olacak şekilde oluşturulmaktadır. Hiperdüzlemi belirleyen bu veri noktalarına "destek vektörleri" denmektedir.

Doğrusal olmayan modellerde iki sınıf arasındaki mesafeyi hesaplamak için "kernel" adı verilen yapılar kullanılmaktadır. Kernel, yüksek boyutlu ve doğrusal olmayan modellerin oluşturulmasını sağlayan kernel fonksiyonuna sahiptir. Doğrusal olmayan bir problemde, işlenmemiş verilere ek boyutlar eklemek ve böylece daha yüksek boyutlu uzayda işlenmemiş veriyi doğrusal bir problem haline getirmek için kernel fonksiyonu kullanılabilir. Kernel fonksiyonu, belirli hesaplamaların daha hızlı yapılmasına yardımcı olacak şekilde tasarlanmıştır.

Şekil 2.2'de gösterildiği gibi doğrusal SVM'ye ait bir karar düzlemi sınıfları çok iyi bir şekilde ayırabilmektedir. Bir diğer ifadeyle, iki sınıf olarak etiketlenen veri kümeleri için bu iki sınıf arasına ayırt edici bir doğru çizilebilir. Denklem 2.61'de belirtildiği üzere etiketleri bilinen (denetimli öğrenme) bir veri seti için, x_i özellik vektörüdür, y_i ise eğitim verisine ait etiketlerdir (+1 veya -1). Bu tez çalışmasında kullanılan WDBC veri setine ait etiketler "malignant" ve "benign" olarak tanımlanmaktadır. *n* değeri veri setindeki toplam özellik sayısını ifade etmektedir.

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$
$$(x_i, y_i) \in \mathbb{R}^n \times \{-1, +1\} \tag{2.61}$$

Optimal hiperdüzlem Denklem (2.62)'de formülize edilmektedir. Burada *w* hiperdüzleme ait normal vektör, xi giriş özellik vektörü ve b ise bias değeridir.

$$wx^T + b = 0$$
$$wx_i^T + b \geq +1 \; eğer \; y_i = +1$$
$$wx_i^T + b \leq -1 \; eğer \; y_i = -1 \tag{2.62}$$

H_1 ve H_2 birbirine paralel iki hiperdüzlem olmak üzere Denklem (2.63) aşağıdaki şekilde verilebilir.

$$H_1: wx_1^T + b = +1$$
$$H_2: wx_2^T + b = -1 \tag{2.63}$$

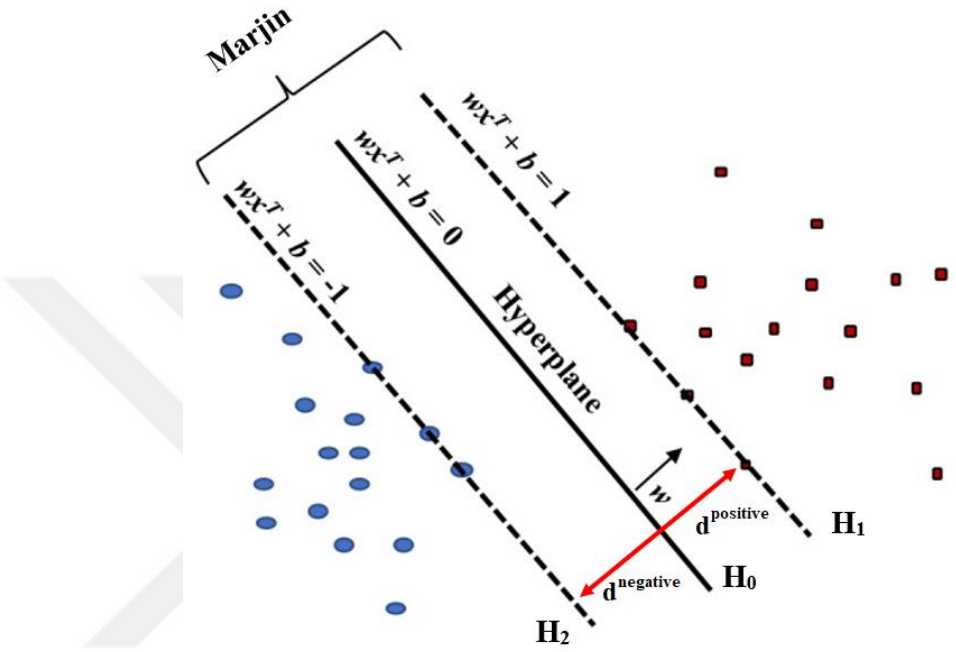
Denklem (2.63) kullanılarak H_1 ve H_2 düzlemlerinin farkı alındığında elde edilen değer bu iki düzlem arası uzaklık olacaktır.

$$wx_1^T + b = +1$$
$$wx_2^T + b = -1$$
$$= w(x_1^T - x_2^T) + b = 2$$
$$= \left(\frac{w}{\|w\|}(x_1^T - x_2^T) + b\right) = \frac{2}{\|w\|} = \frac{2}{\sqrt{w.w}} \tag{2.64}$$

Denklem (2.64) kullanılarak H_1 ve H_2 düzlemleri arası uzaklığın $\frac{2}{\|w\|}$ olduğu görülmektedir. H_1 ve H_2 düzlemlerine eşit uzaklıklı bir H_0 medyanı olduğunu kabul edersek H_0 düzlemi:

$$H_0 : wx_0^T + b = 0 \tag{2.65}$$

olarak formüle edilir. H_0 ve H_1 düzlemleri arası uzaklık $d^{positive}$, H_0 ve H_2 düzlemleri arası uzaklık $d^{negative}$ olmak üzere Denklem (2.64)'e göre bu uzaklık değeri $\frac{1}{\|w\|}$ olmaktadır. Bir SVM modeli eğitilirken temel amaç w ve b değerlerini hesaplamaktır. Böylece hiperdüzlem verileri ayırarak marjı maksimize eder (Şekil 2.2).



Şekil 2.2: Doğrusal SVM karar düzlemi.

2.7 R Dili

Tez çalışmasında yazılım geliştirme platformu olarak R dili kullanılmıştır. R dili, Ross Ihaka ve Robert Gentleman tarafından ilk olarak 1993 yılında ortaya atılmıştır. R dili günümüzde hala R Core Team tarafından geliştirilmeye devam etmektedir. R dili doğrusal ve doğrusal olmayan modelleme, zaman serisi analizi, sınıflandırma, kümeleme gibi birçok teknik sunmaktadır. R yazılımı Comprehensive R Archive Network (CRAN) ağı üzerinden yüklenmektedir. Türkiye'de Pamukkale Üniversitesinde bulunan R sunucusu üzerinden yükleme yapılabilir (http://cran.pau.edu.tr/).

2.8 Sınıflandırıcı Performans Değerlendirme Kriterleri

Bir sınıflandırıcının etkinliğini görebilmek için eğitim aşaması tamamlandıkran sonra sınıflandırıcının tahmin doğrulukları ölçülmelidir. Bu tez çalışması kapsamında hata

matrisi oluşturulduktan sonra doğruluk, duyarlılık, kesinlik ve F1 skoru kriterleri kullanılarak tahmin modellerinin performansları karşılaştırılmıştır.

2.8.1 Hata matrisi (Confusion Matrix)

Hata matrisi, makine öğrenmesi algoritmaları sonucu oluşturulan sınıflandırma modelinin tahmin ettiği değerin, gerçek sınıf değeri ile ne kadarının uyuştuğunu göstermek için kullanılan 2x2 boyutunda bir matristir.

Sınıflandırma modelinin performansı genellikle hata matrisindeki veriler kullanılarak değerlendirilir. Çizelge 2.1 iki sınıflı bir sınıflandırıcı için hata matrisini göstermektedir. Her bir örnek mutlaka iki sınıftan birine atanmaktadır. Gerçek sınıf etiketleri ve sınıflandırıcının tahmin ettiği etiketler doğru ve yanlış olarak değerlendirilir. Bu tez çalışması kapsamında kullanılan veri seti etiketleri “Malignant” (kötü huylu, 0, negatif) ve “Benign” (iyi huylu, 1, pozitif) olarak ele alınmıştır. İki sınıfa sahip olduğumuzdan dolayı gerçek sınıf ve tahmin edilen sınıflara ait dört farklı durum karşımıza çıkmaktadır, (Gerçek etiket, Tahmin edilen etiket):

Gerçek ve tahmin edilen sınıf etiketleri (0, 0) olduğunda, hata matrisinde “True Negative” alanına veri girişi yapılmaktadır. Gerçekte “Malignant” olan verinin, “Malignant” sınıfına doğru olarak atandığını ifade etmektedir.

Gerçek ve tahmin edilen sınıf etiketleri (0, 1) olduğunda, hata matrisinde “False Positive” alanına veri girişi yapılmaktadır. Gerçekte “Malignant” olan verinin “Benign” sınıfına atandığını, aslında bu verinin yanlışlıkla pozitif (False Positive) sınıfa atanmış bir negatif (Malignant-0) değer olduğunu ifade etmektedir.

Gerçek ve tahmin edilen sınıf etiketleri (1, 0) olduğunda, hata matrisinde “False Negative” alanına veri girişi yapılmaktadır. Gerçekte “Benign” olan verinin “Malignant” sınıfına atandığını, aslında bu verinin yanlışlıkla negatif (False Negative) sınıfa atanmış bir pozitif (Benign-1) değer olduğunu ifade etmektedir.

Gerçek ve tahmin edilen sınıf etiketleri (1, 1) olduğunda, hata matrisinde “True Positive” alanına veri girişi yapılmaktadır. Gerçekte “Benign” olan verinin, “Benign” sınıfına doğru olarak atandığını ifade etmektedir.

Doğruluk, duyarlılık, kesinlik ve F1 skoru gibi sınıflandırıcının performanısını belirleyen diğer kriterler hata matrisi kullanılarak hesaplanmaktadır.

Çizelge 2.1: İki sınıflı bir veri setine ait hata matrisi.

		Gerçek Sınıf Etiketi	
		True or Positive (Benign, 1)	False or Negative (Malignant, 0)
Tahmin Edilen Sınıf Etiketi	True or Positive (Benign, 1)	True Positive	False Positive
	False or Negative (Malignant, 0)	False Negative	True Negative

2.8.2 Doğruluk (Accuracy)

Doğruluk, tüm doğru tahminlerin sayısının (TP+TN) veri kümesinin toplam sayısına (TP+TN+FP+FN) bölünmesiyle hesaplanır. En kötü doğruluk değeri 0, en iyi doğruluk değeri 1'dir. (1-hata oranı) ile de hesaplanabilir.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 1 - error\ rate$$

2.8.3 Duyarlılık (Sensitivity or Recall)

Duyarlılık, gerçek pozitif tahminlerin (TP) sayısının toplam gerçek pozitif sınıf sayısına (TP+FN) bölünmesiyle hesaplanır. Bu değere recall veya gerçek pozitif oran (TPR) denir. En kötü duyarlılık değeri 0, en iyi değeri ise 1'dir.

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

2.8.4 Kesinlik (Precision)

Kesinlik, gerçek pozitif tahminlerin (TP) sayısının toplam pozitif tahminlere (TP+FP) bölünmesiyle hesaplanır. Buna pozitif tahmin değeri (Positive Predictive Value) denir. En kötü kesinlik 0, en iyisi 1 değerine sahiptir.

$$Precision = \frac{TP}{TP + FP}$$

2.8.5 F1-skoru (F1-Score)

F1 Score değeri, kesinlik ve duyarlılık değerleri arasındaki dengeyi sağlayan bir kriterdir. Eşit dağılıma sahip olmayan veri kümeleri olması durumunda doğruluk yerine F1 score kullanmak hatalı bir model seçiminin önüne geçecektir. Duyarlılık ve kesinlik değerlerinin harmonik ortalması alınarak F1 Score hesaplanmaktadır. Aritmetik yerine

harmonik ortalama kullanılmasının sebebi uç durumların da hesaba katılmasıdır. 0 ve 1 değerlerine sahip olunması durumunda F1 Score harmonik ortalama ile 0 olarak hesaplanırken, aritmetik ortalama ile 0.5 değeri ile hatalı bir şekilde hesaplanmaktadır.

$$F_1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

3. UYGULAMALAR

Bu bölümde meme kanseri veri seti üzerinde AdaBoost, k-Nearest Neighbor, Naïve Bayes, Random Forest ve Support Vector Machine sınıflandırma algoritmalarının uygulanışı ve sınıflandırma sonuçları verilmiştir. Tez çalışmasının temel amacı; meme kanseri veri seti üzerinde sınıflandırma doğruluğu açısından uygulanan beş makine öğrenmesi modelinden en iyi olan modeli bulmaktadır.

3.1 Meme Kanseri Veri Seti

Bu tez çalışmasında kullanılan WDBC veri seti California Üniversitesi – Irvine Machine Learning Repository'den alınmıştır (Wolberg ve diğ., 1995). Meme kanseri veri seti Wisconsin Üniversitesi Genel Cerrahi Bölümünden Dr William H. Wolberg ve Bilgisayar Bilimleri Bölümünden W. Nick Streetve Olvi L. Mangasarian tarafından 1995 yılında oluşturulmuştur. Veri setinde yarıçap, doku, çevre uzunluğu, alan, pürüzsüzlük, kompaktlık, konkavlık, konkav noktalar, simetri ve fraktal boyut olmak üzere 10 temel nitelik bulunmaktadır. Veri setindeki temel niteliklere ait 10 adet ortalama değer, 10 adet standart hata değeri ve 10 adet maksimum değer ile veri seti oluşturulmuştur. Hasta ID bilgisi ve sınıf etiketiyle birlikte veri seti toplam 32 niteliğe sahip 569 örnekten oluşmaktadır. Veri setinde bulunan 569 örnek içerisinde 357 adet iyi huylu (benign) ve 212 adet kötü huylu (malignant) meme kanseri hücresi bulunmaktadır. Bölümleme oranı olarak eğitim verisi için 0.67 kullanılmıştır. Bu tez çalışması kapsamında makine öğrenmesi sınıflandırıcılarının eğitim adımı için 381 adet (%67), test adımı için 188 adet (%33) meme kanseri verisi kullanılmıştır. Veri setinde eksik bilgiye sahip nitelik bulunmadığından kayıpsız bir veri setidir. Çizelge 3.1'de WDBC veri setine ait özelliklerin tip, ortalama, standart sapma ve maksimum niteliklerine ait minimum ve maksimum değerleri özet olarak verilmiştir.

Veri setinde bulunan yarıçap değeri, hücrelerin yarıçap bilgisidir. Doku, hücrelerin iç yüzeylerinin gri tonlamadaki değişim oranı, standart sapmasıdır. Çevre uzunluğu, herbir hücrenin yarıçapa bağlı çevre uzunluğu değeridir. Alan, hücreye ait yüzey alanıdır. Pürüzsüzlük, hücrenin etrafındaki komşu hücrelerin yarıçaplarının değeri, bir diğer

ifadeyle yarıçap uzunluklarındaki yerel varyasyondur. Kompaktlık, hücrenin çevre uzunluğunun karesinin hücre alanına bölünerek bir çıkarılmasıyla elde edilen hücre yoğunluk ortalamasıdır. Konkavlık, hücre çevresindeki girinti ve çıkıntıların büyüklük değerleridir. Konkav noktalar, hücre çevresindeki girinti ve çıkıntı nokta sayısıdır. Simetri, hücrelerin elips şekil değişikliği değeridir. Fraktal boyut, iç içe geçmiş düzensiz hücrelerin tüm normal hücrelere oranıdır.

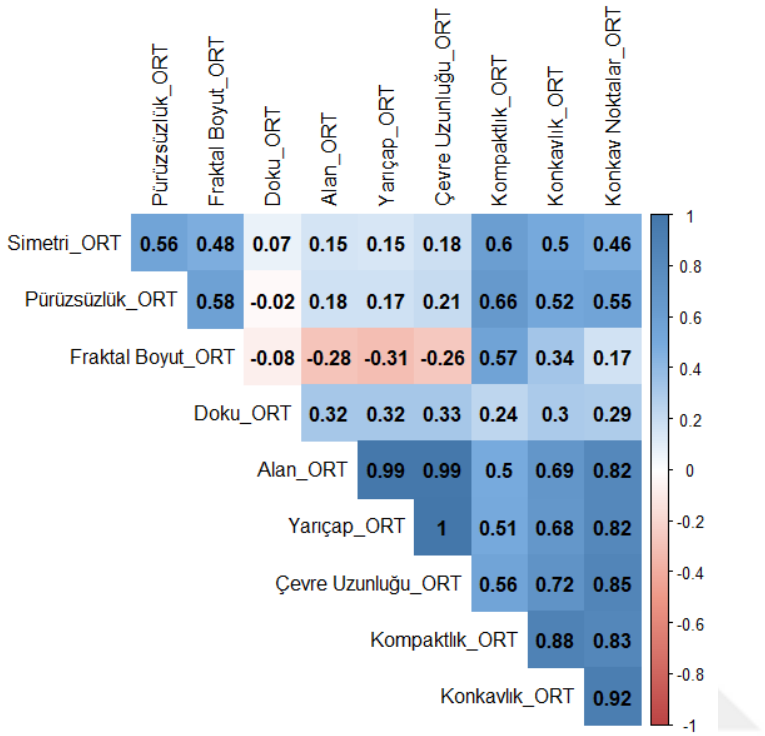
Çizelge 3.1: Veri seti WDBC'nin ortalama, standart sapma ve maksimum değerleri.

#	Özellik	Özellik Tipi	Ortalama	Std. Sapma	Maksimum
1	Hasta ID No.	Sayısal	–	–	–
2	Teşhis	Ayrık	–	–	–
3	Yarıçap	Sayısal	6.981 – 28.11	0.112 – 2.873	7.93 – 36.04
4	Doku	Sayısal	9.71 – 39.28	0.36 – 4.89	12.02 – 49.54
5	Çevre uzunluğu	Sayısal	43.79 – 188.5	0.76 – 21.98	50.41 – 251.20
6	Alan	Sayısal	143.5 – 2501.0	6.80 – 542.20	185.2 – 254.0
7	Pürüzsüzlük	Sayısal	0.053 – 0.163	0.002 – 0.031	0.071 – 0.223
8	Kompaktlık	Sayısal	0.019 – 0.345	0.002 – 0.135	0.027 – 1.058
9	Konkavlık	Sayısal	0.000 – 0.427	0.000 – 0.396	0.000 – 1.252
10	Konkav noktalar	Sayısal	0.000 – 0.201	0.000 – 0.053	0.000 – 0.291
11	Simetri	Sayısal	0.106 – 0.304	0.008 – 0.079	0.157 – 0.664
12	Fraktal boyut	Sayısal	0.050 – 0.097	0.001 – 0.030	0.055 – 0.208

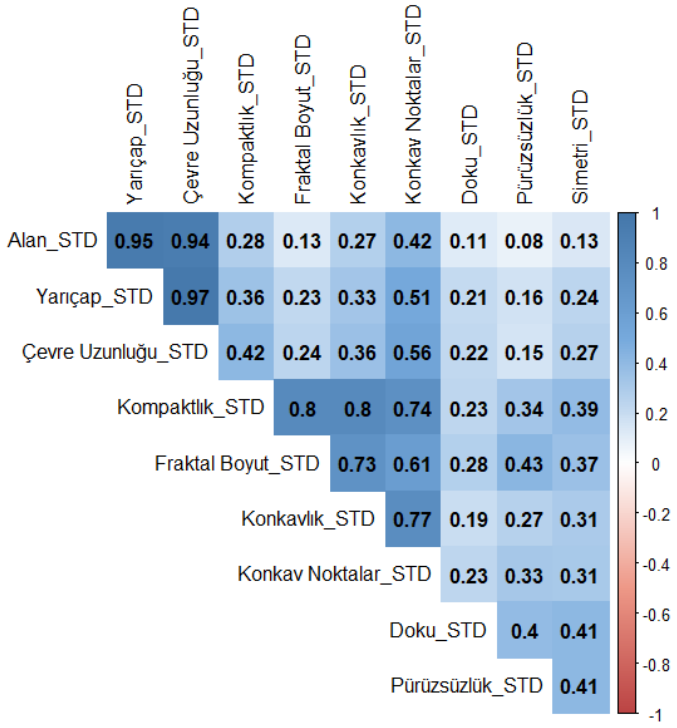
3.2 Meme Kanseri Veri Setine Ait Korelasyon Matrisi

Meme kanseri veri setindeki özellikler ortalama, standart sapma ve maksimum olmak üzere her biri 10 adet sütundan oluşan toplam üç bloğa ayrılmıştır. Korelasyon matrislerinde birbiriyle yüksek korelasyona sahip nitelikler +1 ile, düşük korelasyona sahip olan nitelikler -1 ile temsil edilmektedir. Renk koyu maviye yaklaştığında korelasyon değeri +1, renk koyu kırmızıya yaklaştığında korelasyon değeri -1 olacak şekilde matrisler oluşturulmuştur. Herhangi bir niteliğin kendisi ile olan korelasyonu her zaman +1 olacağından grafiklerde köşegen kaldırılmıştır.

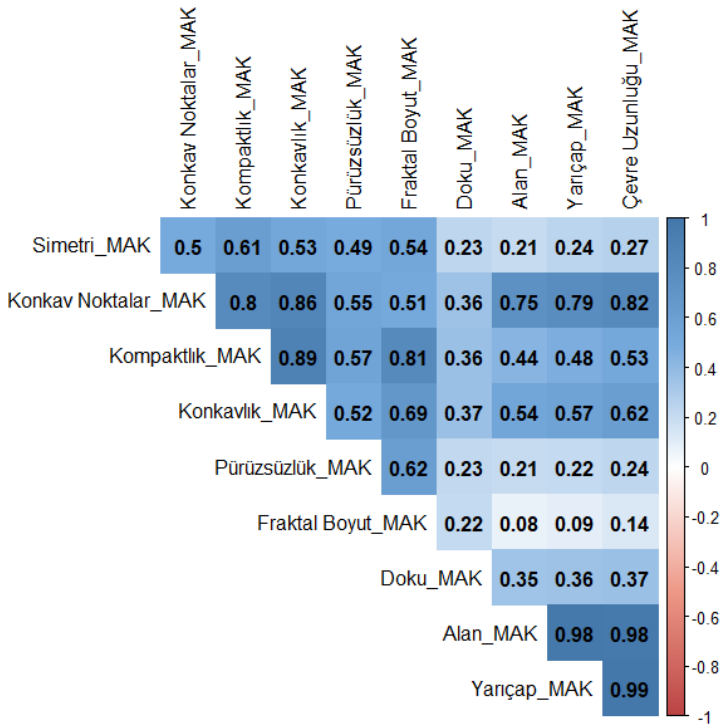
Şekil 3.1'de Yarıçap_ORT ve Çevre Uzunluğu_ORT korelasyon değeri, yuvarlak bir hücrenin çevresini hücrenin yarıçapına bağlı olarak değiştiğinden +1 olarak görülmektedir. Aynı şekilde alan ve çevre uzunluğu da birbirine sıkı derecede bağlı niteliklerdir. Fraktal boyutun yarıçap ile korelasyonu ise oldukça düşüktür.



Şekil 3.1: WDBC veri setinde ortalama değerlere ait korelasyon matrisi.



Şekil 3.2: WDBC Veri setinde standart sapma değerlerine ait korelasyon matrisi.



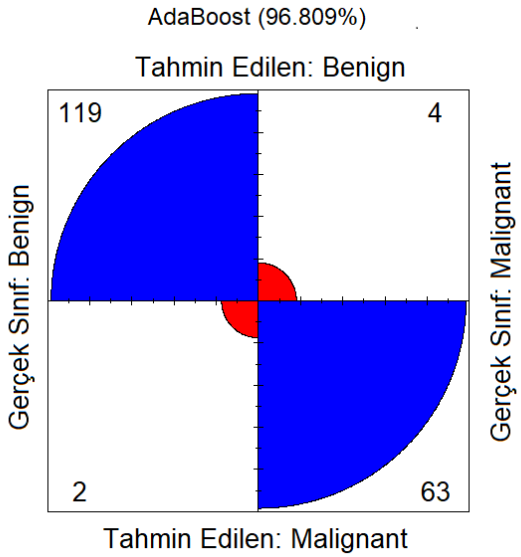
Şekil 3.3: WDBC Veri setinde maksimum değerlere ait korelasyon matrisi.

3.3 Sınıflandırma Yöntemlerinin Veri Setine Uygulanması

Bu bölümde WDBC veri seti üzerinde sınıflandırma yöntemlerinin uygulanış şekli verilmiştir. AdaBoost, kNN, Naïve Bayes, Random Forest ve SVM sınıflandırma algoritmalarının sonuçları ve doğruluk oranları karşılaştırılmıştır.

3.3.1 AB yönteminin uygulanması

AdaBoost algoritması zayıf tahmine sahip sınıflandırıcıların, yüksek sınıflandırma doğruluğuna sahip yeni bir sınıflandırıcıya dönüşümü prensibiyle çalışmaktadır. Şekil 3.4'de uygulama sonucunda AdaBoost algoritmasına ait hata matrisi verilmiştir. Hata matrisine bakıldığında gerçekte iyi huylu olan 119 değer; iyi huylu olarak doğru şekilde tahmin edilmiştir. Gerçek sınıfı kötü huylu olup iyi huylu olarak sınıflandırılan veri sayısı 4'dür. Gerçek sınıfı iyi huylu olup kötü huylu olarak sınıflandırılan 2 adet veri vardır. Gerçekte kötü huylu olan 63 değer; kötü huylu olarak doğru şekilde tahmin edilmiştir. AdaBoost sınıflandırma algoritmasına ait performans değerlendirme kriterleri Çizelge 3.2'de verilmiştir.



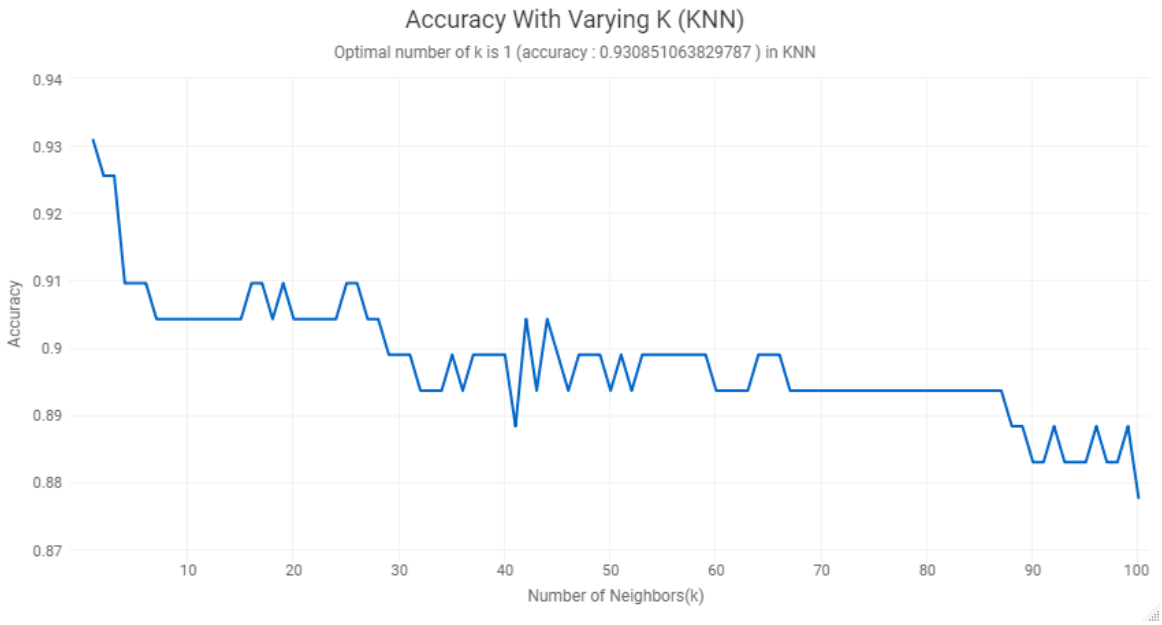
Şekil 3.4: AdaBoost sınıflandırma algoritmasına ait hata matrisi.

Çizelge 3.2: AdaBoost algoritmasına ait performans kriterleri.

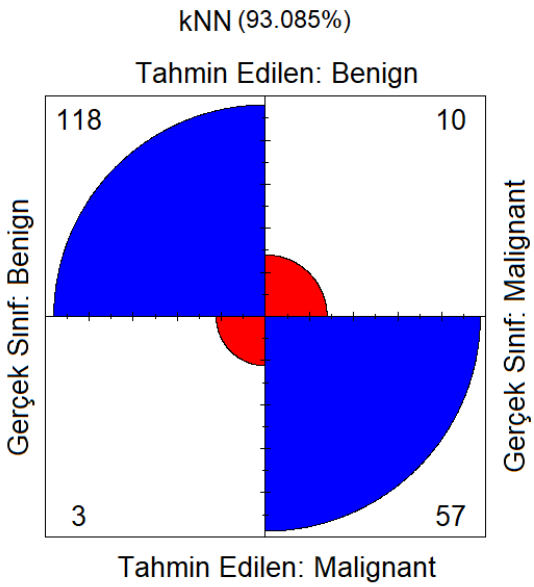
Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
AdaBoost	0.9681	0.9835	0.9675	0.9754

3.3.2 kNN yönteminin uygulanması

kNN algoritmasının uygulanmasında doğru k değerinin seçimi gerekmektedir. Temel olarak kNN algoritması bütün komşu veri noktalarına olan uzaklığı hesaplayarak birbirine en yakın k noktayı bulmaya çalışmaktadır. Doğru k değerini hesaplamak, algoritmanın etkin bir şekilde sınıflandırması için önem taşımaktadır. Şekil 3.5'de 1-100 arası k değerleri ve doğruluk oranları verilmiştir. Şekil 3.5'e göre k=1 değeri 0.9309 ile en yüksek doğruluk oranına sahiptir. kNN sınıflandırma modelinin başarımını daha detaylı görmek için ikinci bölümde bahsedilen hata matrisinden faylanmamız gerekmektedir. Şekil 3.6'da kNN algoritmasına ait hata matrisi verilmiştir. Hata matrisine bakıldığında gerçekte iyi huylu olan 118 değer; iyi huylu olarak doğru şekilde tahmin edilmiştir. Gerçek sınıfı kötü huylu olup iyi huylu olarak sınıflandırılan veri sayısı 10'dur. Gerçek sınıfı iyi huylu olup kötü huylu olarak sınıflandırılan 3 adet veri vardır. Gerçekte kötü huylu olan 57 değer; kötü huylu olarak doğru şekilde tahmin edilmiştir. kNN sınıflandırma algoritmasına ait performans değerlendirme kriterleri Çizelge 3.3'de verilmiştir. 0.9309 doğruluk ve 0.9478 F1 Skoru değerleri ile kNN algoritması sınıflandırma algoritmaları içerisinde en kötü sonucu vermiştir.



Şekil 3.5: Farklı k değerlerine ait doğruluk performansı.



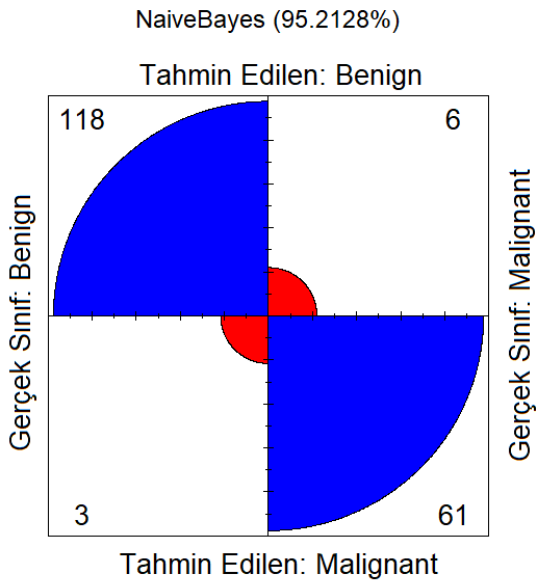
Şekil 3.6: kNN sınıflandırma algoritmasına ait hata matrisi.

Çizelge 3.3: kNN algoritmasına ait performans kriterleri.

Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
kNN	0.9309	0.9752	0.9219	0.9478

3.3.3 NB yönteminin uygulanması

Naïve Bayes algoritması daha önce girişine verilen veriler sonucu oluşturduğu model ile, yeni bir veri geldiğinde bu verinin iyi huylu veya kötü huylu bir kanser hücresi olma olasılığını hesaplamaktadır. NB hızlı bir algoritmadır. Bu tez çalışmasında Gauss dağılımına sahip NB algoritması kodlanmıştır. Şekil 3.7'de Naïve Bayes algoritmasına ait hata matrisi verilmiştir. Hata matrisine bakıldığında gerçekte iyi huylu olan 118 değer; iyi huylu olarak doğru şekilde tahmin edilmiştir. Gerçek sınıfı kötü huylu olup iyi huylu olarak sınıflandırılan veri sayısı 6'dır. Gerçek sınıfı iyi huylu olup kötü huylu olarak sınıflandırılan 3 adet veri vardır. Gerçekte kötü huylu olan 61 değer; kötü huylu olarak doğru şekilde tahmin edilmiştir. Naïve Bayes sınıflandırma algoritmasına ait performans değerlendirme kriterleri Çizelge 3.4'de verilmiştir.



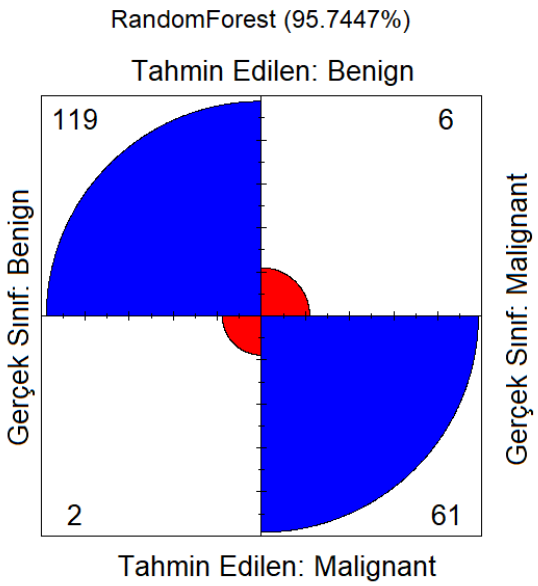
Şekil 3.7: Naïve Bayes sınıflandırma algoritmasına ait hata matrisi.

Çizelge 3.4: Naïve Bayes algoritmasına ait performans kriterleri.

Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
Naïve Bayes	0.9521	0.9752	0.9516	0.9633

3.3.4 RF yönteminin uygulanması

Şekil 3.8'de Random Forest algoritmasına ait hata matrisi verilmiştir. Hata matrisine bakıldığında gerçekte iyi huylu olan 119 değer; iyi huylu olarak doğru şekilde tahmin edilmiştir. Gerçek sınıfı kötü huylu olup iyi huylu olarak sınıflandırılan veri sayısı 6'dır. Gerçek sınıfı iyi huylu olup kötü huylu olarak sınıflandırılan 2 adet veri vardır. Gerçekte kötü huylu olan 61 değer; kötü huylu olarak doğru şekilde tahmin edilmiştir. RF sınıflandırma algoritmasına ait performans değerlendirme kriterleri Çizelge 3.5'de verilmiştir.



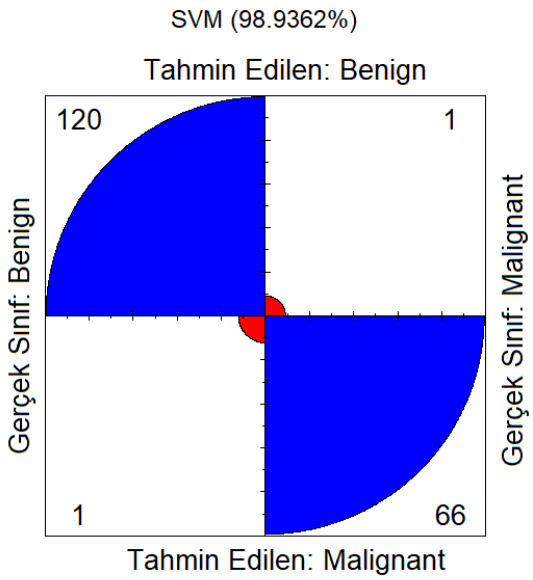
Şekil 3.8: Random Forest sınıflandırma algoritmasına ait hata matrisi.

Çizelge 3.5: Random Forest algoritmasına ait performans kriterleri.

Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
Random Forest	0.9574	0.9835	0.9520	0.9675

3.3.5 SVM yönteminin uygulanması

Şekil 3.9'da SVM algoritmasına ait hata matrisi verilmiştir. Hata matrisine bakıldığında gerçekte iyi huylu olan 120 değer; iyi huylu olarak doğru şekilde tahmin edilmiştir. Gerçek sınıfı kötü huylu olup iyi huylu olarak sınıflandırılan veri sayısı 1'dir. Gerçek sınıfı iyi huylu olup kötü huylu olarak sınıflandırılan 1 adet veri vardır. Gerçekte kötü huylu olan 66 değer; kötü huylu olarak doğru şekilde tahmin edilmiştir. SVM sınıflandırma algoritmasına ait performans değerlendirme kriterleri Çizelge 3.6'da verilmiştir.



Şekil 3.9: SVM sınıflandırma algoritmasına ait hata matrisi.

Çizelge 3.6: SVM algoritmasına ait performans kriterleri.

Algoritma	Doğruluk	Duyarlılık	Kesinlik	F1 Skoru
SVM	0.9894	0.9917	0.9917	0.9917

3.4 Uygulama Sonuçlarının Karşılaştırılması

Bu tez çalışmasının temel amacı, veri setine uygulanan sınıflandırma algoritmalarının sınıflandırma performansını karşılaştırmaktır. kNN algoritması %94.79 F1 skoru ile en kötü başarıma sahip sınıflandırma modelini oluşturmuştur. NB algoritması %96.33 F1 skoru ile dördüncü sıradadır. Random Forest algoritması %96.75 başarım ile üçüncü en iyi sınıflandırma yapan algoritma olarak bulunmuştur. AdaBoost algoritması %97.54 başarım ile en iyi ikinci sınıflandırıcı modeline sahiptir. SVM algoritması %99,17 F1 skoru ile en yüksek başarıma sahip algoritmadır. Çizelge 3.7'de Wisconsin Diagnostic Meme Kanseri (WDBC) veri seti üzerinde farklı sınıflandırma yöntemlerinin uygulanmasından elde edilen sonuçları özetlenmektedir.

Çizelge 3.7: Sınıflandırma algoritmalarına ait performans kriterleri.

Algoritma	**Doğruluk**	**Duyarlılık**	**Kesinlik**	**F1 Skoru**
AdaBoost	0.9681	0.9835	0.9675	0.9754
kNN	0.9309	0.9752	0.9219	0.9478
Naïve Bayes	0.9521	0.9752	0.9516	0.9633
Random Forest	0.9574	0.9835	0.9520	0.9675
SVM	**0.9894**	**0.9917**	**0.9917**	**0.9917**

4. SONUÇ ve GELECEK ÇALIŞMALAR

4.1 Sonuç

Bu tezde öncelikle makine öğrenmesi ve veri madenciliği kavramlarına, tekniklerine, uygulamalarına ve yapay zekâ ile olan ilişkilerine değinilmiştir. Tezin konusu olan meme kanseri ile ilgili genel bilgiler verilip kullanılacak olan algoritmalar hakkında bilgi verilmiştir.

Birinci bölümde, veri madenciliği, temel veriler ve kavramlar ve literatür taramasına yer verilmiştir. Veri madenciliği ve temel veri ve kavramlarının tanımlamaları; ayrıca literatür taraması kısmında ise bu tezde kullanılan sınıflandırma algoritmaları ile yapılan çalışmaların sonuçları ve değerlendirmeleri verilmektedir.

İkinci bölümde, materyal ve yöntemler hakkında detaylı tanımlama ve örneklendirme yapılmıştır. Tez için kullanılan sınıflandırma algoritmaları olan AdaBoost (AB), k-Nearest Neighbors (kNN), Naïve Bayes (NB), Random Forest (RF) ve Support Vector Machine (SVM) sınıflandırma algoritmaları için açıklamalar yapılmıştır.

Üçüncü bölümde ise, tez için kullanılan sınıflandırma algoritmalarının uygulamaları ve değerlendirmeleri bulunmaktadır. Bu sınıflandırma algoritmalarında girdi olarak meme kanseri hücresini, çıktı olarak da iyi huylu veya kötü huylu bir tümör olup olmadığını tahmin eden bir algoritma sistemi kullanıldı. Görseller ile de desteklenen üçüncü bölümde algoritmaların tahmin sonuçları bulunmaktadır. Bu algoritmaların kendi sonuçları değerlendirildi ve sonrasında bu sonuçlar kendi aralarında da değerlendirilerek hangi sınıflandırma algoritmasının daha iyi tahminde bulunduğu tespiti yapıldı. SVM algoritması performans yüzdesi olarak diğer algoritmalaradan daha başarılı bir yüzdeye sahiptir. Sınıflandırma algoritmalarından biri olan kNN algoritması ise SVM sınıflandırmasının aksine en kötü başarıma sahip olan bir sınıflandırma algoritması olmuştur.

Sonuç olarak tez çalışmasının temel amacı; WDBC veri seti üzerinde sınıflandırma yöntemlerini kullanarak meme kanseri hücresinin iyi huylu tümör ya da kötü huylu tümör olacak şekilde iki sınıflandırma tahmininde bulunulmasıdır. Bu amaca uygun olarak meme kanseri hakkında genel bilgiler ve veri madenciliğinin bu alandaki etkileri, kullanılan sınıflandırma algoritmaları, bu algoritmaların tanımları, sonuçları ve karşılaştırılmaları aşama aşama açıklanarak ayrıca görsellerle de desteklenerek ilerlenmiştir.

4.2 Gelecek Çalışmalar

Bu tez en çok kullanılan sınıflandırma algoritmalarının performansını ve karşılaştırmasını elde etmekte ve sunmaktadır. Gelecekteki çalışmalar için ise meme kanseri hücresinin sınıflandırma tahmininde Support Vector Machine (SVM) öğrenme modelinin geliştirilmesini önermekteyiz. Support Vector Machine (SVM) algoritması, karşılaştırdığımız diğer algoritma yapılarından daha iyi tahmin yüzdesi göstermektedir. Eğer bu algoritma yapısı geliştirilip yaygın olarak kullanılan bir algoritma yapısı haline gelirse; kansere yakalanan hastaların teşhis ve tedavi sürecinin olumlu olarak ilerleyeceği ve dolayısıyla kanser hastalığının yenilebilme oranının anlamlı ölçüde artacağı kanısındayız. İyi bir tümör sınıflandırması tahmini ile daha erken tedavi şartları oluşabilir ve böylelikle erken tedavi ile ölüm oranı riski azalabilir.

KAYNAKLAR

Ahmadi, M., Sharifi, A., Jafarian Fard, M., & Soleimani, N. (2021). Detection of brain lesion location in MRI images using convolutional neural network and robust PCA. *International Journal of Neuroscience, 0*(0), 1-12.https://doi.org/ 10.1080/00207454.2021.1883602

Akar, Ö., & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation, 1*(2), 105-112.

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102-4107.

Bhat, S., Saritha, M., Yatakunta, P. R., Naik, P. S., & Bhat, P. (2022). Chronic Kidney Disease Prediction Using Naive Bayesian Classifier and K-NN Machine-Learning Algorithms. *Research & Review: Machine Learning and Cloud Computing*, 1(2), 1-5.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, *9*, 13.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., & Li, K. (2016). A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Transactions on Parallel and Distributed Systems*, *28*(4), 919-933.

Christobel, A., & Sivaprakasam, Y. (2011). An empirical comparison of data mining classification methods. *International Journal of Computer Information Systems*, *3*(2), 24-28.

Collaborative Group on Hormonal Factors in Breast Cancer (2002). Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. *The Lancet, 360* (9328), 187-195. https://doi.org/10.1016/S0140-6736(02)09454-0

Collaborative Group on Hormonal Factors in Breast Cancer (2012). Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *The Lancet Oncology*, *13*(11), 1141-1151. https://doi.org/10.1016/S1470-2045(12)70425-4

Danacı, M., Çelik, M., & Akkaya, A. E. (2010). Veri madenciliği yöntemleri kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi. *Akıllı sistemlerde Yenilikler ve Uygulamaları Sempozyumu (ASYU'2010)*, 21-24.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC medical informatics and decision making*, *19*(1), 1-17.

Gc, S., Kasaudhan, R., Heo, T. K., & Choi, H. D. (2015). Variability measurement for breast cancer classification of mammographic masses. *Proceedings of the 2015 Conference on research in adaptive and convergent systems*, 177-182.

Gopalsamy, A., & Radha, B. (2022). Machine Learning-Based Ensemble Classifier Using Naïve Bayesian Tree with Logit Regression for the Prediction of Parkinson's Disease. Içinde N. Marriwala, C. C. Tripathi, S. Jain, & D. Kumar (Ed.), *Mobile Radio Communications and 5G Networks* (ss. 451-469). Springer Nature. https://doi.org/10.1007/978-981-16-7018-3_34

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician*, *63*(4), 308-319.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), 349-360.

Hazra, A., Mandal, S. K., & Gupta, A. (2016). Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and ensemble algorithms. *International Journal of Computer Applications*, *145*(2), 39-45.

Kachhia, P., & Rathod, D. (2022). Kidney Disease Detection Using Supervised Machine Learning Techniques. Içinde Y.-D. Zhang, T. Senjyu, C. So-In, & A. Joshi (Ed.), *Smart Trends in Computing and Communications* (ss. 357-365). Springer. https://doi.org/10.1007/978-981-16-4016-2_34

Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: A Systematic Review. *Acta Informatica Medica*, *24*(5), 364-369. https://doi.org/10.5455/aim.2016.24.364-369

Larose, D. T., Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd Edition, Wiley.

Mashudi, N. A., Rossli, S. A., Ahmad, N., & Noor, N. M. (2021). Comparison on Some Machine Learning Techniques in Breast Cancer Classification. *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 499-504.

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, *1*(6).

Nkikabahizi, C., Cheruiyot, W., & Kibe, A. (2022). Chaining Zscore and feature scaling methods to improve neural networks for classification. *Applied Soft Computing*, *123*, 108908. https://doi.org/10.1016/j.asoc.2022.108908

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, *26*(1), 217-222.

Pawlovsky, A. P., & Nagahashi, M. (2014). A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 189-192.

Rumgay, H., Shield, K., Charvat, H., Ferrari, P., Sornpaisarn, B., Obot, I., Islami, F., Lemmens, V. E. P. P., Rehm, J., & Soerjomataram, I. (2021). Global burden of cancer in 2020 attributable to alcohol consumption: A population-based study. *The Lancet Oncology*, *22*(8), 1071-1080. https://doi.org/10.1016/S1470-2045(21)00279-5

Seddik, A. F., & Shawky, D. M. (2015). Logistic regression model for breast cancer automatic diagnosis. *2015 SAI Intelligent Systems Conference (IntelliSys)*, 150-154.

Selvi, S. T., & Malmathanraj, R. (2006). Segmentation and SVM classification of mammograms. *2006 IEEE International Conference on Industrial Technology*, 905-910.

Sharma, A., & Suryawanshi, A. (2016). A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. *International Journal of Computer Applications*, *136*(6), 28-35.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(1), 7-33. https://doi.org/10.3322/caac.21708

Singh, H., & Raj, R. (2021). Breast Cancer Analysis and Prediction by Using Machine Learning. *International Journal of Research in Engineering and Science (IJRES)*, *9*(6), 69-73.

The Endogenous Hormones and Breast Cancer Collaborative Group (2002). Endogenous Sex Hormones and Breast Cancer in Postmenopausal Women: Reanalysis of Nine Prospective Studies. *JNCI: Journal of the National Cancer Institute*, *94*(8), 606-616. https://doi.org/10.1093/jnci/94.8.606

Todorovski, L., & Džeroski, S. (2003). Combining classifiers with meta decision trees. *Machine learning*, *50*(3), 223-249.

UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. (t.y.). Erişim tarihi 23 Mayıs 2022, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

Vapnik, V. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, *24*, 774-780.

Wilkinson, L., & Gathani, T. (2022). Understanding breast cancer as a global health concern. *The British Journal of Radiology*, *95*(1130), 20211033. https://doi.org/10.1259/bjr.20211033

Wolberg, W.H., Street W.N. & Mangasarian, O. L. (1995). *Breast Cancer Wisconsin (Diagnostic)*. UCI Machine Learning Repository.

Xu, J., & Zhang, H. (2014). Design and Application of Adaboost Algorithm Classifier. *Journal of Sichuan University of Science & Engineering (Natural Science Edition)*.

Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745-758.

Zheng, J., Lin, D., Gao, Z., Wang, S., He, M., & Fan, J. (2020). Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis. *IEEE Access*, *8*, 96946-96954. https://doi.org/10.1109/ACCESS.2020.2993536

ÖZGEÇMİŞ

Ad-Soyad **:** Ayça ACET

ÖĞRENİM DURUMU:

- **Lisans** **:** 2017, TOBB ETÜ Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölüm

MESLEKİ DENEYİM:

Mayıs 2016 - Ağustos 2016 İnönü Üniversitesi Bilgisayar Mühendisliği Bölümü

• Uçuş kontrol ve veri işleme laboratuvarlarında akademik çalışmalar gerçekleştirdim. Kinect'den alınan görüntü verileri ile eclipse üzerinde java dilini kullanarak alınan görüntüyü algılamak.

Mayıs 2017 - Ağustos 2017 Bilgi İletişim ve Teknolojileri Kurumu

• Spektrum İzleme Bölümü'nde çalışma deneyimi kazandım. Frekans ve baz istasyonları ile ilgili çalışmalar gerçekleştirdim. Baz istasyonlarından alınan verileri depolayan NoSQL veritabanı sistemi geliştirdim.

Eylül 2017 - Aralık 2017 Rakun Yazılım A.Ş

• TOBB ETÜ Veri Madenciliği Laboratuvarı'nda çalışmalarımı sürdürdüm. Natural Language Processing Tool'ları ile ilgili bir projede aktif yer aldım. Bu proje bir web uygulaması üzerinden analizler gerçekleştirmektedir.

Ağustos 2021

•İstanbul SAP MDG Ana Veri Yönetiminde Fiz Bilişim şirketinde çalışmaktayım.

YÜKSEK LİSANSTA TÜRETİLEN ÇALIŞMALAR
(Makaleler, Bildiriler, Patentler v.b.)

• Güldoğan, E., Tunç, Z., Acet, A. & Çolak, C. (2020). Performance evaluation of different artificial neural network models in the classification of type 2 diabetes mellitus. *The Journal of Cognitive Systems, 5* (1), 23-32.

• Acet, A, & Akkaya, A. E. (2020). A Deep Learning Image Classification Using Tensorflow for Optical Aviation Systems. *The Journal of Cognitive Systems, 5* (1), 1-4.