

← Adınızı soyadınızı giriniz

Tez kabul edildikten sonra yapılan **sabit ciltte sırt yazısı** bu şablona göre yazılacak. Yazılar tek satır olacak
Cilt sırtı yazıların yönü yukarıdan aşağıya
(sol yandaki gibi) olacak .



← Tez, Yüksek Lisans'sa, YÜKSEK LİSANS TEZİ;
Doktora ise DOKTORA TEZİ ifadesi kalacak

← Tez Sınavının yapılacağı yılı yazınız

**T.C.
İSTANBUL ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

(DOKTORA TEZİ)

**SOLUNUM SİSTEMİ HASTALIKLARININ
SINIFLANDIRILMASINDA MAKİNE ÖĞRENMESİ
YÖNTEMLERİNİN KULLANIMI**

ERKUT BOLAT

**DANIŞMAN
PROF. DR. ERAY YURTSEVEN**

**BİYOİSTATİSTİK
BİYOİSTATİSTİK VE TIP BİLİŞİMİ PROGRAMI**

İSTANBUL-2021

İTHAF

Sevgili aileme ithaf ediyorum.



TEŞEKKÜR

Tezimin hazırlanmasında yakın ilgi ve yardımlarıyla daima destek olan değerli hocam ve danışmanım Prof. Dr. Eray YURTSEVEN'e, bilgi ve tecrübesiyle yardımcı olan hocam Prof. Dr. Ahmet DİRİCAN'a içten teşekkürlerimi sunarım. Tezimde kullanılan uygulama verilerini sağlayan ve her konuda hiçbir zaman yardımını esirgemeyen değerli hocam Prof. Dr. SEDAT ALTIN'a, ayrıca değerli arkadaşım Dr. Hasan YILDIRIM'a ve çalışma arkadaşlarıma çok teşekkür ederim.



İÇİNDEKİLER

BEYAN	iii
İTHAF	iv
TEŞEKKÜR	v
İÇİNDEKİLER.....	vi
TABLolar LİSTESİ	viii
ŞEKİLLER LİSTESİ.....	ix
ÖZET	xii
ABSTRACT	xiii
1. GİRİŞ VE AMAÇ	1
2. GENEL BİLGİLER.....	4
2.1. Solunum Sistemi.....	4
2.2. Bazı Solunum Sistemi Hastalıkları.....	5
2.2.1. Astım	5
2.2.2. KOAH.....	5
2.2.3. Pnömoni (Zatürre)	6
2.3. Makine Öğrenmesi	7
2.3.1. Makine Öğrenmesi Metotları	10
2.3.1.1. Danışmanlı Öğrenme	10
2.3.1.2. Danışmansız Öğrenme.....	11
2.3.1.3. Yarı Danışmalı Öğrenme	12
2.3.1.4. Takviyeli Öğrenme	13
2.4. Sınıflandırma	13
2.5. Kullanılan Makine Öğrenmesi Yöntemleri	14
2.5.1. Naif Bayes	14
2.5.2. Lojistik Regresyon.....	15
2.5.3. Yapay Sinir Ağları.....	16
2.5.4. Destek Vektör Makineleri	19
2.5.5. K –En Yakın Komşu	23
2.5.6. Karar Ağaçları	24
2.5.7. Bagging (Torbalama).....	25
2.5.8. Rastgele Orman	26
2.5.9. Gradient Boosting Makinesi (GBM)	28

2.5.10. C5.0	29
2.6. Performans Ölçümü.....	29
2.6.1. Çapraz Doğrulama(Cross Validation)	30
2.6.2. Karışım Matrisi (Confusion Matrix)	31
2.6.3. İşlem Karakteristik Eğrisi (Receiver Operating Characteristics-ROC).....	32
2.7. Hiperparametrelerin Ayarlanması	34
3. GEREÇ VE YÖNTEM	36
4. BULGULAR	39
4.1. Değişkenler ve özellikleri.....	39
4.2. Grupların tanımlayıcı istatistikleri.....	39
4.3. Grupların karşılaştırılması ve Korelasyon analizi	43
4.3.1. Nitel değişkenlerin karşılaştırılması	43
4.3.2. Nicel değişkenlerin karşılaştırılması	45
4.3.2.1. Solunum fonksiyon testlerinin karşılaştırılması	45
4.3.2.2. Acil ve poliklinik başvuru sayılarının karşılaştırılması.....	47
4.4. Makine öğrenmesi algoritmalarının astım, KOAH, pnömoni hastalık grupları ve sağlıklı grup sınıflandırılmasında performanslarının karşılaştırmaları	48
4.5. Astım,KOAH, pnömoni grupları ve sağlıklı grup sınıflandırmasında değişken önemlilikleri	52
4.6. ROC Eğrileri.....	55
5. TARTIŞMA	58
KAYNAKLAR.....	61
İNTİHAL RAPORU İLK SAYFASI	66

TABLOLAR LİSTESİ

Tablo 4-1 Değişkenler ve açıklamalarının listesi	39
Tablo 4-2 Sağlıklı grup için nitel değişkenlerin tanımlayıcı istatistikleri	40
Tablo 4-3 Astım grubu için nitel değişkenlerin tanımlayıcı istatistikleri	40
Tablo 4-4 KOAH grubu için nitel değişkenlerin tanımlayıcı istatistikleri	40
Tablo 4-5 Pnömoni grubu için nitel değişkenlerin tanımlayıcı istatistikleri	41
Tablo 4-6 Sağlıklı grup için nicel değişkenlerin tanımlayıcı istatistikleri	41
Tablo 4-7 Astım grubu için nicel değişkenlerin tanımlayıcı istatistikleri	42
Tablo 4-8 KOAH grubu için nicel değişkenlerin tanımlayıcı istatistikleri	42
Tablo 4-9 Pnömoni grubu için nicel değişkenlerin tanımlayıcı istatistikleri	43
Tablo 4-10 Sağlıklı grup ile astım grubu karşılaştırılması	44
Tablo 4-11 Sağlıklı grup ile KOAH grubu karşılaştırılması	44
Tablo 4-12 Sağlıklı grup ile pnömoni grubu karşılaştırılması	45
Tablo 4-13 Sağlıklı grup ile astım grubu karşılaştırılması	45
Tablo 4-14 Sağlıklı grup ile KOAH grubu karşılaştırılması	46
Tablo 4-15 Sağlıklı grup ile pnömoni grubun karşılaştırılması	46
Tablo 4-16 Sağlıklı grup ile astım grubu karşılaştırılması	47
Tablo 4-17 Sağlıklı grup ile KOAH grubu karşılaştırılması	47
Tablo 4-18 Sağlıklı grup ile pnömoni grubu karşılaştırılması	47
Tablo 4-19 Sağlıklı grup ve astım grubu sınıflandırması için performans karşılaştırmaları...	48
Tablo 4-20 Sağlıklı grup ve KOAH grubu sınıflandırması için performans karşılaştırmaları.	49
Tablo 4-21 Sağlıklı grup ve pnömoni grubu sınıflandırması için performans karşılaştırmaları	50
Tablo 4-22 Astım, KOAH, pnömoni grupları ve sağlıklı grup sınıflandırması için performans karşılaştırmaları	51

ŞEKİLLER LİSTESİ

Şekil 2-1 İnsan vücudundaki solunum sistemi	4
Şekil 2-2 Normal ve astımlı akciğer	5
Şekil 2-3 Normal ve KOAH'lı akciğer	6
Şekil 2-4 Normal ve pnömonili akciğer	7
Şekil 2-5 Danışmalı Öğrenme	11
Şekil 2-6 Danışmanlı Öğrenme	12
Şekil 2-7 Yarı danışmanlı öğrenme	12
Şekil 2-8 Takviyeli öğrenme	13
Şekil 2-9 Doğrusal regresyon ve lojistik regresyon	15
Şekil 2-10 Bir yapay sinir hücresi yapısı	17
Şekil 2-11 Bir yapay sinir ağı	18
Şekil 2-12 Sınıfları ayıran doğru	19
Şekil 2-13 İki sınıflı bir veri kümesinde, hiperdüzlem, marjin ve destek vektörleri	20
Şekil 2-14 Doğrusal olmayan durumda DVM	21
Şekil 2-15 Eşleşme uzayı	22
Şekil 2-16 Örnek KNN algortiması	23
Şekil 2-17 İki dallı ve çok dallı karar ağaçları	25
Şekil 2-19 D veri kümesini $k=10$ için çapraz doğrulama işlemi	31
Şekil 2-20 Örnek ROC eğrisi	34
Şekil 4-1 Sağlıklı grup ve astım grubu sınıflandırması için Rastgele Orman algoritmasına dayalı değişkenlerin önemi	52
Şekil 4-2 Sağlıklı grup ve KOAH grubu sınıflandırması için GBM algoritmasına dayalı değişkenlerin önemi	53
Şekil 4-3 Sağlıklı grup ve pnömoni grubu sınıflandırması için Rastgele Orman algoritmasına dayalı değişkenlerin önemi	53
Şekil 4-4 Astım, KOAH, pnömoni grupları ve sağlıklı grup sınıflandırması için C5.0 algoritmasına dayalı değişkenlerin önemi	54
Şekil 4-5 Sağlıklı grup ve astım grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri	56
Şekil 4-6 Sağlıklı grup ve KOAH grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri	56

Şekil 4-7 Sağlıklı grup ve pnömoni grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri.....	57
---	----



SEMBOLLER / KISALTMALAR LİSTESİ

YSA: Yapay Sinir Ağları

DVM: Destek Vektör Makinesi

GBM: Gradient Boosting Makinesi

KNN: K- En Yakın Komşu

SFT: Solunum Fonksiyon Testi

RF: Random Forest

SVM: Support Vector Machine

CART: Classification and Regression Trees

CIT: Conditional Inference Trees



ÖZET

Bolat, E (2020), Solunum Sistemi Hastalıklarının Sınıflandırılmasında Makine Öğrenmesi Yöntemlerinin Kullanımı, İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik ABD. Doktora. İstanbul

En sık görülen solunum yolu hastalıkları olan astım, KOAH ve pnömoni (zatürre) gibi hastalıkların tanısı ve sınıflandırılması hayati öneme sahiptir. Son zamanlarda, bu önem nedeniyle sağlık verilerinin işlenmesine büyük ilgi gösterilmiştir. Makine öğrenmesi algoritmaları, bu amaç için önemli bir yere sahiptir. Önceki çalışmaların odağı astım, KOAH ve pnömoni hastalıklarının teşhisinde en öngörücü algoritmayı elde etmektir. Ancak, en önemli değişkenlerin belirlenmesi çok az ilgi görmüştür. Bu çalışmanın temel amacı, en önemli makine öğrenmesi algoritmalarını değişken önemine sahip sonuçlarla ayrıntılı olarak sunmaktır. Makine öğrenme algoritmalarına ve değişken önemi değerlendirmelerine dayalı kapsamlı bir karşılaştırma çalışması, doğruluk, precision(kesinlik), recall(duyarlılık), Kappa istatistikleri, F-ölçüsü, ROC eğrisi ve AUC değeri gibi farklı performans kriterleri dikkate alınarak yapılmıştır. Sağlıklı grup ile hastalıkların sınıflandırılmasında, astım hastalığı için rastgele orman (CART öğrenen ile) ve C5.0, KOAH için SVM (doğrusal olmayan çekirdek ile) ve GBM ve pnömoni hastalığı için Bagging ve Random Forest (CART öğrenen ile) en iyi algoritmalar olarak bulunmuştur. Ayrıca, tüm gupların birlikte sınıflandırılmasında en iyi test edilen algoritma GBM'dir. Değişken önemlilikleri dikkate alındığında sağlıklı grup ile hastalıklar arasında astım vakası sınıflandırma sonuçlarına göre, en önemli değişkenler FEV3, FVC, FEV1, MEF50 dır. KOAH'da MEF50, FVC, FEV3, PIF ve FEV1, pnömonide ise FVC, FEV3, MEF50 ve FEV1 dikkate değer değişkenlerdir. Tüm grupların birlikte sınıflandırılmasında ise neredeyse tıbbi değişkenlerin hepsi büyük öneme sahiptir. Önemli değişkenlerin sırası, korelasyon analizi ve istatistiksel anlamlılık testleriyle desteklenmiştir. Ek olarak, ROC eğrileri ve AUC değerleri görsel olarak benzer ve destekleyici sonuçlar vermiştir. Sonuç olarak, bu çalışma sağlık bilimleri ve makine öğrenimi uygulamaları konusunda uzman sistemler üzerinde çalışan uygulayıcılar ve araştırmacılar için değerli olacaktır.

Anahtar Kelimeler: astım, KOAH, pnömoni, hastalık tahmini, makine öğrenmesi, özellik seçimi.

ABSTRACT

Bolat. E. (2020). Use of Machine Learning Methods in Classification of Respiratory System Diseases, İstanbul University, Institute of Health Science, Department of Biostatistics. PhD Thesis. İstanbul.

The diagnosis and classification of asthma, COPD and pneumonia, which are the most common respiratory diseases, has been vital importance. Recently, there has been great interest for processing health data due to this importance. Machine learning algorithms have an important place for this purpose. The focus of previous studies has been to obtain the most predictive algorithm for diagnosing asthma, COPD and pneumonia diseases. However, the determination of the most significant variables has received little attention. The principal objective of this study is to present the most predictive machine learning algorithms with variable importance results in details. A comprehensive comparison study based on the machine learning algorithms and variable importance evaluations have been carried out by considering different performance criteria including accuracy, precision, recall, Kappa statistics, F-measure, ROC curve and AUC value. The findings have been indicated that Random forest (with CART learner), C5.0 for asthma disease and SVM (with non-linear kernel), GBM for COPD disease and Bagging ve Random Forest (with CART learner) for pneumonia have been found as the best algorithms. Also, GBM is the best tested algorithm for the classification of all groups together. Considering the variable significance, the most important variables were FEV3, FVC, FEV1, MEF50 according to the classification results of asthma cases between healthy group and diseases. MEF50, FVC, FEV3, PIF and FEV1 in COPD and FVC, FEV3, MEF50 and FEV1 in pneumonia are remarkable variables. Almost all medical variables are of great importance in the classification of all groups. The rank of important variables has been supported by correlation analysis and statistical significance tests. Additionally, ROC curves and AUC values have provided similar and supportive results as visually. Consequently, this study will be of value to practitioners and researchers studying on expert systems on health sciences and machine learning applications.

Keywords: asthma, COPD, pneumonia, disease prediction, machine learning, feature selection.

1. GİRİŞ VE AMAÇ

Son yıllarda, veri miktarı giderek artmaktadır. Big data dediğimiz büyük veri birikimi ile yapay zeka ve veri bilimi önem kazanmaktadır. Makine öğrenimi, geçmiş deneyimleri ve örnek verileri kullanarak bir performans ölçütünü optimize etmek için bilgisayarların programlandığı yapay zeka alanıdır. En basit anlamda veriden üretilen yazılımdır.

Veri boyutundaki artışın ve veri işlemenin gerekliliğinin bir sonucu olarak, makine öğrenmesi algoritmaları çok çeşitli alanlarda araştırmacılar tarafından kullanılmaktadır. Sağlık bilimleri, hastalık gruplarını tanımlamayı ve sınıflandırmayı amaçlayan en etkili ve faydalı uygulama alanlarından biridir. Özellikle, en sık görülen ve ölüm oranlarında da yüksek görünüme sahip olan solunum yolu hastalıkları olan astım, KOAH, pnömoni (zatürre) tanısı ve sınıflandırılması çok önemlidir.

Literatürdeki birçok çalışma, tanı ve tahmin için en önde gelen özellikleri saptamak veya önemli özellikleri tanımlamak yerine en öngörücü algoritmayı elde etmek için solunum verisine sahip eğitim sistemlerine odaklanmaktadır.

Uzman sistemlerde astım tanısı için farklı makine öğrenme algoritmaları araştırılmıştır (Prasad ve ark., 2011). Bu çalışmada, Bayesian ağı, bazı karar ağaçları yöntemlerini, otomatik ilişkisel bellek sinir ağları (AMNN) karşılaştırılmıştır. Elde edilen sonuçlara göre, Sinir Ağları diğer yöntemlerden daha iyi performans göstermiştir.

Zorlu salınım ölçümlerini kullanarak kronik obstrüktif akciğer hastalığının (KOAH) teşhisine yardımcı olmak için makine öğrenme algoritmalarına dayalı bir klinik karar destek sistemi geliştirilmiştir (Amaral ve ark.,2011). Bunun sonucunda, Lineer Bayes Normal Sınıflandırıcısı, K-en yakın komşusu, karar ağaçları, yapay sinir ağları ve destek vektörü makinelerine dayanan sınıflandırma algoritmalarının performansları en iyi sınıflandırıcı araştırması için karşılaştırılmıştır.

Solunum anomalileri uyarlanabilir bir yapay sinir ağı ve spirometriye dayalı bulanık mantık çıkarım sistemi kullanılarak sınıflandırılmıştır (Asaithambi ve ark., 2013).

Şüpheli astımın belirsizliği ile başa çıkmak için kural tabanlı bir uzman sistem önerilmiş ve oluşturulmuştur (Khalid ve ark., 2014).

Başka bir çalışmada, lojistik regresyondan rasgele ormanlara kadar, astım ve egzamayı tahmin etmek için doğrusal ve doğrusal olmayan makine öğrenme algoritmaları kullanılmıştır (Prosperi ve ark., 2014).

Mohktar ve ark., (2015) sınıflandırma ve regresyon ağacı (CART) algoritmasını kullanmışlardır. Önerilen algoritma, evde yaşayan KOAH seviyelerine sahip hastalardan kaydedilen (telehealth) ölçüm verileri kullanılarak doğrulanmıştır. CART algoritması, ev içi (telehealth) ölçüm verilerini % 71.8 doğruluk, % 80.4 özgüllük ve % 61.1 duyarlılık ile düşük riskli veya yüksek riskli bir kategoriye ayırmıştır.

Alevlenmelerin erken tespiti ve ardından triyaj için bir makine öğrenmesi temelli strateji sunulmuştur (Swaminathan ve ark., 2017). Bu çalışmanın uygulanmasında, denetlenen algoritmaları eğitmek için istatistiksel ve klinik olarak kapsamlı bir hasta vakaları setinde doktor görüşü değerlendirilmiştir. Modelin doğruluğu, örnek bir hasta doğrulama setinde aynı vakalar bir doktor paneline karşı değerlendirilmiştir.

Spathis ve Valamos (2017) bazı makine öğrenme yöntemleriyle astım ve KOAH'ı farklı kontrol gruplarıyla teşhis etmiştir. Çalışmalarında, rastgele orman algoritmasının, KOAH ve astım tanısında yüzde 97.7 ve yüzde 80.3 kesinliğe sahip diğer yöntemlere göre en iyi sınıflandırıcı olduğunu göstermiştir.

Badnjevic ve ark. (2018) astım, KOAH ve sağlıklı arasında ayrım yapabilen bir uzman tanı sistemi geliştirmiştir. Spahtis ve Valamos yöntemlerine ek olarak Gradyan Boosting kullanmışlardır. Gradyan Boosting, 98.33 ile en yüksek doğruluk yüzdesine sahip bulunmuştur.

Çeşitli yapay sinir ağ mimarileri ve bulanık mantık sınıflandırıcılarının yanı sıra her iki metoda dayanan uzman sistem yapıları kullanılarak solunum yolu hastalıklarının teşhisi için çeşitli çalışmalar yapılmıştır. Bu çalışmalar, çeşitli özelliklere sahip farklı veri setleri kullanılarak, astım ve KOAH da dahil olmak üzere çeşitli solunum yolu hastalıklarının teşhisi için farklı sinir ağı yapıları uygulamıştır (Price ve ark, 2001; Hanania ve ark, 2010; Badnjevic ve ark, 2018; Granulo ve ark., 2016; Gurbeta ve ark, 2018).

İlk defa, uyarlanabilir destek vektör makinesi algoritması astım ve KOAH'ı teşhis etmek için incelenmiştir (Yahyaoui ve Yumuşak, 2018). Sınıflama doğruluğu astım için % 98.45 ve KOAH için % 92.63 bulunmuştur.

Astım alevlenmelerini erken tahmin etmek için özel ayırma yapılmıştır (Finkelstein ve Jeong, 2018). Çalışma, tahmine dayalı modelleme, tabakalı eğitim veri

setlerinin hazırlanmasını, öngörücü özellik seçimini ve sonuçlanan sınıflandırıcıların değerlendirilmesini içerir. Naive Bayes sınıflandırıcısı, uyarlanabilir Bayesian ağı ve destek vektörü makineleri, sırasıyla 0.80, 1.00, 0.84 duyarlılık, 0.77, 1.00, 0.80 özgüllük, ve 0.77, 1.00, 0.80 doğruluk ile astım alevlenmesini tahmin edebilmiştir.

Son olarak, normal ve KOAH hastalarının solunum ses analizi temel alınarak sınıflandırılması yapılmıştır (Haider ve ark.,2019). Solunum ses parametrelerinin yanı sıra spirometri parametrelerine göre sınıflandırma değerlendirilmiştir. Ayrıca, Lineer ve lineer olmayan diskriminant sınıflandırıcıları farklı olarak incelenmiştir. Lineer diskriminant sınıflandırıcı, % 94.4'lük bir sınıflandırma doğruluğu elde etmiştir.

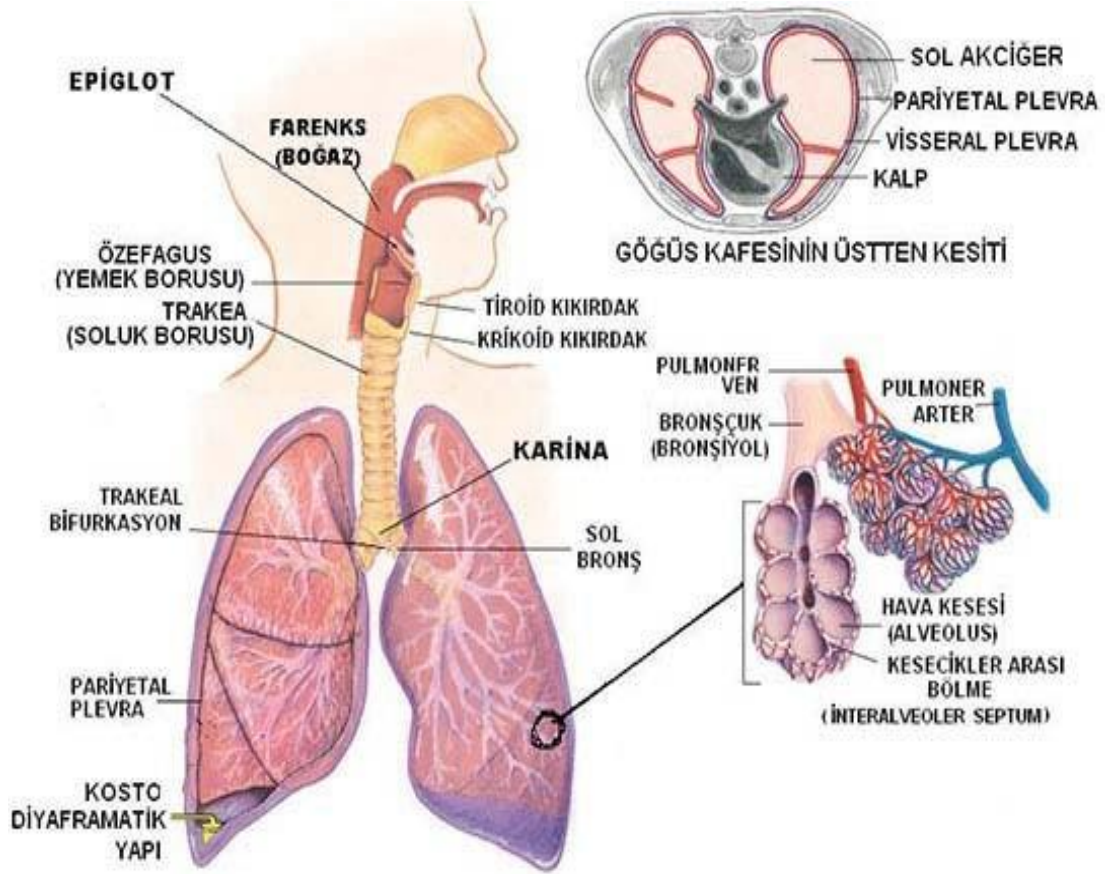
Bu çalışma, makine öğrenmesi algoritmaları ile solunum sistemi hastalıklarının erken tahmini ve teşhisi için astım, KOAH, pnömoni (zatürre) ve sağlıklı kontrol sınıflamasını karakterize eden faktörlere odaklanmaktadır. Astım, KOAH, pnömoni (zatürre) hastalıklarının teşhisine ilişkin mevcut literatürle karşılaştırıldığında, bu çalışmanın ana katkıları ve amaçları aşağıdaki gibi özetlenebilir.

1. Hiperparametre uzayı tarama da dahil olmak üzere en başarılı ve iyi bilinen makine öğrenme algoritmalarına dayalı kapsamlı karşılaştırmalar sağlanmıştır.
2. Astım, KOAH, pnömoni hastalıkları için gerçek ve güncel veri setleri dikkate alınmıştır.
3. Yukarıda belirtilen hastalıkları etkileyen en önemli belirleyiciler ayrı ayrı belirlenmiş ve sıralanmıştır. Bulguların bu alanda çalışan araştırmacılar için bazı yararlı bilgiler sağlayabileceği düşünülmektedir.

2. GENEL BİLGİLER

2.1. Solunum Sistemi

Solunum sistemi, solunum yolları (burun, farinks, larinks, trakea, bronşlar), akciğerler, mediastinum, plevra ve solunum kasları (diafragma ve diğerleri) ile bu yapılarla ilgili afferent ve efferent sinirlerden oluşur (Guyton ve Hall, 2010).



Şekil 2-1 İnsan vücudundaki solunum sistemi

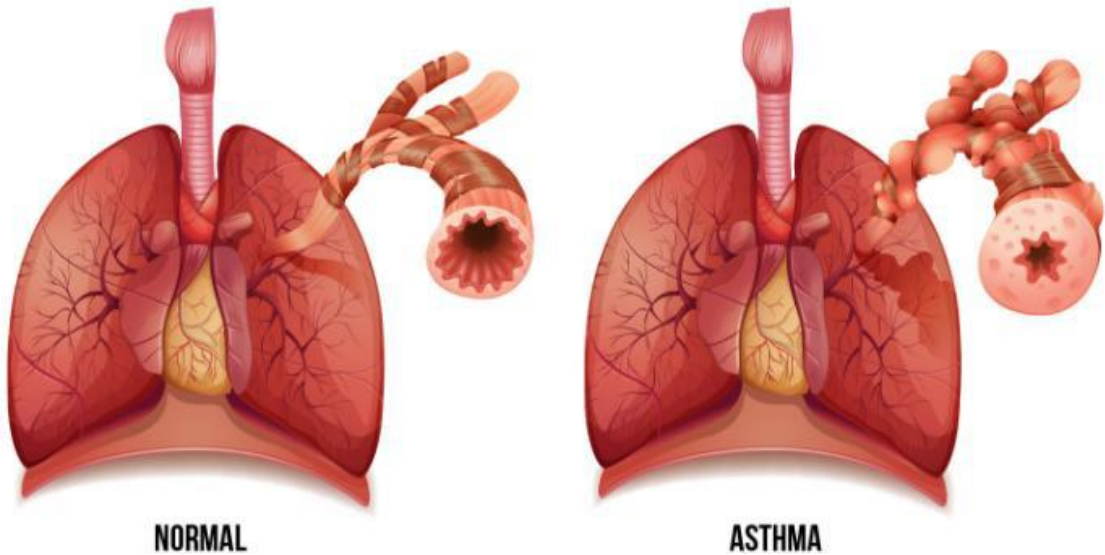
Akciğerlerin içerisi alveol denilen küçük hava keseciklerinden oluşmaktadır. Bu kesecikler üzerinde bulunan kapiller damarlar aracılığıyla kan ve hava arasında gaz alışverişi sağlanır. Akciğerler kendi başına kontraktil yapıda olmadığından, çevresindeki dokular aracılığıyla hareket edebilmektedir. Solunum sistemini oluşturan organların bir özelliği de dış kısımlarında kıkırdak doku bulunmasıdır. Bunun sayesinde ise tamamen kapanmaları önlenmiş olup içlerinde sürekli hava bulunur.

2.2. Bazı Solunum Sistemi Hastalıkları

2.2.1. Astım

Astım hastalığı solunum yollarının kronik bir iltihabıdır. Astım, insan vücudunda ilerlemesine bağlı olarak günde veya haftada birkaç kez çeşitli sürelerde gözlemlenen ataklarla her yaşta insanı etkileyen ve dünyada yaygın bir şekilde görülen bir hastalıktır.

Dünya çapında, 300 milyondan fazla insanın astımı olduğu tahmin edilmektedir (GINA, 2018). 2025'te, muhtemelen 100 milyondan fazla astım hastası olacak (Masoli, 2004). Aynı zamanda, Dünya Sağlık Örgütü tahminlerine göre, her yıl yaklaşık 250.000 kişi astımdan erken ölmektedir (WHO, 2007). Doktorlar astımın nedenlerinin iyi tespit edilmediğini doğrulasa da, hastalığın genetik ve çevresel faktörlerin karmaşık bir etkileşiminin sonucu olduğunu belirtiyorlar. Ayrıca, çevre kirliliği ve bazı alerjiler astımın nedenleri arasındadır.



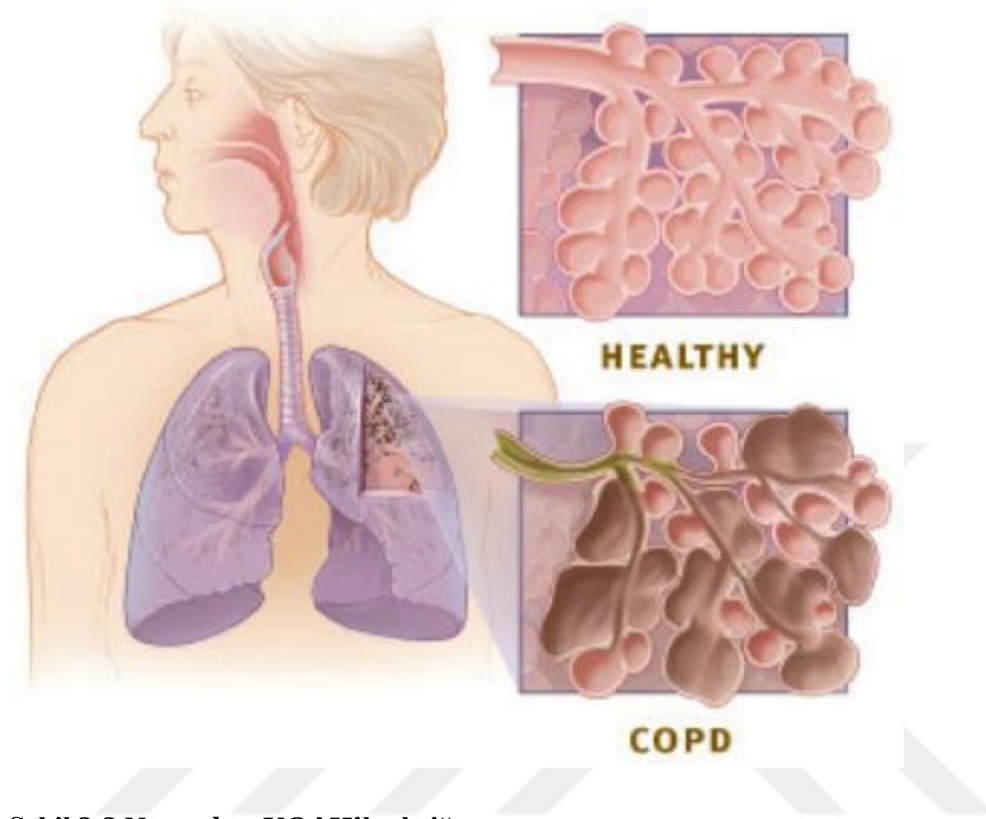
Şekil 2-2 Normal ve astımlı akciğer

2.2.2. KOAH

KOAH (Kronik Obstrüktif Akciğer Hastalığı) ; zararlı gaz ve partiküllere karşı havayolları ve akciğerin artmış kronik inflamatuvar yanıtı ile ilişkili ve genellikle ilerleyici özellikteki kalıcı hava akımı kısıtlaması ile karakterize, yaygın, önlenebilir ve tedavi edilebilir bir hastalıktır. Alevlenmeler ve ek hastalıklar KOAH'ın şiddetine katkıda bulunur.

Dünya çapında yaklaşık 330 milyon kişinin KOAH'ı olduğu tahmin edilmektedir (Eisner ve ark., 2011). KOAH, son zamanlarda ölümlerin hızla artışına neden olan kronik

hastalıklardandır (Afonso ve ark., 2011) ve 15 yılda KOAH'ın önde gelen ölüm nedeni olması beklenmektedir (Alvan, 2010).



Şekil 2-3 Normal ve KOAH'lı akciğer

2.2.3. Pnömoni (Zatürre)

Pnömoni (zatürre) göğüs enfeksiyonunun sebep olduğu salgın bir hastalıktır. Pnömoni, bir mikroorganizmanın akciğerlere ulaşması ve çoğalması sonucu akciğerlerde oluşan yangısal süreçtir. Bu hastalık, özellikle iki ay ile yedi yaş arasındaki çocuklar başta olmak üzere her yaşta insanı etkileyebilmektedir. Hastalık, genellikle grip şeklinde meydana gelmektedir. Pnömoni tüm dünya üzerinde ve ülkemizde önemli ölçüde mortalite nedenidir. Ulusal Hastalık Yüğü Raporu' na göre, halen ülkemizde en yüksek mortaliteye neden olan beşinci hastalıktır. Yaş grupları ile değerlendirildiğinde, 60 yaş üstü grupta 6. sırada olmasına karşın, 0-14 yaş grubunda ikinci sırada yer almaktadır. WHO raporuna göre pnömoni sebebiyle 2015 yılında yaklaşık olarak 922 bin çocuk kaybedilmiştir.



Şekil 2-4 Normal ve pnömonili akciğer

2.3. Makine Öğrenmesi

Öğrenme, ilk olarak bir süreçtir ve süreç içerisinde kişinin davranışlarında değişiklik olması beklenir. Gerçekleşen bu değişiklikler olumlu olabileceği gibi olumsuz da gerçekleşebilir. Süreç içerisinde birey edindiği tecrübeler ışığında mevcut davranışlarını değiştirebileceği gibi yeni davranışlarda edinebilir. Öğrenme kelimesi sözcük anlamı ile ifade edildiğinde; öğrenim, talimat ve deneyim yoluyla beceri, anlama ve bilgi kazanımı olarak tanımlanabilir (Nilsson,1996).

Makine Öğrenmesi, sistemin geçmişteki deneyimlerinden elde edilen öğrenmelerini kullanarak bir model oluşturmaya ve gelecekte karşılaşacağı durumlar karşısında bir tahminde bulunmasını sağlayan bir yapay zeka alanıdır. Bilgisayarın eldeki veriler üzerinden geliştirilen bir model yardımıyla oluşturdukları öğrenmeleri, ileride karşılaştıkları yeni veriler üzerinde kullanarak kararlar verebilmesini ve ilgili problemlere çözümler üretmesi olarak tanımlanabilir (Öğücü, 2006). Bilgisayar oyunları ve yapay zeka alanında öncü kişilerden sayılan Arthur Lee Samuel 1959 yılında makine öğrenmesini, bilgisayarların yeniden programlamaya ihtiyaç duymadan görev yapmasını sağlayan bilim olarak tanımlamıştır (Samuel,1996). Daha resmi tanımı ise

Michalski ve arkadaşları yapmıştır. Bir bilgisayar yazılımının, D deneyimini kullanarak istenilen bir görev olan G yi gerçekleştirmek için performans P ölçüsü ile eğitilmesidir. Burada performans P ölçütü ile oluşturulan görev G, deneyim D'yi geliştirmektir (Michalski ve ark., 1986).

Makine öğrenmesi, bilgisayarların öğrenmelerine yönelik yöntemleri içeren, veri madenciliği ile oldukça ilgili geniş bir bilgisayar bilimi alanıdır (Carbonell ve ark.,1983). Makine öğrenmesi ile bilgisayar yazılımlarının, geçmişteki tecrübelerle yani önceden yaşanan durumlarda oluşturdukları model yardımıyla yeni durumları öğrenmeleri sağlanır. Eğitilmek istenen yazılım, öncelikle bir örnek üzerinden bazı bilgiler çıkarılarak bir model oluşturulur. Daha sonra gelen diğer örnekler yardımıyla model üzerinde düzeltmeler ve iyileştirmeler gerçekleştirilir. Gerçekleştirilen bu işleme tecrübeden öğrenme de denilebilir (Öztemel,2003). Makine öğrenmesi bir yapay zeka alanıdır ve sadece veriler üzerinde çalıştırılmaz. Makine öğrenmesinin amacı elimizdeki veri içerisindeki değerli bilgiyi ortaya çıkarmak için algoritmaların iyileştirilmesini sağlamaktır.

Bir makine öğrenmesi görevinin gerçekleştirilmesi için beş temel adımı şöyle ifade edebiliriz (Lantz, 2015).

Veri toplama: Çeşitli ortamlardan (veri tabanları, web sayfaları ve metin dosyaları vb.) ham verilerin toplanmasıdır. Geçmişte kayıtlı verilerin toplanması gelecekte öğrenmelerin temelini oluşturur. Verilerin çeşitliliği, yoğunluğu ve hacmi ne kadar iyi olursa makine öğrenmesi için öğrenme o derece iyi olacaktır.

Verilerin Hazırlanması: Verilerin eğitim modelini oluşturmak amacıyla sisteme verilmeden önce yapılan ön işlemleri kapsar. Örneğin bir doğal işleme için durma kelimelerinin veriden arındırılması, eğer gerekli görülüyorsa noktalama işaretlerinin veriden ayıklanması gibi işlemleri kapsar.

Modeli eğitmek: Modeli oluşturmak için uygun algoritmanın seçilmesini ve verileri temsil etmek için uygun bir model oluşturma sürecini kapsar. Eğer danışmanlı eğitim yapılacak ise bu aşamada veriler eğitim ve test kümelerine bölünerek ikiye ayrılır. Eğitim kümesi ile sistem çalıştırılıp test aşamasında değerlendirilmek üzere verilere uygun bir model oluşturulur.

Modelin Değerlendirilmesi: Sistemin doğruluğunun test edilmesi için bir önceki aşamada geliştirilen model test kümesi üzerine uygulanır ve sonuçlar kaydedilir. Bir modelin

doğruluğunun en iyi şekilde test edilmesi için modelin daha önce hiç görmediği veriler üzerindeki başarısı ölçülmelidir.

Performansı Artırmak: Bu adımda tamamen farklı bir modelin seçilmesi veya modelin verimliliğinin artırılması için daha fazla değişken sunmak gerekebilir. Bu nedenle önemli ölçüde veri toplanması ve model oluşturmak için ön işleminden geçmesi gerekebilir.

Bilgisayarlar, makine öğrenmesi algoritmaları ile toplumun işgücüne katkı sağladıkları gibi toplumun beyin gücüne de katkılar sağlamaya başlamıştır. Çeşitli alanlarda ve çok miktardaki verinin işlenmesi, analiz edilmesi ve analiz sonuçlarının yorumlanması makine öğrenmesi algoritmaları ile kolay hale gelmiştir (Alpaydın,2004). Makine öğrenmesi algoritmaları, tümevarım yöntemini kullanmaktadır. Tümevarım yönteminin amacı parçaları kullanarak bütüne ulaşmaya çalışmaktır. Parçalardan bütüne ulaşmak için oluşturulan çıkarımlar, bir sonraki adımda tahminde bulunmak için veya bir tanımlama yapmak için kullanılacaktır. Elimizdeki veri, geleceğe yönelik bir bilgiyi tahmin etmek için kullanılacaksa buna tahmin edici model denmektedir. Tahmin edici modeller eğitim ve test aşaması olmak üzere iki bölümden oluşmaktadır. Eğitim safhasında, elimizdeki eğitim verisi denilen ve etiketli olan veriler kullanılarak bir model oluşturulur. Eğitim verisinin bir kısmı test aşamasında kullanılmak üzere ayrılır. Eğitim verisi kullanarak sistemin bir model oluşturması sağlanır. Modelin oluşturulmasının ardından eğitim verisi içinden test için ayrılan veri kullanılarak sistem test edilir.

Makine öğrenmesinin amaçlarını üç farklı yönden ele alabiliriz:

1. Hedef Tabanlı Çalışmalar: Öğrenme sistemini geliştirmek ve analizi belirlenmiş görevleri yerine getirmek için gerçekleştirilir. Bu yaklaşım “makine öğrenmesine” mühendis yaklaşımı olarak tanımlanır.
2. Bilişsel Simülasyon: İnsanın öğrenme sürecini araştırıp bilgisayar ortamında benzetimini gerçekleştirmek olarak tanımlanmıştır. Bu ise makine öğrenmesine bilişsel bir modelleme yaklaşımıdır.
3. Teorik Analiz: Uygulama alanlarından bağımsız olarak teorik olabilecek öğrenme metotları ve algoritmaları incelemek içindir.

Makine öğrenmesiyle ilgilenen kişiler önerilen üç yaklaşımdan sadece birisini veya hibrit bir şekilde kullanarak çalışmalar gerçekleştirmiştir. (Bilgin,2108). Nilsson, makine öğrenmesinin önemini şu şekilde sıralamıştır (Nilsson, 1996).

- Bazı görevlerde girdi/çıkış belirlense bile arasındaki ilişki belirtilemeyebilir. Bu durumda, makinenin kendi iç yapısını ayarlayarak, büyük veri yığınlarından giriş/çıkış fonksiyonunu bulup, arasındaki ilişkiyi tahmin etmesi beklenir.
- Büyük veri yığını içinde gizli kalmış önemli ilişkiler ve bağıntılar olabilir. Makine öğrenmesi metotlarından, bu ilişkileri seçip çıkarmak için yararlanılır. Bu konu veri madenciliği olarak adlandırılır.
- Bazı görevlerde bilgi miktarı, insanın kodlaması için oldukça fazla olabilir. Bu gibi durumda makine insanın yapabileceğinden fazlasını yapabilir.
- Makine, değişikliklere kolayca uyum sağlayabilir. Bazı görevlerde bilgiler değişebilir. Bu durumlarda, yapay zeka istemini tekrar tasarlamak pratik değildir. Makine öğrenmesi metotlarını kullanarak bu gibi değişiklikler gözlenebilir.

2.3.1. Makine Öğrenmesi Metotları

Makine öğrenmesi algoritmaları eldeki verileri kullanarak en iyi modeli oluşturmaya çalışır. Yeni veriler geldiğinde önceki verilerden oluşturulmuş olan modeli kullanarak yeni gelen verileri analiz etmeye çalışırlar. Eldeki verilerin işlenerek içerisinden anlamlı ve kullanışlı bilgilerin oluşturulmasına veri madenciliği denilmektedir (Çentik, 2013).

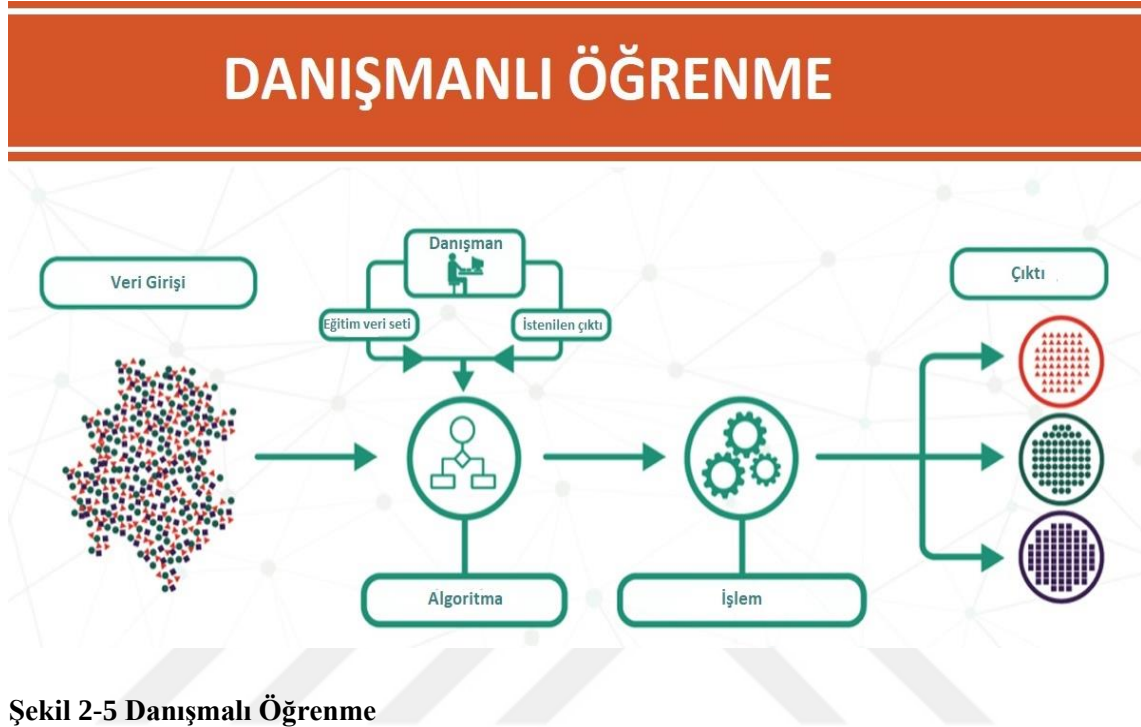
2.3.1.1. Danışmanlı Öğrenme

Danışmanlı Öğrenme’de sistemin öğrenme işlemini gerçekleştirilebilmesi için bir danışman gerekmektedir. Danışmanın görevi, öğrenilmesi istenen model için ilgili eğitim olayları ve olaylarla ilgili örnekleri Girdi/Çıkış sisteme verilmesidir. Danışmanlı öğrenme ile, girişler ile çıkışlar eşleştirilerek işlemin sistem tarafından öğrenilmesi ya da ilgili modelin oluşturulması sağlanmaktadır (Nilsson, 1996). Danışmanlı öğrenme çalışma prensibi Şekil 2.5’te görülmektedir.

Danışmanlı öğrenme modelinde problem, sınıflandırma problemi olarak ele alınmakta ve eğitilmiş sistem tarafından oluşturulan model kullanılarak test seti üzerinde tahmin ve tanıma yapmak üzere kullanılmaktadır (Chao, 2011). En iyi bilinen ve kullanılan makine öğrenme algoritmaları destek vektör makinesi, yapay sinir ağları, lojistik regresyon, basit bayes, multinom basit bayes, k-En yakın komşu, rasgele orman ve karar ağaçları algoritmaları olarak verilebilir (Bilgin, 2018).

Sınıflandırma problemlerinde kullanılan yöntemler de uygulanan problem veya beklentisindeki farklılığa göre değişiklik gösterebilmektedir. Bu nedenle sınıflandırma sonuçlarının tek etiketli veya çok etiketli olarak üretilmesi gerekebilmektedir. Tek etiketli

sınıflandırmada test örneğine mevcut etiket kümesinden sadece bir etiket atanırken, çok etiketli sınıflandırmada birden fazla etiket atanmaktadır. Ayrıca tek etiket sınıflandırmanın özel bir durumu olan ikili sınıflandırmada ise mevcut iki kategori arasından tek bir etiket atanarak sınıflandırma yapılmaktadır (Sebastiani, 2002).

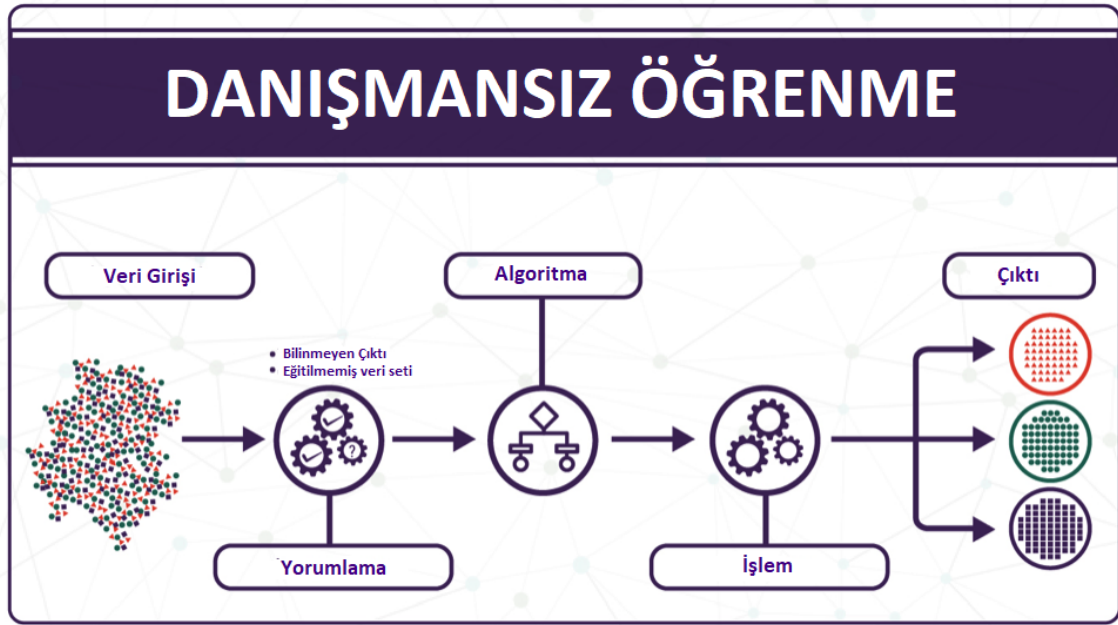


Şekil 2-5 Danışmanlı Öğrenme

2.3.1.2. Danışmansız Öğrenme

Danışmansız öğrenmede sistemin öğrenme işlemini gerçekleştirebilmesi için danışmanlı öğrenmeden farklı olarak bir danışmana ihtiyacı yoktur. Sisteme çıkışlardan bağımsız olarak sadece girişler verilmektedir. Kullanılan algoritmalar sisteme girilen örnekler arasındaki ilişkilerin bulunması sağlayacak bir model oluşturmaya çalışır (Nilsson,1996). Danışmansız sistemin çıkış hakkında herhangi bilgisi yoktur. Danışmansız öğrenme çalışması Şekil 2.6’ da görülmektedir.

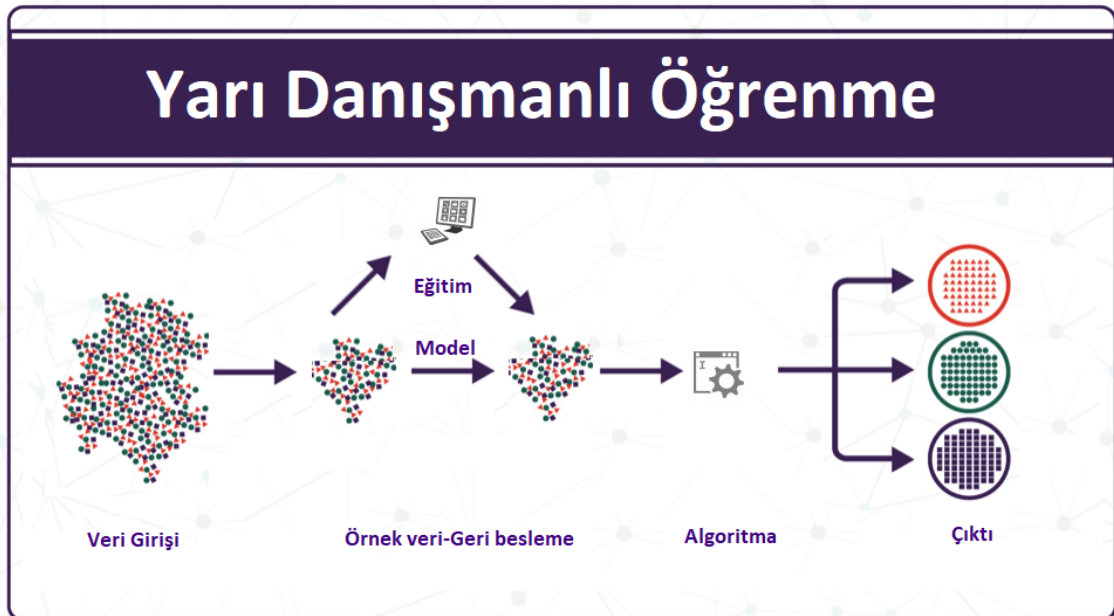
Danışmansız öğrenmenin amacı veri setindeki örneklerin çıkışları bilinmediği için tanıma veya sınıflandırma değildir. Danışmansız öğrenmenin kullanılmasının amacı, kümeleme, olasılık yoğunluk tahmin, öznitelikler arasındaki ilişkilerin bulunması ve boyut indirgeme gibi alanlardır. İstenirse danışmansız öğrenme algoritması ile elde edilen model ve sonuçlar danışmanlı öğrenme için de kullanılabilir (Chao, 2011). Bunlarında yanı sıra parçalayıcı ve hiyerarşik kümeleme algoritmaları da danışmansız öğrenme modeli oluşturmak üzere kullanılan algoritmalar (Özgür, 2004).



Şekil 2-6 Danışmanlı Öğrenme

2.3.1.3. Yarı Danışmanlı Öğrenme

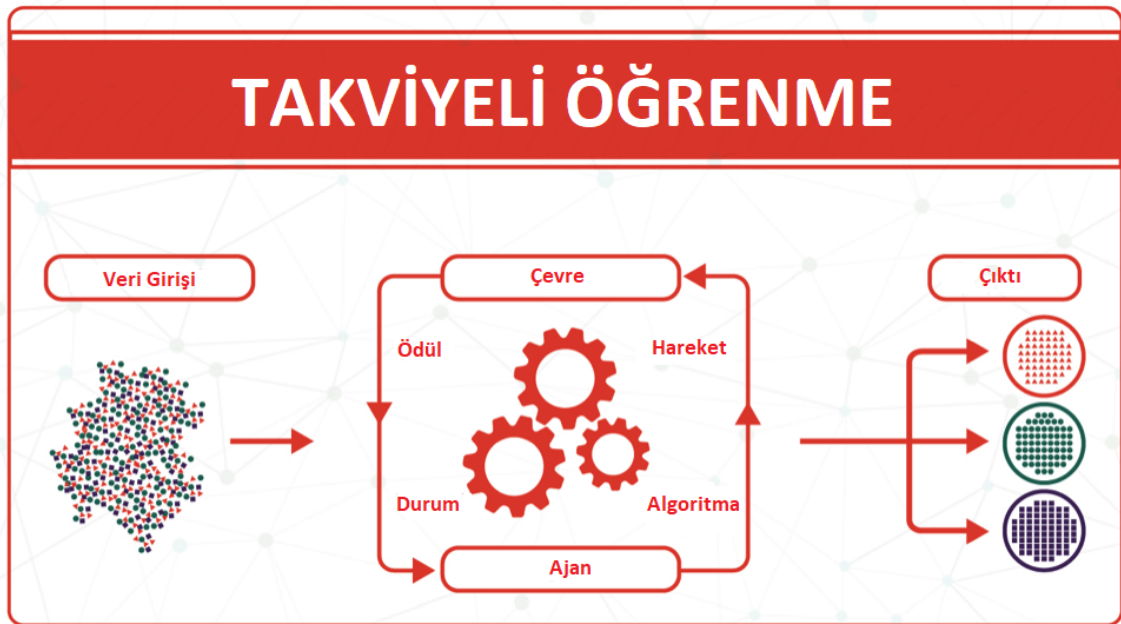
Yarı danışmanlı öğrenmede sistemin girişinde etiketlenmemiş veriler ile beraber az miktarda etiketlenmiş veri girilmektedir (Witten ve ark., 2011). Bu yöntem, genellikle etiketlenmiş verinin az olduğu durumlar ile etiketlenmemiş veriyi elde etmenin kolay olduğu durumlarda tercih edilmektedir. Yarı danışmanlı öğrenme Şekil 2.7’ de görülmektedir.



Şekil 2-7 Yarı danışmanlı öğrenme

2.3.1.4. Takviyeli Öğrenme

Takviyeli öğrenmenin çalışma sisteminde, danışmanlı öğrenmeninki gibi sisteme bir danışman yardımcı olur. Danışmanın görevi, sisteme verilen girdinin beklenen çıktısı ile hesaplanan çıktısını karşılaştırarak sonucun doğru ya da yanlış olduğunu sisteme dönüt olarak vermektir. Genellikle sistemin hesapladığı ya da oluşturduğu sonucun doğru olması durumunda sisteme ödül, yanlış olması durumunda da sisteme ceza verilerek sistemin öğrenmesini sağlayan yönteme takviyeli öğrenme denmektedir (Alpaydın, 2004). En yaygın bilinen takviyeli öğrenme algoritması olarak Q-learning algoritmasıdır. Q-learning algoritması bir çeşit pekiştirmeli algoritmadır. Watkins tarafından 1989 yılında doktora tezinde yayınlanmıştır. Q-öğrenmenin amacı, rasgele bir ortam içinde etmene en uygun yolu nasıl öğrenebileceğini göstermek üzere tasarlanmıştır. Takviyeli öğrenme algoritması Şekil 2.8’de gösterilmiştir.



Şekil 2-8 Takviyeli öğrenme

2.4. Sınıflandırma

Sınıflandırma elimizdeki verilerin etiketli yani sınıflarının belli olduğu durumlarda etiketsiz yani sınıfları belli olmayan yeni verilerin sınıflarının bulunması işlemine denilmektedir (Amasyalı, 2008). Sınıflandırma yöntemlerinde çıktı olarak üretilen her bir eleman birer sınıf olarak isimlendirilirken, sınıflandırma probleminin çözümü için kullanılan algoritmaya da sınıflayıcı denilmektedir. Makine öğrenmesi algoritmaları

içerisinde sınıflandırma için kullanılan bir çok algoritma mevcuttur. Bunlardan en iyi bilinenleri lojistik regresyon, karar ağaçları, Naif Bayes sınıflandırma, yapay sinir ağları, destek vektör makineleri, k-en yakın komşu, rastgele orman algoritması gibi algoritmalarıdır. Bir sınıflandırıcı bir başka ifade ile bir sınıflandırma modeli örneklerden tahmin edilen sınıflara doğru bir haritalamadır (Fawcett, 2006). Sınıflandırma makine öğrenmesinin en bilinen problem türlerinden birisidir. Sınıfları belli olan bir veri grubuna sınıfları belli olmayan bir veri grubunun eklenmesi için kullanılan bir yöntemdir (Harrington, 2012). Sınıflandırma yapılmasındaki amaç, elimizdeki sınıfları belli olan verilerin kullanılmasıyla oluşturulan modellerin, sınıfları belli olmayan ve sistem tarafından daha önceden karşılaştırılmamış örnekler üzerinde yüksek doğrulukla yapmasını sağlamaktır (Çakır, 2005).

2.5. Kullanılan Makine Öğrenmesi Yöntemleri

2.5.1. Naif Bayes

Naif Bayes (Naive Bayes, Basit Bayes) danışmanlı öğrenme ve yapay zeka ile veri madenciliği içinde yer alan en basit metottur (Medhekar ve ark., 2013). KNN, Karar ağaçları, destek vektör makineleri ve yapay sinir ağları gibi yöntemler bir sınıflandırma kuralını doğrudan modellerler. Buna karşılık, Bayes kuralı veya Naif Bayes yönteminde ise sınıflar için olasılığa dayanan model oluştururlar. 18. Yüzyılda bir matematikçi olan Thomas Bayes, Bayesyen yöntemlerinin temelini kuran Bayes teoremini ortaya atan kişidir. Makine öğrenmesinde Bayesyen yöntemlere dayanan sınıflandırıcılar, eğitim verilerinden yararlanarak, değişken değerlerine yani özelliklere göre her bir sınıf için gözlenen bir olasılık hesaplar ve daha sonra bu gözlenen olasılıkları etiketlenmiş test verilerinde sınıfların tahmin edilmesinde (kestirilmesinde) kullanırlar. Naif Bayes, karmaşık sorunların olduğu durumlarda problemleri çözmek için doğal araçlara dayanan, açık, anlaşılır ve özel bir yöntemdir. Özellikle birbirinden bağımsız özelliklerin var olduğu veritabanının önemli olduğu durumlarda kullanılır. Ayrıca, metin sınıflaması yaparak, çöp e-postalar filtrelenebilir, bir metnin yazarı belirlenebilir. Metinlerin siyaset, spor, magazin gibi konularda sınıflandırılmasında kullanılabilir. Öneri sistemlerinde ve tıbbi teşhis koymada da kullanılabilir.

Bayes teoremi aşağıdaki eşitlikle ifade edilebilir:

X ve Y rastgele iki olay olsun. Bu durumda;

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

$P(X) \Rightarrow X$ olayının olasılığı (bağımsız)

$P(Y) \Rightarrow Y$ olayının olasılığı (bağımsız)

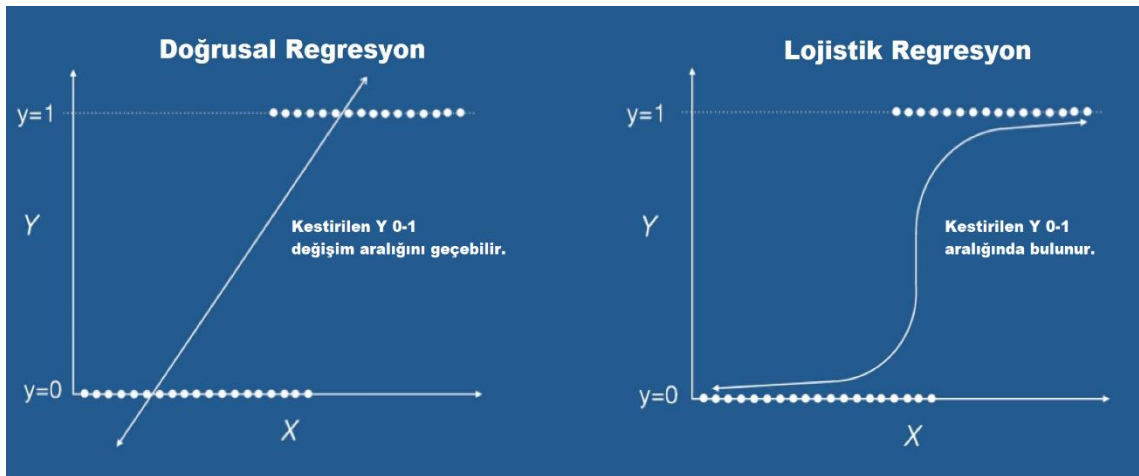
$P(X|Y) \Rightarrow Y$ 'nin olduğu bilindiğinde X 'in olma olasılığı

$P(Y|X) \Rightarrow X$ 'in olduğu bilindiğinde Y 'nin olma olasılığı

Eşitlikten yararlanılarak $P(X|Y)$ 'yi maksimum yapan durumlar hesaplanarak Naif Bayes algoritması çalıştırılmış olur. $P(Y)$ bütün sınıflar için sabit olduğundan $P(X|Y) = P(Y|X)P(X)$ olasılığı için en büyük değer aranır.

2.5.2. Lojistik Regresyon

Lojistik regresyon bir sınıflandırma algoritmasıdır. Lojistik regresyon tıp ve sosyal bilimler araştırmalarında hem olanakları hem de kullanım kolaylıkları açısından çok sık kullanılan bir değerlendirme tekniğidir. Lojistik regresyon, doğrusal regresyonun bağımlı değişkeninin iki düzeyli (1/0, Evet/Hayır, Doğru/Yanlış) veya daha fazla (0/1/2, az/orta/çok) olduğu özel bir durum olarak düşünülebilir (Gürsakaç, 2019). Yani lojistik regresyonda hedef değişkeni ikili kategorik, çoklu kategorik veya sıralı olmak zorundadır. Lojistik regresyonda bir olayın gerçekleşme olasılığı, veriye lojit fonksiyonu uydurularak tahmin edilebilir.



Şekil 2-9 Doğrusal regresyon ve lojistik regresyon

Şekil 2-9’da da görüldüğü gibi doğrusal regresyon ile lojistik regresyon arasındaki en temel fark doğrusal regresyonda kestirilen Y’nin her hangi bir değer alması mümkün iken lojistik regresyonda Y’nin sadece 0-1 değerini alabilmesidir (Gürsakal, 2019).

Lojistik regresyon modelinde bağımlı değişken için alınan nitelik yanıt aynen kullanılmaz, dönüşüme tabi tutulur. Bu dönüşüm, bağımlı odds’unun (üstünlük oranı) doğal logitması alınarak yapılır. Dönüşümden sonra en çok olabilirlik (maximum likelihood) yöntemi ile bağımlı değişken tahmin edilir. Burada dikkat edilmesi gereken husus şudur: lojistik regresyonda bağımlı değişkenin log odds’unun değişimi hesaplanır, bağımlı değişkenin kendi değişimi hesaplanmaz, bu değişim doğrusal regresyon analizinde hesaplanır.

Lojistik regresyon modelinin temeli, üstünlük oranı dediğimiz Odds Ratio’ya dayanmaktadır. Bir olayın meydana gelmesi olasılığının, meydana gelmeme olasılığına oranlanmasıyla üstünlük oranı elde edilir.

Lojistik regresyon modeli, başarı olasılığının “p” lojit dönüşümüdür.

$$p' = \ln \left[\frac{p}{1-p} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Eşitlikte:

p : olayın görülme olasılığı (başarı olasılığı)

p/(1-p) : odds ratio

$\ln(p/(1 - p))$: odds ratio’nun doğal logaritması ya da lojit

2.5.3. Yapay Sinir Ağları

Yapay sinir ağları (YSA) insan beyninin öğrenme özelliği yolu ile yeni bilgiler üretme ve keşfetme becerilerini, otonom ve kendi başına yerine getirebilecek bir bilgisayar sistemi olarak tanımlanabilir (Bilgin,2018). Bilgisayarların öğrenebilen ve herhangi bir bilgisi olmadığı veya eksik bir bilgiye sahip olduğu bir konuda yorum üretebilecek sistemlerdir (Öztemel, 2012). Diğer bir ifade ile eğitilebilir, uyarlanabilir ve kendi kendine organize olup öğrenebilen ve değerlendirme yapabilen yapay sinir ağları ile insan beyninin öğrenme yapısı modellenmeye çalışılmaktadır.

YSA, insan beynindeki nöronların işlevini yapay olarak gerçekleştirilen basit hesap hücrelerinin birbirleriyle olan bağlantılarını kullanarak en iyi performansa ulaşmaya çalışır.

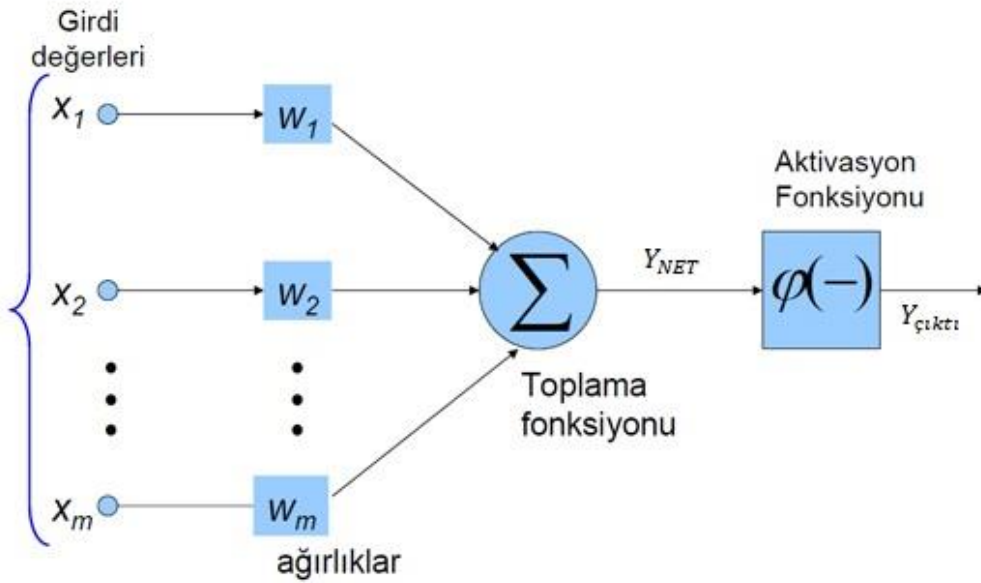
Genel olarak yapay sinir hücresi kendisinden önceki sinir hücrelerinden aldığı sinyalleri kendisinde topladığı ve toplam sinyal birikiminin belli bir eşik değerini aştığı anda kendi sinyalini bir başka sinir hücresine iletilmesi olarak söylenebilir.

Yapay sinir ağları üç unsurdan oluşur:

Girdi Katmanı: Bir ağa girdi olarak verilen öğrenilmesi istenen örneğin özniteliklerinin veya değişkenlerin giriş olarak verildiği katmandır. Girdi katmanında öğretilecek örneklerin öznitelik sayısı kadar giriş nöronu, diğer bir ifade ile giriş düğümü bulunmalıdır.

Gizli Katman: Giriş katmanı ile çıkış katmanları arasındaki katmanlardır. Katman sayısı ve katman üzerindeki nöron (düğüm) sayısı problemten probleme değişebilmektedir. Bu katmanlarda ileri yönlü hesaplamalar ve geri yönlü hata yayılımı yapılır. Katman sayısı fazla olduğunda hesaplama karmaşıklığına ve hesaplama süresinin artmasına sebep olur. Karmaşık problemlerin çözümü için genelde katman sayıları ve katmanlardaki nöron (düğüm) sayıları fazladır.

Çıktı Katmanı: Yapay ağıdaki öğrenilmesi istenen örneklerin sınıf bilgisinin veya etiket değerinin çıkış olarak hesaplandığı katmandır.



Şekil 2-10 Bir yapay sinir hücresi yapısı

Şekil 2-10'daki yapay sinir hücresinin yapısını aşağıdaki gibi ifade edebiliriz.

Ağırlıklar (Weights): Girdi değerlerin çıktılar üzerindeki etkisini ayarlayabilmek için kullanılan parametrelerdir. Ağırlıklar girdi değerleri ile çarpılarak ileri doğru iletilir. Ağıdaki en önemli nokta, verilen ağırlıkların eğitim kümesine göre hatayı yayararak olması

gereken uygun ağırlık değerlerini hesaplamaktır. Bu ağırlıklar pozitif, negatif veya sıfır olabilir. Ağırlığı sıfır olan nöronların (düğümlerin) çıktı üzerinde etkisi olmamaktadır.

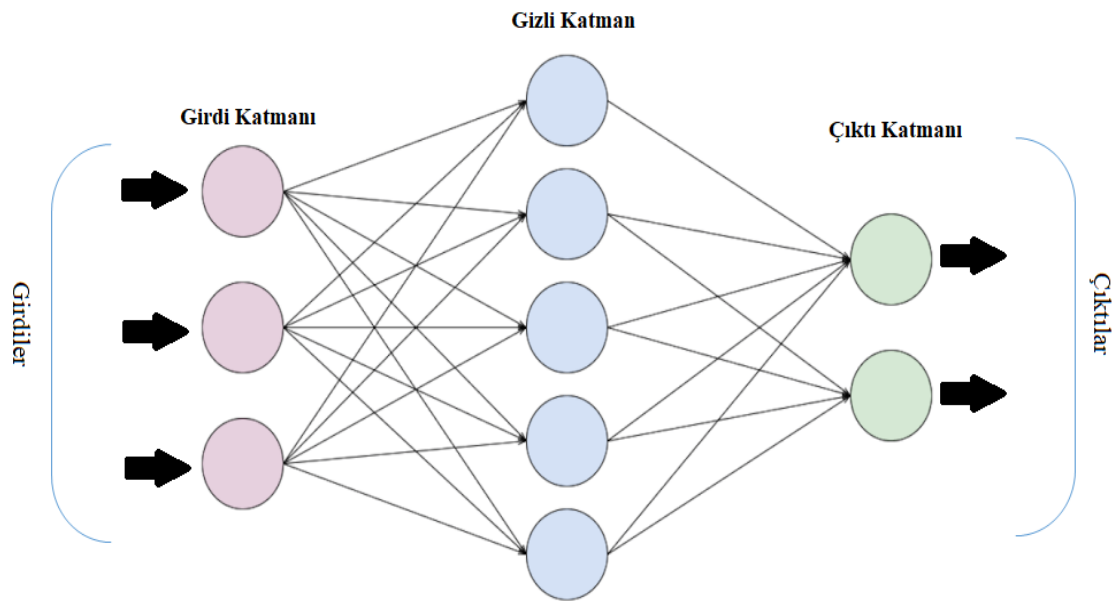
Transfer Fonksiyonu: Bir düğüme gelen net girdiyi hesaplar. Bunu hesaplamak için değişik fonksiyonlar kullanılmaktadır. En yaygın kullanılan ise ağırlıkların girdiler ile çarpımının toplamıdır.

Aktivasyon Fonksiyonu: Hücreye gelen net değeri işleyerek nöronun (düğümün) bu girdiye karşılık aktivasyon çıktısını oluşturur. Aktivasyon fonksiyonunun probleme göre uygun seçilmesi ağırlık performansını ve başarı oranını önemli derecede etkiler. Hatayı geri yayarken her katmanda türev alma işlemi yapılır. Dolayısıyla seçilecek aktivasyon fonksiyonlarının türevlenebilir ve türevlerinin kolay alınabilen fonksiyonlardan olması gerekmektedir. Aksi takdirde eğitim süresi uzayacaktır. Ayrıca, aktivasyon fonksiyonun seçimi, ağırlık verilerine ve neyin öğrenmesi istendiğine de bağlı olarak değişebilir.

En sık kullanılan aktivasyon fonksiyonları sigmoid fonksiyon ve hiperbolik tanjant fonksiyonlarıdır.

Şekil 2-10 incelendiğinde toplam ağırlıklı Y fonksiyonu ele alınırsa w ifadeleri x girişlerindeki ağırlıkları belirlemektedir. Ağırlıklı toplam matematiksel olarak aşağıdaki formülde ifade edilmiştir.

$$Y_{NET} = w_1x_1 + w_2x_2 + \dots + w_nx_n$$



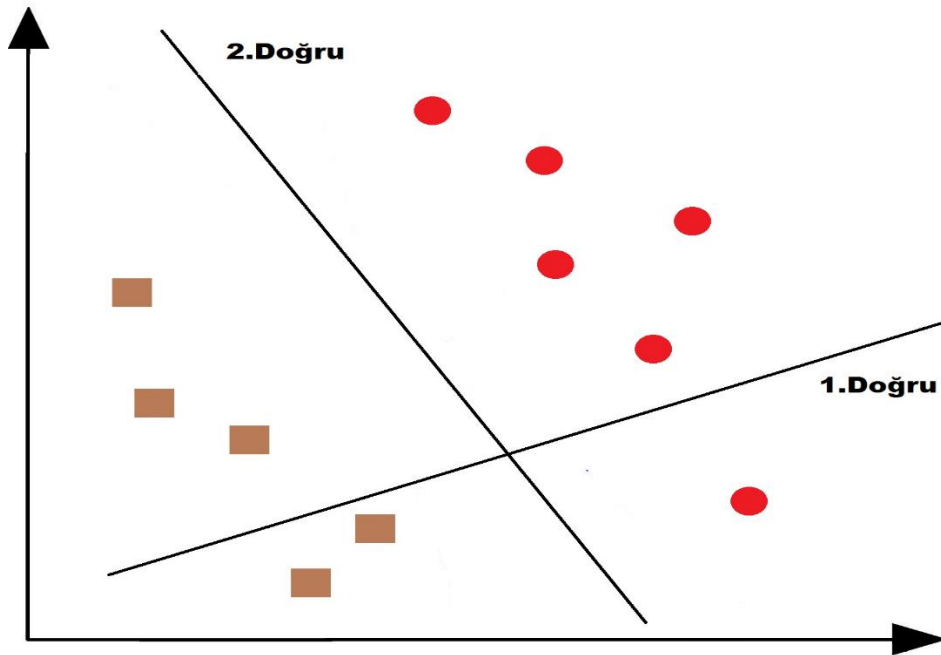
Şekil 2-11 Bir yapay sinir ağı

Şekil 2-11’de girdi, gizli ve çıktı katmanı ve her bir katmandaki düğüm sayıları 3-4-2 olan bir yapay sinir ağı görülmektedir. Bu ağda girdi katmanının düğümleri I1, I2 ve I3; gizli katmanının düğümleri H1, H2, H3 ve H4; çıktı katmanının düğümleri ise O1 ve O2 olarak ifade edilebilir. Düğümler arasındaki bağlantıların ağırlıkları ise $w(I1H1)$ şeklinde ifade edilebilir.

2.5.4. Destek Vektör Makineleri

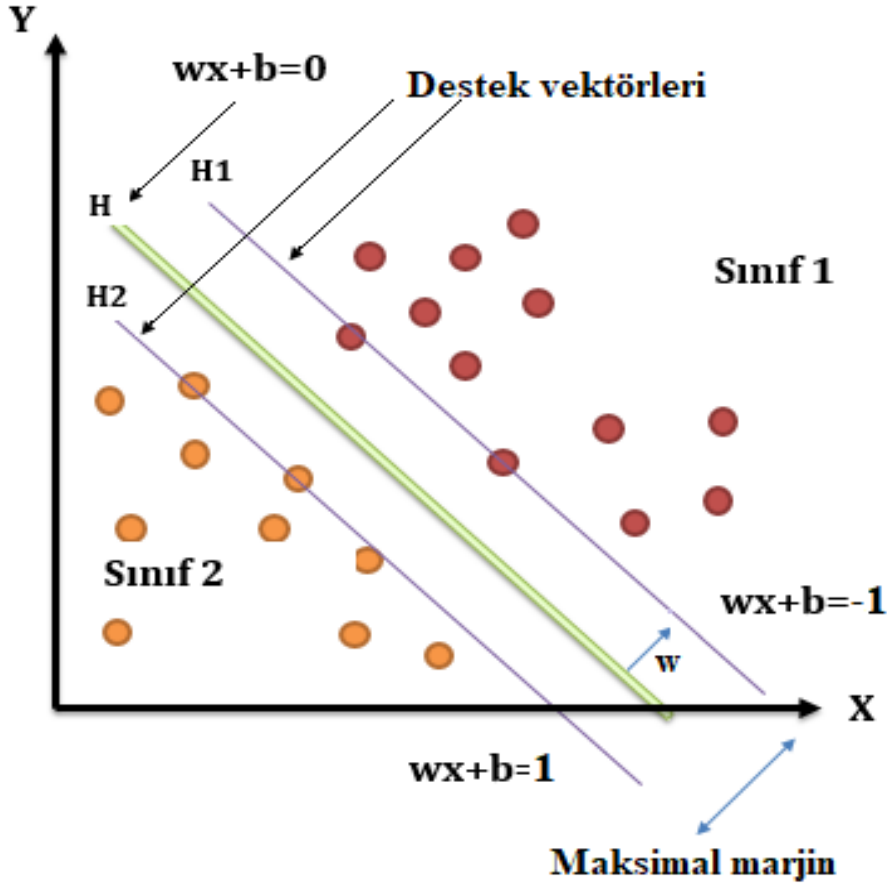
Destek vektör makineleri (Support Vector Machine(SVM)), Vapnik tarafından geliştirilen temeli istatistiksel öğrenme kuramına dayanan sınıflandırma algoritmalarından biridir. İlk olarak ayrıklık ve regresyon analizlerini çözümlemeye kullanılmıştır. Günümüzde tıbbi tanı, ses tanıma, resim ve nesne vb. çalışma alanlarında kullanılmaktadır. Destek vektör makinelerinin temel çıkış noktası sonsuz sayıdaki doğru içinden marjini en yüksek olan doğruyu seçip sınıflandırmayı gerçekleştirmeye çalışmaktır. Oluşan marjin doğrusu sınıf üyelerinin seçilen doğruya en yakın olan üyelerine paralel olmalıdır. Burda çizilen doğrulara hiperdüzlem adı verilmektedir.

Destek vektör makineleri verilerin doğrusal olarak ayrılabilmesine, doğrusal olarak ayrılmamasına göre ve çoklu sınıf problemleri gibi başlıklarda ele alınmaktadır. Doğrusal olarak ayrılabilen iki sınıflı bir veri kümesini ayırmak için çizilen iki hiperdüzlem şeklinde görülmektedir. Şekil 2-12’den de anlaşılacağı üzere 1.doğru verileri düzgün ayıramaz iken 2. doğru verileri düzgün bir şekilde ayırmayı başarmıştır.



Şekil 2-12 Sınıfları ayıran doğru

Destek vektör makinelerinde sınıfları birbirinden ayırmak için hiperdüzlemler kullanılır. Oluşturulan hiperdüzlemler marjini en büyük olan seçilerek işleme devam edilir. En büyük marjine sahip hiperdüzlemi seçebilmek için her iki sınıfa yakın noktalardan geçen paralel doğrular çizilebilir. Paralel olarak çizilen ve destek vektör adı verilen bu doğrular arasındaki uzaklığa marjin denilmektedir (Akarsu, 2016).



Şekil 2-13 İki sınıflı bir veri kümesinde, hiperdüzlem, marjin ve destek vektörleri

Şekilde 2-13'te görüldüğü üzere iki sınıfı birbirinden ayıran H , hiperdüzlemi için $w \cdot x + b = 0$ olurken, destek vektörleri olan $H1$ ve $H2$ ise $w \cdot x + b = -1$ ve $w \cdot x + b = 1$ olmaktadır.

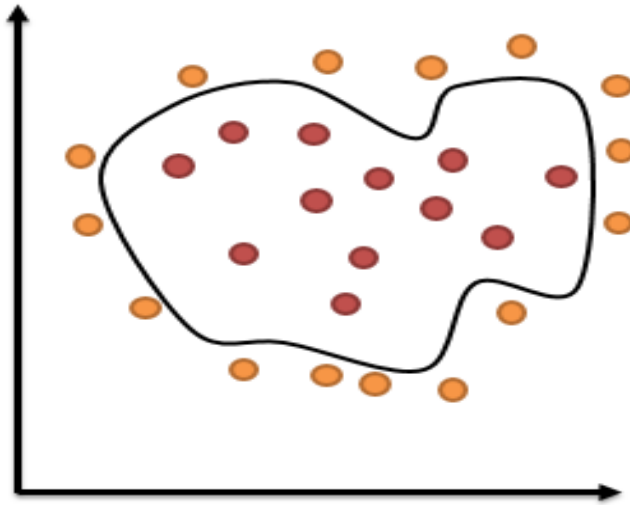
Destek vektörleri arasındaki uzaklığı ifade eden marjin genişliği ise $2/w$ şeklinde ifade edilmektedir. w değeri küçüldüğünde marjin genişliği o kadar artacak, aksi durumda ise azalacaktır. Destek vektör makinesi algoritmasındaki amaç en büyük marjin değerini elde etmek olduğundan w değeri minimize etmeye çalışmaktır (Bilgin, 2018).

Destek vektör makinelerinin ilk çıkış noktası iki sınıflı problemlerin çözümüne yöneliktir. Uygulamalarda karşılaşılan problemler genellikle ikiden fazla sınıfa sahip problemlerdir. İkiden fazla sınıfa sahip problemlerin çözümünde ikili sınıflandırıcıların birleştirilmesi

ile sonuca gidilmektedir. Bunun için önerilen pek çok başarılı yöntemden bahsedilebilir. En başarılı olan yöntemler bire-karşı-diğer (BKD) ya da bire-karşı-bir (BKB) yöntemleridir (Pöyhönen, 2004). BKD yönteminde, elimizdeki eğitim verisi ikili destek vektör makinesi ile ayrılabilir şekilde m adet sınıflandırıcıya bölünür. Oluşturulan k adet sınıflandırıcıların her biri, bir sınıf kendisi dışındaki diğer sınıflardan ayırmak için kullanılır. Gerçekleştirilen bu işlem elimizdeki tüm sınıflandırıcılar için tekrar edilir. En son aşamada elimizdeki sınıflandırıcılardan elde ettiğimiz sonuçlar birleştirilerek en son karar oluşturulmuş olur (Pöyhönen, 2004). BKB yönteminde, m adet sınıflı bir sınıflandırma problemi çözebilmek için $M=m(m-1)/2$ adet farklı sınıflandırıcı oluşturulur. Bu sınıflar x . ve y . sınıfa ait veriler ile eğitilerek başlanır. Eğitim örneklerini sınıflandırabilmek için yarışma yöntemi kullanılmaktadır. Eğitim bittiğinde n eğitim örneği x . sınıfa ait ise x . sınıfın puanı bir artırılır, değil ise y . sınıfın puanı bir artırılır. İşlem bittiğinde n örneği en yüksek puanı elde eden sınıfa eklenir (Bilgin, 2018). Doğrusal olarak ayrılamayan veri kümelerinde sınıflandırma yapılabilmesi için ilgili veriler daha yüksek boyutta bir uzaya taşınarak, sınıflandırma işlemi oluşturulan bu yeni uzayda gerçekleştirilebilir.

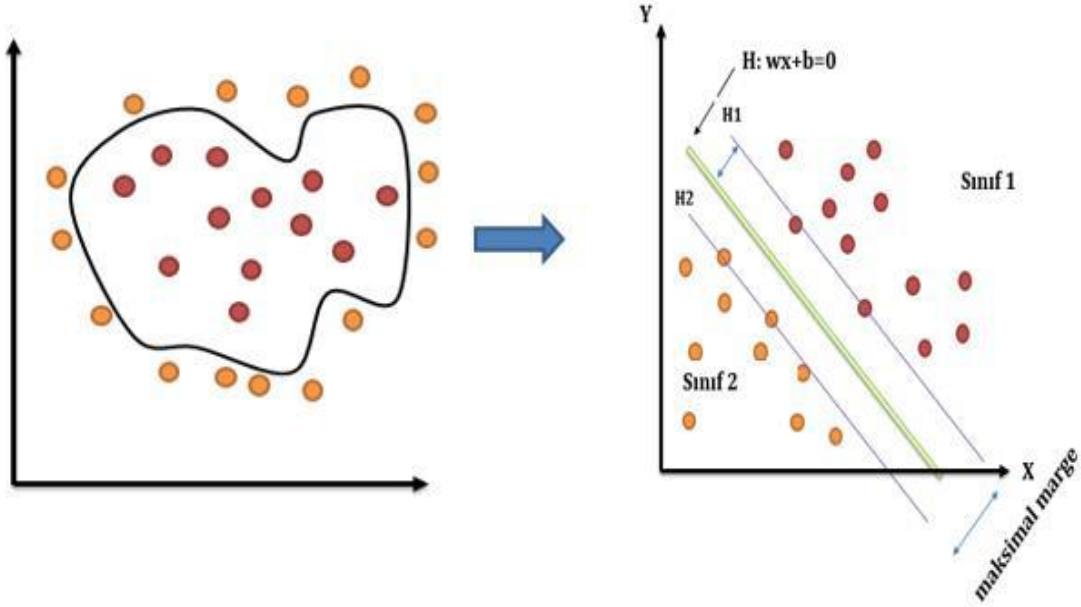
Şekil 2-14'te görüldüğü gibi veriler gibi doğrusal olarak ayrılabilir değildir. Bu nedenle, doğrusal olmayan uzayı yeni doğrusal yüksek boyutlu uzayla eşleştirerek eğitim seti doğrusal olarak ayrıştırılabilir. Denklemde görüldüğü gibi çekirdek fonksiyonuyla ifade ederek eşleştirilmektedir (Amani, 2017).

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$



Şekil 2-14 Doğrusal olmayan durumda DVM

Dolayısıyla, bu dönüşüm uzayı, SVM çekirdeği seçilirken dolaylı olarak tanımlanan bir $F = \{\Phi\{x\} | x \in X\}$ “eşleştirme fonksiyonu” kullanılarak gerçekleştirilir. Yeni uzaya “özellikler uzayı” denilmektedir (Amani, 2017).



Şekil 2-15 Eşleşme uzayı

SVM uygulanırken birden çok çekirdek fonksiyonu kullanılmaktadır. Bunlardan en sık kullanılan fonksiyonlar aşağıdaki gibidir (Djeffal,1999).

Doğrusal çekirdek:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

Polinom çekirdeği:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$$

RFB çekirdeği:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2, \gamma > 0)$$

Sigmoid çekirdeği:

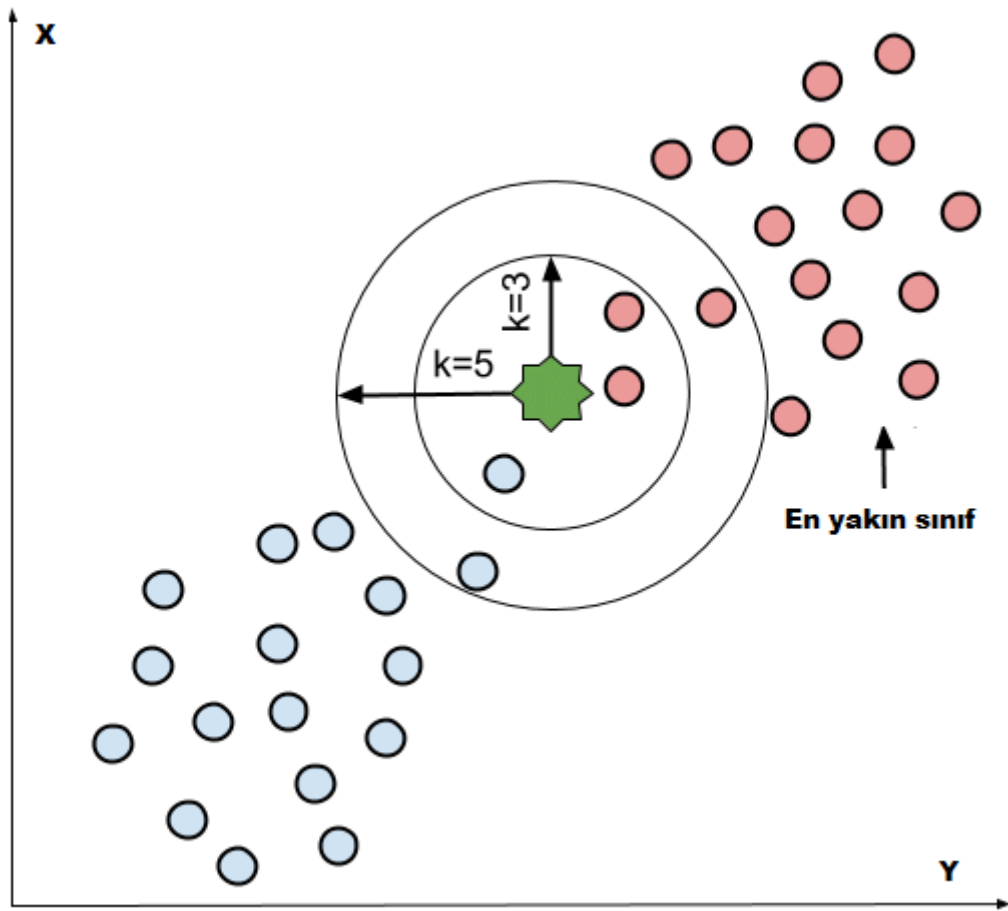
$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$$

SVM çekirdeğinin seçimi henüz çözüme kavuşmamış bir problemdir. SVM seçimi, performansı sınırlayan bir faktördür. Doğru çekirdeği seçmek için araştırmacılar

genellikle testler uygulanır. Buradaki tez çalışmasında öncelikle standart doğrusal çekirdek kullanılmış, diğer SVM çekirdekleri üzerinde de bazı testler yapılmıştır.

2.5.5. K –En Yakın Komşu

K- en yakın komşu (K-nearest neighbor-KNN) algoritması en çok bilinen sınıflandırma algoritmalarından birisidir. Seçilen bir özelliğin kendisine en yakın olan özellikler arasındaki yakınlığı kullanarak sınıflandırma işlemi gerçekleştirilir (Bilgin, 2018). K parametresi komşu sayısını temsil etmektedir ve genellikle 1'den büyük ve tek sayılar tercih edilmektedir.



Şekil 2-16 Örnek KNN algoritması

Bu algoritma, 1950 yıllarda geliştirilmeye başlanmış olsa da 1960 yıllarda daha çok tanınır hale gelmiştir (Bilgin, 2018). Gelişiminin geç olmasının sebebi büyük hesaplama gücünün sağlanmasındaki yetersizliklerin olmasından kaynaklanmaktadır (Han ve Kamber, 2006). KNN algoritması sınıflandırma için geliştirilmiş olsa da kestirim ve tahminleme için de kullanılan bir yöntemdir (Larose, 2005). KNN algoritmasının

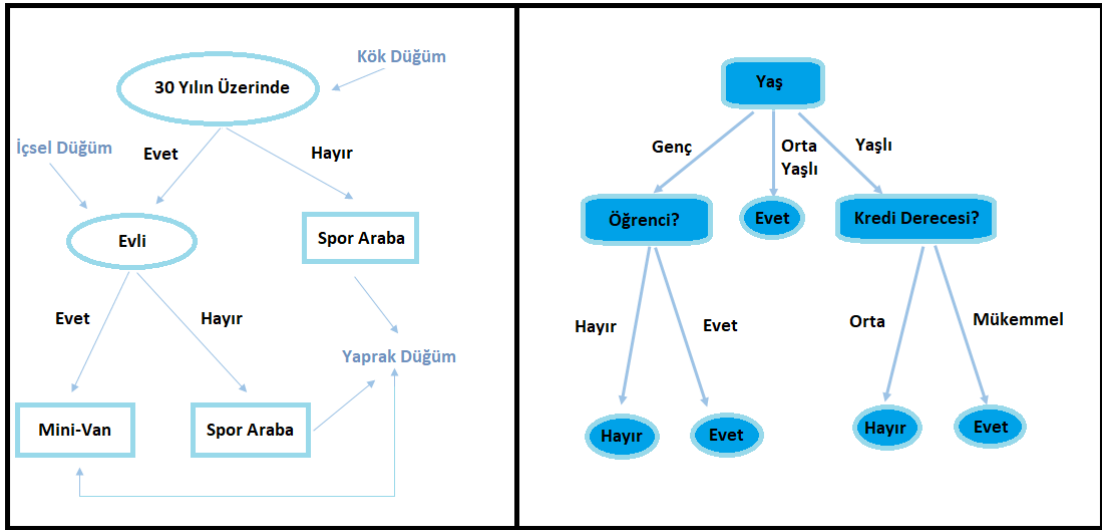
kullanım alanları olarak; örüntü tanıma, veri madenciliği, metin sınıflandırma ve görüntü işleme gibi alanlar örnek verilebilir. KNN algoritmasında sınıfı belli olmayan bir örnek geldiğinde en yakın verilerin bulunması için çeşitli uzaklık hesaplama ölçütleri kullanılmaktadır. Bu ölçütlerden en fazla kullanılanlar Öklid, Manhattan ve Minkowskidir.

2.5.6. Karar Ağaçları

Karar ağaçları (decision tree) sınıflandırma problemlerinde en çok kullanılan algoritmalarından birisidir; çok sayıda veri içeren bir veritabanını, bazı teknikler kullanarak alt bölümlere ayırmak için geliştirilmiş bir algoritmadır. Diğer sınıflandırma algoritmalarına kıyasla yapılandırılması ve anlaşılması oldukça kolaydır.

Karar ağaçları sınıflama ve regresyon amacı ile kullanılır. “Eğer öyle ise sonra ne ?” mantığı içinde bu ağaçlar, akış diyagramı mantığı içinde ardışık soruları cevaplarlar. Karar ağaçları görsellik sağladıkları için kolay yorumlanabilirler, sayısal ve kategorik veriler üzerinde uygulanabilirler. Regresyon ağaçları uygulamalarında “ardışık ikili ayırma” (recursive binary splitting) yöntemi ile eğitim verileri üzerinde büyük bir ağaç oluşturulur. Daha fazla ayrılamayan terminal düğümlerde minimum sayıda gözlemde daha az gözlem kaldığında işlem tamamlanır. Bir sınıflama ağacı da aynı regresyon ağacı gibidir. Ancak burada nicel bir tepki yerine, nitel bir tepki kestirilir (Gürsakar, 2018) .

Karar ağaçları çeşitli hareket biçimleri arasında tercih yapmamızı sağlayan araçlardır. Anlaşılması kolay olan karar ağaçları akış diyagramına benzeyen bir şekilde sahiptirler ve bu diyagram çeşitli kararların sonuçlarını bize gösterirler. Üç ana parçaya sahip olan karar ağaçlarında kök düğüm ağacın başlangıç noktasıdır. Hem kök düğüm hem de yaprak düğümler cevaplanacak sorular veya kriterler içerirler. Dallar ise düğümler arasındaki bağlantıları gösterirler. Örneğin, eğer ilk düğümdeki soru evet ve hayır cevaplarını içeriyorsa bir yaprak düğüm evet ve bir yaprak düğüm ise hayır cevabı ile ilgili olacaktır. Karar ağaçları danışmanlı öğrenmenin bir türü olduğundan iki aşamalı şekilde çalışır. Karar ağaçlarında makine öğrenmesinin uygulanması için genelde CART (Classification and Regression Trees) modeli uygulanmaktadır (Gürsakar, 2018). İlk kısmı öğrenme dediğimiz elimizdeki verilerden bir model oluşturma aşamasıdır. İkinci kısmı ise test verisi model üzerinde denenerek test verisinin sınıfı belirlenmeye çalışılır (Metin, 2018).



Şekil 2-177 İki dallı ve çok dallı karar ağaçları

2.5.7. Bagging (Torbalama)

Bagging (Torbalama, Bootstrap Aggregating), sınıflandırma ve regresyon görevleri için kullanılan ve Breiman tarafından önerilen bir grup makine öğrenme algoritmasıdır (Breiman, 1996). Bazı makine öğrenme algoritmaları, yüksek varyans problemine çare gibi görünen istatistiksel bootstrap (önyükleme) yaklaşımını kullanır. Karar ağacı gibi klasik algoritmalar oldukça değişken tahmin sonuçları üretebilir. Bu, iki ayrı ama benzer veri seti arasındaki sonuçların oldukça farklı olduğu anlamına gelir. Önyükleme (bootstrap) yaklaşımı art arda değiştirilen eğitim verilerinden rastgele örnekler oluşturmaya dayanır. Bagging'de, her model yeni oluşturulan bootstrap (önyükleme) örnekleri kullanılarak bağımsız olarak oluşturulur, her bir modelin tahmini hesaplanır ve tahminlerin ortalaması nihai tahmin olarak belirlenir. Toplama fikri, Bagging'i daha düşük varyansa sahip modeller üretme perspektifinden daha üstün kılar. Sınıflandırma ortamlarında, her bir ağaç sonucu dikkate alınarak nihai tahminler, sınıf tahminlerinin çoğunluğu ile elde edilir. Ağaç sayısı, performans üzerinde çok etkili bir parametre değildir (James ve Witten, 2013). Diğer topluluk yaklaşımlarında olduğu gibi, Bagging (Torbalama)'de genellikle kararsız bir algoritma olan bir temel öğrenici kullanır. Sınıflandırma ve regresyon ağacı (CART) gibi karar ağaçları yüksek varyansa maruz kalır ve bu nedenle CART, Bagging algoritmasında yaygın olarak kullanılmaktadır. Ancak, CART dışındaki herhangi bir algoritma da toplama için kullanılabilir. Bagging (Torbalama)'de temel öğrenenler olarak çok değişkenli adaptif spline (MARS) veya klasik lineer regresyon gibi daha düşük varyanslı bazı algoritmalar kullanıldığında

performans iyileşmesinin anlamlı olamayacağı unutulmamalıdır (Kuhn ve Johnson, 2013).

2.5.8. Rastgele Orman

Topluluk sınıflandırma yöntemlerinin çalışma ilkesi, bir sınıflandırıcının yerine birden fazla sınıflandırıcının ürettiği sonuçlar kullanarak yeni bir verinin sınıflandırılmaya çalışmasıdır (Akar ve Güngör, 2012). Topluluk sınıflandırma algoritmaları içerisinde en yaygın kullanılanlardan birisi olan Rastgele Orman (Random Forest), Breiman ve Cutler tarafından önerilen yöntem, mevcut durumlardan rastgele oluşturulan örnekleri kullanarak çoklu ağaçlardan oluşan bir sınıflayıcı türüdür. Bir örneği sınıflandırmak için, ormandaki her ağaca bir giriş vektörü verilmekte ve her ağaç için bir sonuç elde edilmektedir (Bilgin, 2018). Rastgele orman algoritması verilen sonuçlardan en çok oyu alan sınıfı seçer (Doğan, 2006).

Breiman, Rastgele Orman algoritmasını, tahmin edilen ağaçların bir birleşimi olarak ifade etmiştir. Tahmin edilen her ağaç (kestirim yapılan), ağaç kümesi içerisinde birbirinden bağımsız olarak gösterilen raslantısal bir vektörün değerine karşılık gelmektedir (Bilgin,2018). Breiman'a göre topluluk sınıflandırıcılarında oluşturulan ağaç kümesinin genelleme hatası, ağaçların her birinin küme içerisindeki bireysel güçlerine ve ağaçlar arasındaki korelasyona bağlıdır (Breiman, 2001).

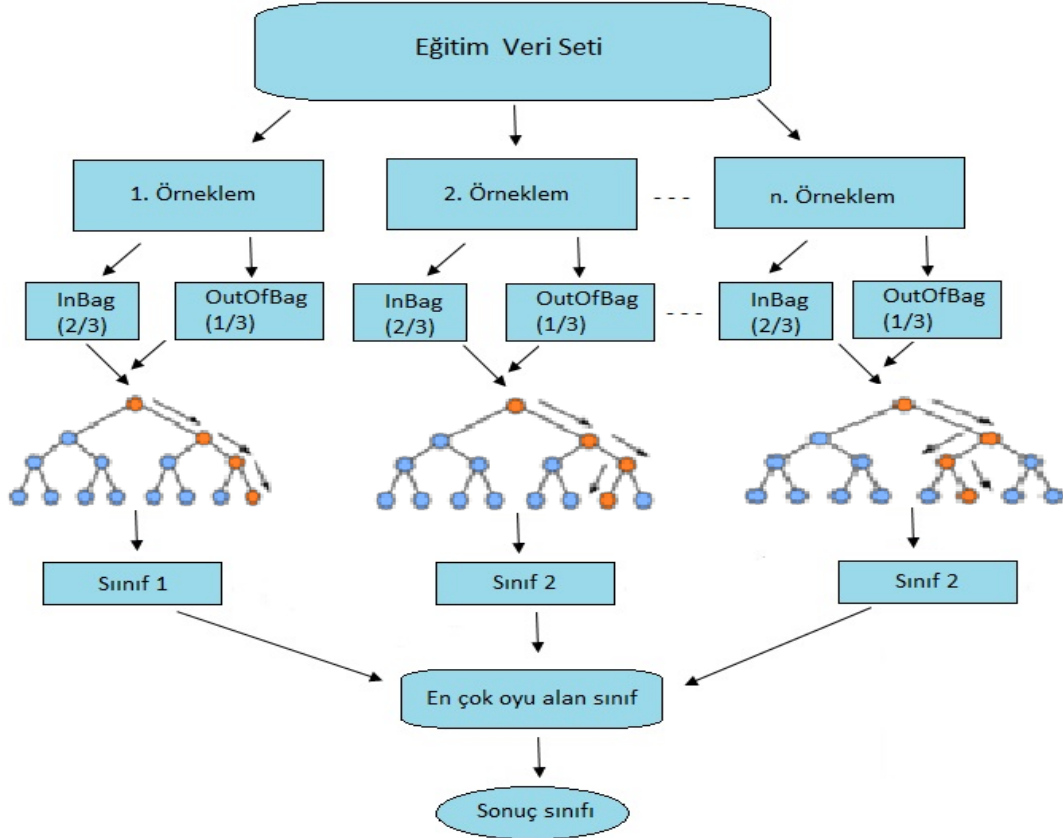
Rastgele Orman tüm değişkenler arasından en iyi dalı kullanarak her bir düğümü dallara ayırmaktansa, her bir düğümde rastgele olarak seçilen değişkenler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Her bir veri kümesi orjinal veriler içerisinde yer değiştirmeli olacak şekilde seçilerek oluşturulur. Ardından rasgele özellikler seçilerek ağaç yapıları oluşturulur. Rastgele özellikler ile oluşturulan ağaçlarda budama işlemi gerçekleştirilmez (Archer,2008; Breiman 2001). Geliştirilen bu yöntem ile Rastgele Orman'ın doğruluğu artırılmış olmaktadır. Bu algoritmanın artılarından diğerleri; çok hızlı olması, aşırı uyuma karşı dayanıklı bir yapısının olması ve istenildiği kadar ağaç ile çalışabilmesidir.

Rastgele Orman algoritmasında eğitim setindeki veriler yardımıyla ağaçları oluşturabilmek için kullanıcının belirlediği ağaç sayısına (N) ve her düğümde kullanılması düşünülen özellik sayısına (m) diyelim. İlk yapılması gereken işlem Inbag (IB) adı verilen eğitim veri kümesinin üçte ikisi bootstrap (ön yükleme) verisi olarak kullanılmak üzere ayrılır. Ardından Outofbag (OOB) adı verilen, eğitim veri kümesinden

kalan üçte birlik bölüm sistemi test etmek için ayrılır. Yapılan bu işlemlerin ardından her bir ön yükleme verisi yardımıyla bir ağaç modeli oluşturulur. Her düğüm noktasında m özellikleri tüm özellikler arasından rastgele şekilde ve seçilen özellikler arasından en iyi dal bulunmaya çalışılır. Özellikler sayısının seçimi, düşük korelasyon elde etmek için büyük öneme sahiptir. Şekil 2-19'da Rastgele Orman algoritmasının çalışma şekli verilmektedir.

Koşullu çıkarım ağacı (CIT) ve CART, Rastgele Orman algoritmasında temel öğrenen olarak yaygın kullanılmaktadır. CIT, klasik karar ağacının bir çeşididir ve saflık benzeri ölçümler yerine istatistiksel olarak anlamlılık testleri yaparak değişkenlerin ve bölünmelerin seçimine dayanır. En düşük anlamlılık değeri sağlayan değişken ve bölme, ağacın inşasının her adımında kullanılır (Kabacoff,2015).

Rastgele orman kısaca ifade edilecek olursa, CART modeli uygulanan bootstrap benzeri bir karar ağacı algoritmasıdır. Varsayalımki elimizde 1000 gözlem ve 20 değişkenden oluşan bir veri kümesi olsun. Rastgele Orman rassal olarak 100 gözlem ve 10 değişken seçerek farklı örneklem ve farklı değişkenler için çok sayıda CART modeli bulup sonunda her model için yaptığı kestirimlerin ortalamasını alarak nihai bir kestirim yapar.



Şekil 2-18 Rastgele orman algoritması (Ayas, 2014)

Bagging (Torbalama), tüm özelliklerin ağacı bölmek için kullanılmasına izin verirken, Rastgele Orman belirli ve rastgele seçilmiş bir değişken sayısını göz önünde bulundurur. Bu fikir Dietterich tarafından geliştirilmiştir (Dietterich, 2000). Ağaçların yapımında rastgele değişken alt kümelerinin kullanılmasına dayanan benzer bir yaklaşım (Amit ve Geman, 1997) ve (Ho, 1998) tarafından kullanılmaktadır. Breiman, bahsi geçen bu ağaç yapım sürecini Rastgele Orman olarak bilinen genel bir çerçeve ile sunmuştur (Breiman, 2001). Bölünmek için rastgele değişken örneği düşünmeye dayanan fikir, dekorla ilişkili ağaçları sağlar ve Bagging'den daha iyi ve daha az değişken sonuçlar sunar. Bagging algoritmasında, her ağaç, ağaçlarda bölünen aynı değişkenleri seçtiği için benzer sonuçlar üretebilir. Bu ağaçlar kullanılarak elde edilen tahminler yüksek oranda korelasyon gösterecektir. Her ne kadar yüksek korelasyonlu sonuçların ortalaması belli bir derecede etkili olsa da, Rastgele Orman genellikle daha etkili bir çözüm sunar (James ve Witten, 2013).

2.5.9. Gradient Boosting Makinesi (GBM)

GBM, sınıflandırma ve regresyon problemleri için çok uygun olan popüler ve gelişmiş bir makine öğrenme algoritmasıdır. Bagging (Torbalama) ve Rastgele Orman gibi, GBM yöntemi karar ağaçları gibi zayıf bir öğreniciyi geliştirmeye ve bir topluluk modeli oluşturmaya dayanır. GBM algoritması, ikili sınıflandırma görevleri için tanımlanan AdaBoost adlı başka bir yükseltme (artırma) algoritması ile yakından ilgilidir. AdaBoost algoritmasının arkasındaki fikir, her örneğe başlangıç ağırlıkları atamak ve bu ağırlıkları eğitim sürecinin her yinelenmesinde güncellemektir (Freund ve Schapire, 1996). Güncelleme hata oranı kullanılarak gerçekleştirilir. Doğru sınıflandırılmış örnekler daha düşük ağırlık değerlerine sahipken, yanlış sınıflandırılmış örnekler daha fazla ağırlık almaktadır. Yeni güncellenen ağırlıklardaki örnekler bir sonraki yinelemede kullanılır. İşlem birkaç yineme için tekrarlanır ve nihai model ağırlıklı modellerin kombinasyonu kullanılarak oluşturulur (Kuhn ve Johnson, 2013). Başka bir deyişle, karar ağaçları gibi temel öğrenenler önceki yinelemelerde elde edilen bilgiler dikkate alınarak sırayla inşa edilirler (James ve Witten, 2013). GBM, AdaBoost algoritmasının bir genellemesidir ve amacı, gradyan inişine dayanan yaklaşımla bir kayıp (maliyet) fonksiyonunu en aza indirmektir. Eğitim sürecinde keyfi bir farklılaştırılabilir kayıp işlevi kullanılır. Böylece GBM'yi AdaBoost'un aksine çok sınıflı sınıflandırma veya regresyon amacıyla kullanmak mümkün olacaktır. Breiman, artırmanın uygun bir maliyet işleviyle kullanımını

sunmuştur (Breiman, 1998). Daha sonra, Friedman, gradyan boosting (artırma) makinelerini ve stokastik gradyan artırma algoritmalarını birleşik yükseltme yaklaşımı çerçeveleri olarak önermiştir (Friedman, 2001; Friedman, 2002). Ağaç sayısı, büzülme, etkileşim derinliği ve minimum terminal düğüm boyutunun değeri GBM'nin ayarlama parametreleri olarak tanımlanır ve çapraz doğrulama yaklaşımları ile belirlenmelidir. Ağaç sayısı çok büyükse aşırı yüklenmeye neden olabilir. Büzülme parametresi öğrenme oranında etkilidir. Bu parametrenin değeri, daha iyi performans elde etmek için gereken ağaç sayısını değiştirebilir. Etkileşim derinliği, modelin karmaşıklığını açıklar ve her bir ağaçtaki yarık sayısını gösterir (James ve Witten, 2013). Minimum terminal düğüm boyutunun değeri, gerekenden daha küçük veya daha büyük olmaktan kaçınarak optimum boyut oluşturmayı garanti edebilir.

2.5.10. C5.0

C5.0, sınıflandırma görevleri için gelişmiş makine öğrenmesi algoritmasıdır ve Quinlan tarafından önerilen C4.5 algoritmasına dayanmaktadır (Quinlan, 1992). C4.5 eğitim süreci C5.0 ile yaklaşık olarak aynıdır. C4.5'e ek olarak, C5.0 artırma özelliğine sahiptir. Hata değerleri ile orantılı olarak farklı maliyetler (ağırlıklar) atar (Kuhn, 2013). Bu yükseltme işlemi AdaBoost algoritmasındaki ile aynıdır. AdaBoost'un aksine, her ağacın boyutu birinciyle aynıdır ve etkileşim derinliği bir ayar parametresi değildir. Entropiye dayanan bilgi kazanımı kriteri, her tekrarda en iyi bölünmeyi belirlemek için kullanılır. Daha genelleştirilebilir ağaçlar elde etmek ve fazla takmaktan kaçınmak için bir budama seçeneği vardır. Budama öncesi ve sonrası seçenek C5.0 eğitiminde dikkate alınabilir. Bu özellikler dışında C5.0, hız, bellek kullanımı ve verim nedeniyle büyük veri kümeleri için C4.5'ten daha uygun sınıflandırma algoritmasıdır. Her türlü veri yapısı C5.0 sürecinde dikkate alınabilir. Bu özelliklerin bir sonucu olarak, C5.0, SVM veya Yapay Sinir Ağları gibi rakipleriyle benzer performans sunar (Lantz, 2015). Ağaç ve C5.0 model sürümü budama seçeneği ile kullanılabilir. Bu çalışmada C5.0 ağaç versiyonu tercih edilmiştir. Takviye yineleme sayısı ayar parametresidir ve çapraz doğrulama yaklaşımı ile seçilmelidir.

2.6. Performans Ölçümü

Bu bölümde makine öğrenmesi algoritmalarının sınıflandırmada performans başarılarını ölçmek için ölçüm metrikleri ifade edilecektir.

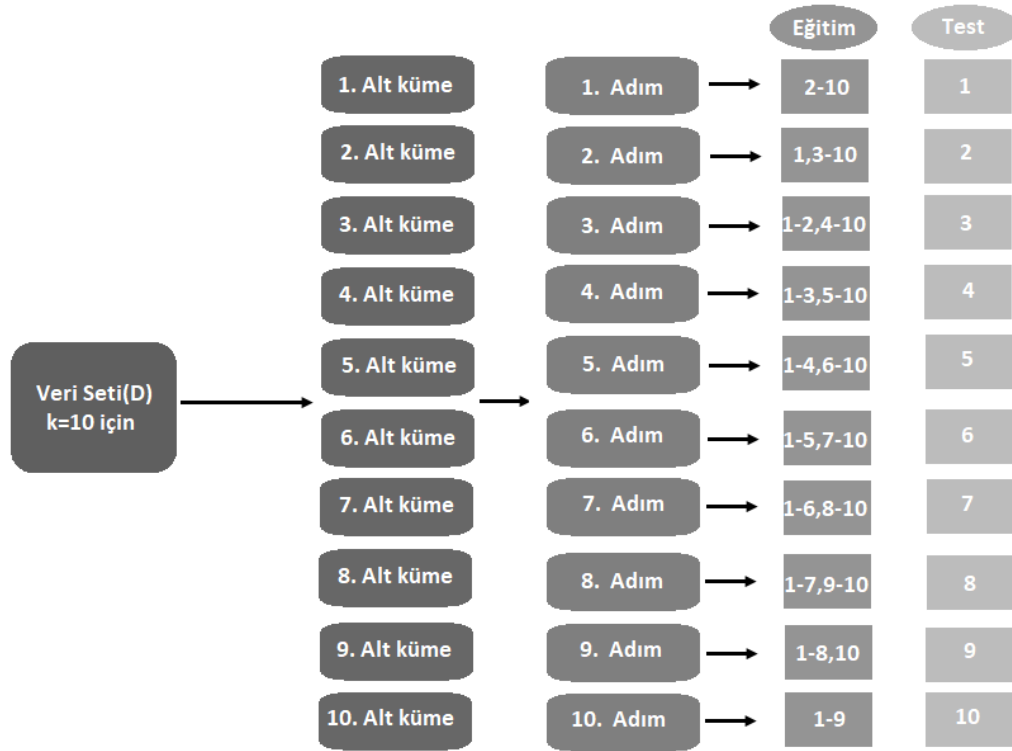
2.6.1. Çapraz Doğrulama (Cross Validation)

Çapraz doğrulama sınıflandırmanın güvenilirliğini artırmak için geliştirilmiş bir yöntemdir. Elimizde var olan veri setinin belirli sayıda alt kümeye bölünmesi işlemiyle başlanır. Alt küme sayısı için k parametresi kullanılır. Kullanılan k parametresi veri setinin kaç tane alt kümeye ayrılacağını belirtir. Genellikle k parametresi için 10 değeri tercih edilir (Bilgin,2018).

Bir D veri kümesi için k parametresi 10 olsun. Bu durumda sistem şu şekilde çalışır:

- D veri kümesi $k=10$ olduğundan 10 adet alt kümeye bölünür.
- İlk iterasyonda 1. alt küme test kümesi olarak ayrılıp, kalan 9 alt küme ($k-1$) ise eğitim kümesi olarak kullanılır.
- Ardındaki her iterasyonda sırasıyla 1 alt küme test kümesi olarak ayrılırken kalan 9 küme eğitim kümesi olarak kullanılır.
- İşlem 10 (k) kez tekrarlanır. Test kümesi her tekrarda farklı bir alt küme olacak şekilde gerçekleştirilir.
- Elde edilen 10 (k) test kümesine ait olan sonuçların ortalaması alınarak işlem sonlandırılır.

Şekilde sistemin çalışma şekli görülmektedir.



Şekil 2-19 D veri kümesini k=10 için çapraz doğrulama işlemi

2.6.2. Karışım Matrisi (Confusion Matrix)

		Gerçek	
		Pozitif	Negatif
Tahmin	Pozitif	TP (Doğru Pozitif)	FP (Yanlış Pozitif)
	Negatif	FN (Yanlış Negatif)	TN (Doğru Negatif)

Doğru Pozitif (TP): Gerçekte hasta olanlardan modelin hasta olarak tahmin ettiği kişilerin sayısıdır.

Doğru Negatif (TN): Gerçekte hasta olmayanlardan modelin hasta değildir biçimde tahmin ettiği kişilerin sayısıdır.

Yanlış Pozitif (FP): Gerçekte hasta olmayanlardan modelin hasta olarak tahmin ettiği kişilerin sayısıdır.

Yanlış Negatif (FN): Gerçekte hasta olanlardan modelin hasta değildir biçimde tahmin ettiği kişilerin sayısıdır.

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

$$\text{Kappa} = \frac{(\text{Gözlenen Doğruluk} - \text{Beklenen Doğruluk})}{(1 - \text{Beklenen Doğruluk})}$$

$$\text{Duyarlılık(Recall)} = \frac{TP}{TP + FN}$$

$$\text{Kesinlik(Precision)} = \frac{TP}{TP + FP}$$

$$\text{F - Ölçütü (F - measure)} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Doğruluk (Accuracy): Modelin başarılı olup olmadığını gösteren en basit ve en sık kullanılan ölçüt doğruluktur.

Duyarlılık (Sensitivity, recall): Gerçekte pozitif bir sonuç elde edildiğinde, test sonucunun da pozitif olması olasılığıdır.

Kesinlik (Precision): Pozitif bir test sonucu elde edildiğinde gerçek sonucunda pozitif olması olasılığıdır.

Özgüllük(Specificity): Gerçekte negatif bir sonuç elde edildiğinde, test sonucunun da negatif olması olasılığıdır.

F-Ölçütü: Duyarlılık ile kesinliğin harmonik ortalamasıdır ve her iki değeri de bir arada ele alma imkanı tanımaktadır.

Kappa: Kestirilen değerler ile gerçek değerlerin uyumunun güvenilirliğinin ölçütüdür.

2.6.3. İşlem Karakteristik Eğrisi (Receiver Operating Characteristics-ROC)

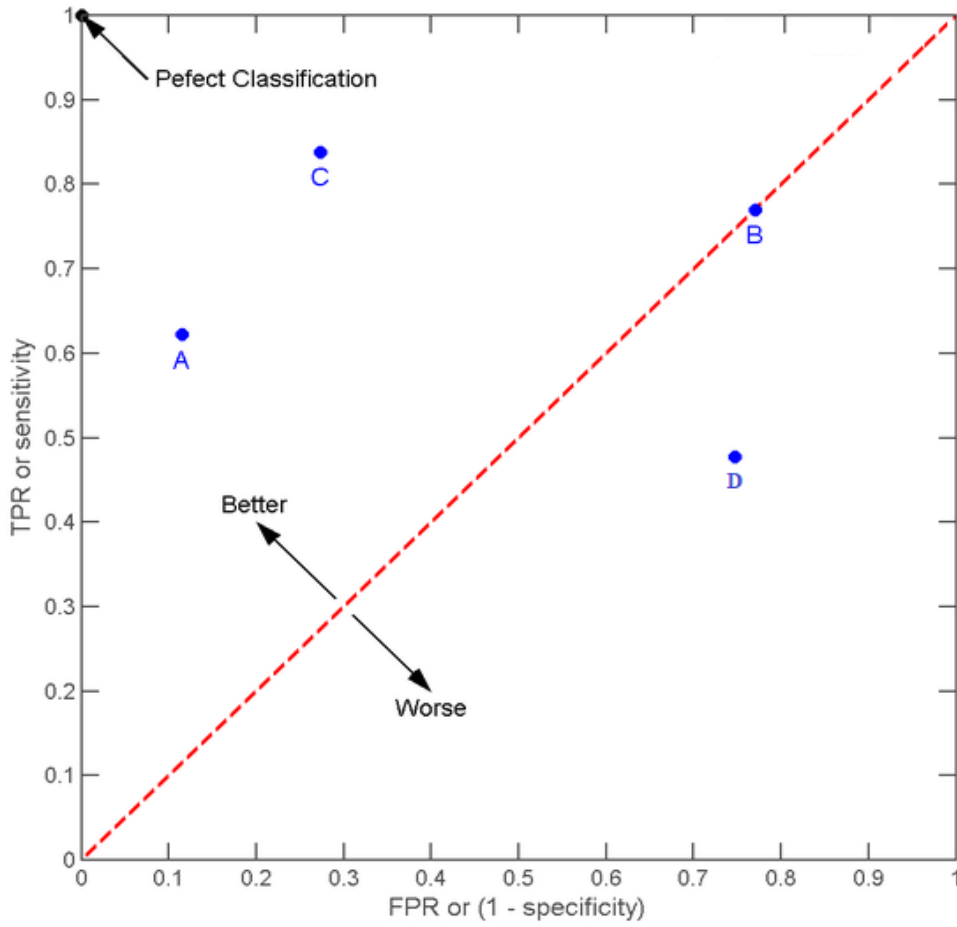
ROC eğrisi ilk olarak sinyal belirleme teorisinde kullanılmış olup daha sonra birçok farklı alanda özellikle sağlık alanında yapılan karar tabanlı çalışmalarda kullanılmıştır. Makine öğrenmesi ve veri madenciliğinde model değerlendirilmesi için tercih edilen önemli bir değer olmuştur.

ROC eğrisinde X-ekseni doğru kabul oranını (true positive rate) gösterirken, Y eksenini yanlış kabul oranını (false positive rate) göstermektedir. ROC eğrisi, sınıflandırıcının doğru kabulleri için hesaplanan yanlış kabul oranlarını gösteren grafikdir (Küçükbay, 2015).

Area Under Curve (AUC, Eğri altındaki alan) ifadesi ROC eğrisindeki sonuçların yorumlanarak sonucun bir sayısal ifade ile açıklanmasına denmektedir. AUC, 0 ve 1 arasında değer alabilir. ROC eğrisi (0,0) ve (1,1) noktaları arasında çizilen bir köşegen şeklindedir. AUC değeri 0.5'ten aşağıda ise bu sistemin hesaplamalarının değersiz olduğu anlamına gelmektedir. Kısaca AUC değerine doğru sınıflandırılan pozitif örneklerin yanlış sınıflandırılan pozitif örneklere bölümü denilebilir (Küçükbay, 2015).

Şekilde gösterilen ROC grafiğinde A noktası şekilden de görüleceği üzere Y eksenine yakın durumdadır. A noktası için doğru kabul oranı ise düşük değerlerde seyretmektedir. B noktası ise daha yüksek yanlış kabul oranına sahiptir. Ama B noktasının doğru kabul oranı A noktasına göre daha yüksek olduğu şekilden görülmektedir. Buradan A değerine sahip bir sistemin doğru kabul işlemi için, belirli özellikler üzerinden hesaplama yaparken doğru sınıflandırma yaptığı sonuca ulaşılabilir. B değeri için ise daha belirsiz özellikler kullanılarak yapılan hesaplamaların daha doğru sonuçlar üretebileceği sonucuna ulaşılabilir (Küçükbay, 2015).

ROC grafikleri kullanılarak mevcut iki sistem arasında seçim yapılabilir. Sistem kurgulanırken nelerin gözardı edilebileceği ve nelere ulaşılabilceği tespit edilebilir. Örneğin şekilde görüldüğü gibi A değeri seçildiğinde yani daha düşük yanlış kabul değeri seçildiğinde, doğru kabulü düşük bir sistem seçilmiş olacaktır (Küçükbay,2015).



Şekil 2-20 Örnek ROC eğrisi

2.7. Hiperparametrelerin Ayarlanması

Hiperparametreler, değerleri verilerden kestirilemeyen parametrelerdir. Model parametreleri doğrusal regresyondaki eğim ve kesme (sabit) gibi modelin eğitimi sırasında öğrenilirken; hiperparametreler eğitmeden önce daha model kurulurken belirlenir. Rastgele Orman modelinde parametreler, değişkenler ve eğitim sırasında her bir düğüm bölmede kullanılan eşikler olurken; hiperparametreler ormandaki karar ağacı sayısı ve düğümü bölmede kullanılacak özellik sayısı olarak karşımıza çıkar (Gürsaka,2018).

Eğitim verileri ile bir modeli optimize ettiğimizde, model eğitim verileri için çok iyi ama test verileri için iyi olmayabilir. Bu durumda, modelin yeni verilere uygulanamaması, genelleştirilememesi sorunu “aşırı uyum” sorunu olarak adlandırılır. Aynen bir radyonun düğmelerini ayarlayarak sesini en net, en cızırtısız ses haline getirmek istersek; hiperparametreleri optimize ederek, ayarlayarak modelimizin performansını artırmayı

amaçlarız. Bu iş, farklı hiperparametre bileşimlerinin deneme-yanılma yoluyla belirlenmesi gibidir (Gürsakal,2018).

Bir rastgele orman algortimasında 4 hiperparametre vardır:

- Ormandaki ağaç sayısı
- Her bölmede göz önüne alınacak özellik sayısı (Varsayılan olarak bu toplam özellik sayısının kareköküdür)
- Düğüm sayısı, ağacın maksimum derinliği
- Ağacın dibinde bir yaprak düğüm (terminal düğüm) olması için minimum örneklem sayısı

Hiperparametre optimizasyonunda “rastgele arama”(random search) ve en sık kullanılan tekniklerden birisi ise, “ızgara arama” (grid search) tekniğidir. Rastgele arama tekniğinde hiperparametre kombinasyonları teker teker denenmek yerine rastgele olarak seçilerek denenir (Gürsakal,2018).

3. GEREÇ VE YÖNTEM

Bu çalışmada astım, KOAH ve pnömoni hastalık grupları ve bir sağlıklı gruptan oluşan veri seti kullanılmıştır. Grupların her biri 1000'er gözlemden oluşmak üzere veri seti toplam 4000 gözlemden oluşmaktadır. Çalışmada 14 ayrı açıklayıcı değişken kullanılmıştır. Tüm veriler 2015-2018 yılları arasında İstanbul'da Yedikule Göğüs Hastalıkları ve Göğüs Cerrahisi Eğitim ve Araştırma Hastanesi'ne başvuran kişilerden toplanmıştır. S. B. Ü. İstanbul Eğitim ve Araştırma Hastanesi Klinik Araştırmalar Etik Kurulu'ndan çalışma izni alınmıştır. Tıbbi ölçümler (açıklayıcı değişkenler) havanın soluma hacmini sağlayan cihaz olan spirometre kullanılarak elde edilmiştir. Kalan değişkenler kişilerin özelliklerini tanımlar ve hastane veri tabanından alınmıştır. Bağımlı değişken olarak astım, KOAH ve pnömoni hastalıkları ile sağlıklı grup düşünülmüştür. Spirometrenin ölçtüğü tıbbi değişkenler yani solunum fonksiyon testleri şunlardır: FEV1, FVC, PEF, MEF75, MEF50, MEF25, MMEF, FEV3, PIF, TIFF dir. Diğer açıklayıcı değişkenler ise cinsiyet, yaş, acil başvuru sayısı ve poliklinik başvuru sayısıdır.

Solunum fonksiyon testlerinde mutlak değerler litre olarak, o yaş grubundaki erkek veya kadın cinsine göre olması gereken mutlak değere göre öngörülen yüzdesi ((ölçülen değer ml/ olması gereken değer ml)*100) olarak verilmiştir. Gözlemler takipteki hastalardan seçildiği için solunum fonksiyon testleri postbronkodilatör değil bazal solunum fonksiyon testi değerleri olarak çalışmada değerlendirilmiştir.

Spirometre ölçümleri hastanenin SFT laboratuvarında spirometre cihazlarıyla otomatik olarak ölçülmüş ve kaydedilmiştir. Statik ve dinamik ölçümler yapılmıştır. Hastanın volümetrik ölçümleri litre olarak ölçülen değer ve yaş ve cinsiyete göre öngörülen değere göre de yüzdelere kaydedilmiştir.

Çalışmada sağlıklı grup ile hastalık gruplarının nitel değişkenlerin karşılaştırmada Ki Kare testi kullanılmıştır. Sürekli değişkenlerin karşılaştırması için öncelikle normal dağılıma uyup uymadığı Kolmogorov Smirnov testi ile kontrol edilmiştir. Normal dağılım koşulunu sağlayan değişkenler için Welch t testi, sağlamayan değişkenler için Mann-Whitney U testi kullanılmıştır. Değişkenler arasındaki ilişkiler Spearman korelasyon analizi ile yapılmıştır.

Çalışmanın asıl amacında, danışmanlı öğrenme yöntemlerinden en başarılı ve tanınmış on makine öğrenmesi algoritması değerlendirilmiştir. Bu algoritmalar Naif Bayes, Lojistik regresyon, Yapay sinir ağları, Suport Vector Machine (Destek Vektör Makinesi

(doğrusal ve doğrusal olmayan çekirdeklerle)), KNN (K-en yakın komşu), Karar ağacı, Bagging (Torbalama), Random Forest (Rastgele Orman (CART temelli öğrenen ile)), Random Forest (Rastgele Orman (CIT temelli öğrenen ile)), Gradient Boosting Makinesi ve C5.0 dır. Deney sonuçları R (Versiyon 3.6.0) ve IBM SPSS (Versiyon 21.0) yazılımları ile elde edilmiştir.

Bu algoritmaları uygulamadan önce, veri setleri, sırasıyla % 80 ve % 20 oranlarıyla eğitim ve test verileri olarak iki kısma ayrılmıştır. Naif Bayes, Lojistik regresyon ve Bagging (Torbalama) dışındaki tüm algoritmalar, dikkatlice ayarlanması gereken bazı hiperparametrelere sahiptir. Hiperparametreleri ayarlamak için eğitim veri setinde ızgara arama yaklaşımı kullanılmıştır. Uygun veri parametrelerini belirlemek için eğitim veri kümeleri kullanılarak tekrarlanan k-kat çapraz doğrulama yaklaşımı kullanılmıştır. Bu yaklaşımda, her veri seti k-katlara bölünür ve bu işlem belirli sayıda tekrarlanır. Tekrarların sonuçları kullanılarak elde edilen ortalama performans değeri nihai model performansı olarak alınır. Bu çalışmada beş kez tekrarlanan 10 kat çapraz doğrulama süreci göz önünde bulundurulmuştur. Eğitim veri setine dayalı olarak her algoritma için en iyi olarak bulunan hiperparametreler, görünmeyen veri setini (test verileri) kullanarak modelleri doğrulamak için kullanılmıştır.

Algoritmaların hiperparametre tarama uzayı aşağıda listelenmiştir.

- Yapay Sinir Ağları: Gizli katman birimleri boyutu (size) ve zayıflama (decay) değeri aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
Size: {5,10, ...,50}
Decay: {0,0.005,0.010, ...,0.095,1.0}
- Destek Vektör Makinesi (doğrusal çekirdek): Maliyet (Cost) değeri (C) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
C: {0.01,0.02, ...,0.05,0.10,0.15, ...,0.95,1.0,2, 2², ..., 2⁵}
- Destek Vektör Makinesi (Doğrusal olmayan çekirdek): Maliyet (Cost) değeri (C) ve sigma değeri aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
C: {0.01,0.02, ...,0.05,0.10,0.15, ...,0.95,1.0,2, 2², ..., 2⁵}
Sigma: {0,0.005, ...,0.095,1.0}
- KNN: Komşu sayısı (k) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
k: {1,2, ...,40}
- Karar ağacı: Karmaşıklık parametresi (c_p) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
c_p: {0,0.01, ...,0.09, 0.1,0.2, ... 0.9,1.0}

- Rasgele Orman (CART temelli öğrenen ile): Her ağaç için rasgele seçilen tahmin edici sayısı (m_{try}) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
 $m_{try}: \{1,2, \dots, 14,15\}$
- Rasgele Orman (CIT temelli öğrenen ile): Her ağaç için rasgele seçilen tahmin edici sayısı (m_{try}) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
 $m_{try}: \{1,2, \dots, 14,15\}$
- Gradient Boosting Makinesi: Boosting iterasyon sayısı ($n.trees$), maksimum ağaç derinliği değeri ($interaction.depth$), büzülme değeri ($shrinkage$) ve minimum terminal düğüm büyüklüğü değeri ($n.minobsinnode$) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
 $n.trees: \{10,20, \dots, 200\}$
 $interaction.depth: \{1,2, \dots, 5\}$
 $shrinkage: \{0,0.1, \dots, 0.5\}$
 $n.minobsinnode: \{5,10,15\}$
- C5.0: Boosting iterasyon sayısı ($trials$), model yapısı türü ($model$) ve eleme koşulu ($winnow$) aşağıdaki tarama uzayı göz önünde bulundurularak çapraz doğrulama ile seçilir:
 $trials: \{1,2, \dots, 5,10, \dots, 50\}$
 $model: \{Tree\}$
 $winnow: \{TRUE\}$

4. BULGULAR

4.1. Değişkenler ve özellikleri

Veri setinin özellikleri ve detayları Tablo 4-1'de verilmiştir. Değişkenler türleri ve özellikleri ile açıklanmıştır.

Tablo 4-1 Değişkenler ve açıklamalarının listesi

No	Değişkenler	Türü	Açıklamalar
1	Yaş	Kategorik	Yaş grupları ve kodları: 1: (0-14) / 2: (15-29) / 3: (30-44) 4: (45-59) / 5: (60-74) / 6: (75+)
2	Cinsiyet	Kategorik	Kadın /Erkek
3	Acil	Nümerik	Acile başvuru sayısı
4	Poliklinik	Nümerik	Polikliniğe başvuru sayısı
5	FEV1	Nümerik	1 saniyede zorlu ekspirasyon hacmi
6	FVC	Nümerik	Zorlu vital kapasite
7	PEF	Nümerik	Tepe ekspiratuar akış
8	MEF75	Nümerik	FVC eğrisinin% 75'inde maksimum ekspiratuar akış
9	MEF50	Nümerik	FVC eğrisinin% 50'sinde maksimum ekspiratuar akış
10	MEF25	Nümerik	FVC eğrisinin% 25'inde maksimum ekspiratuar akış
11	MMEF	Nümerik	Maksimum-orta ekspirasyon akışı
12	FEV3	Nümerik	3 saniyede zorlu ekspirasyon hacmi
13	PIF	Nümerik	Tepe inspiratuar akış
14	TIFF	Nümerik	Tiffeneau–Pinelli indeksi (FEV1/FVC)

4.2. Grupların tanımlayıcı istatistikleri

Makine öğrenme algoritmalarını uygulamadan önce, veri seti hakkında daha iyi bir anlayış sağlamak amacıyla astım, KOAH ve pnömoni hastalık grupları ile sağlıklı grup veri seti için bazı tanımlayıcı istatistikler sunulmuştur. Tablo 4-1, Tablo 4-2, Tablo 4-3 ve Tablo 4-4' te her bir grubun nitel değişkenleri için bazı tanımlayıcı istatistikleri verilmiştir. Tablo 4-5, Tablo 4-6, Tablo 4-7 ve Tablo 4-8'de ise grupların nicel değişkenleri için tanımlayıcı istatistikler sunulmuştur.

Tablo 4-2 Sağlıklı grup için nitel değişkenlerin tanımlayıcı istatistikleri

Değişkenler	Sıklık	Yüzde
Yaş	0-14	1
	15-29	14
	30-44	108
	45-59	362
	60-74	372
	75+	143
Cinsiyet	Erkek	538
	Kadın	462

Sağlıklı grubun nitel değişkenleri istatistikleri tablosunda yaş bakımından 0-14 yaş aralığında 1 kişi olup en sık olan 60-74 aralığıdır. Ayrıca erkek sayısı kadınlara göre daha fazladır.

Tablo 4-3 Astım grubu için nitel değişkenlerin tanımlayıcı istatistikleri

Değişkenler	Sıklık	Yüzde
Yaş	0-14	0
	15-29	11
	30-44	79
	45-59	327
	60-74	473
	75+	110
Cinsiyet	Erkek	595
	Kadın	405

Astım grubu nitel değişken tablosunda yaş değişkeninin 0-14 aralığında vaka olmayıp en sık olduğu aralık 60-74 aralığıdır. Erkek sayısı kadınlardan fazladır.

Tablo 4-4 KOAH grubu için nitel değişkenlerin tanımlayıcı istatistikleri

Değişkenler	Sıklık	Yüzde
Yaş	0-14	0
	15-29	6
	30-44	45
	45-59	300
	60-74	497
	75+	152
Cinsiyet	Erkek	689
	Kadın	311

KOAH grubu nitel değişken tablo istatistiklerinde yine yaşın 0-14 aralığında vaka olmayıp en sık olan aralık 60-74 aralığıdır. Yine erkek sayısı kadınlardan fazladır.

Tablo 4-5 Pnömoni grubu için nitel değişkenlerin tanımlayıcı istatistikleri

Değişkenler	Sıklık	Yüzde
Yaş	0-14	3
	15-29	40
	30-44	108
	45-59	298
	60-74	388
	75+	163
Cinsiyet	Erkek	579
	Kadın	421

Pnömoni grubu nitel değişkenleri tablosu ele alındığında yaşın 0-14 aralığında 3 vaka olup en sık olduğu aralık yine 60-74 aralığıdır. Yine erkek sayısı kadınlardan fazladır.

Tablo 4-1, 4-2 ,4-3 ,4-4 incelendiğinde dört grupta da yaş bakımından en yüksek yüzdeye sahip 60-74 aralığıdır. Ayrıca cinsiyette erkek yüzdesi kadınlara göre biraz daha fazladır.

Tablo 4-6 Sağlıklı grup için nicel değişkenlerin tanımlayıcı istatistikleri

	Ortalama	Standart Hata	Minimum	Maksimum	Değişim Aralığı	Standart Sapma	Varyans
FEV1	84.42	0.55	22.91	147.00	124.10	17.62	310.51
FVC	84.64	0.61	18.70	129.20	133.80	19.57	383.30
PEF	68.24	0.63	18.40	130.30	111.90	20.06	402.51
MEF75	68.21	0.64	5.76	139.10	133.34	20.28	411.28
MEF50	67.39	0,70	14.03	164.20	150.17	22.37	500.81
MEF25	66.07	1.14	9.80	235.00	225.20	36.33	1320.22
MMEF	64.62	0.71	15.45	155.70	140.25	22.68	514.81
FEV3	96.93	0.11	87.63	100.98	13.35	3.55	12.63
PIF	99.39	1.03	24.77	215.61	190.84	32.69	1068.94
TIFF	101.00	0.34	60.23	135.16	74.92	11.02	121.50
Acil	1.43	0.09	0.00	40.00	40.00	3.15	9.96
Poliklinik	4.69	0.16	0.00	42.00	42.00	5.25	27.58

Sağlıklı grubun nicel değişkenleri incelendiğinde solunum fonksiyon testlerinden TIFF değeri ortalaması en yüksek olup MMEF değeri ortalaması en düşük görülmektedir. Ayrıca acil başvuru sayısı ortalaması poliklinik başvuru sayısı ortalamasına göre daha düşüktür.

Tablo 4-7 Astım grubu için nicel değişkenlerin tanımlayıcı istatistikleri

	Ortalama	Standart Hata	Minimum	Maksimum	Değişim Aralığı	Standart Sapma	Varyans
FEV1	73.90	0.28	40.70	126.30	85.60	8.96	80.36
FVC	76.78	0.38	33.96	152.50	95.24	12.02	144.70
PEF	60.18	0.54	21.80	122.10	100.30	17.21	296.45
MEF75	56.32	0.53	13.60	135.40	121.80	16.81	282.84
MEF50	51.58	0.59	10.00	134.00	124.00	18.73	351.00
MEF25	54.07	0.94	8.40	229.94	221.54	29.95	897.33
MMEF	63.00	0.75	10.70	155.70	145.00	23.77	565.25
FEV3	95.73	0.16	56.50	100.99	44.49	5.37	28.85
PIF	86.71	1.05	21.87	215.61	193.74	33.27	1107.18
TIFF	97.55	0.38	57.07	133.54	76.47	12.24	150.02
Acil	1.44	0.09	0.00	30.00	30.00	2.86	8.22
Poliklinik	4.14	0.14	0.00	42.00	42.00	4.45	19.84

Astım grubu nicel değişkenleri tablosuna göre en düşük ortalamaya sahip olan MEF75 olup en yüksek ortalaması olan değer TIFF değeridir. Acil başvuru sayısı ortalaması poliklinik başvuru sayısı ortalamasına göre daha düşüktür.

Tablo 4-8 KOAH grubu için nicel değişkenlerin tanımlayıcı istatistikleri

	Ortalama	Standart Hata	Minimum	Maksimum	Değişim Aralığı	Standart Sapma	Varyans
FEV1	61.45	0.25	53.50	96.40	42.90	7.94	63.05
FVC	68.04	0.37	44.60	110.70	66.10	11.75	138.12
PEF	51.49	0.49	16.20	112.70	96.50	15.55	241.98
MEF75	44.02	0.49	15.20	111.10	95.90	15.70	246.53
MEF50	36.83	0.48	9.40	104.64	95.24	15.33	235.29
MEF25	41.52	0.78	6.90	235.70	228.80	24.66	608.9
MMEF	56.71	0.81	13.00	155.70	142.70	25.85	668.3
FEV3	95.35	0.16	74.89	100.98	26.09	5.25	27.62
PIF	87.17	0.84	52.55	180.27	127.72	26.63	709.40
TIFF	139.36	1.31	52.17	286.77	234.60	41.67	1736.41
Acil	1.55	0.39	0.00	263.00	263.00	12.42	154.34
Poliklinik	2.76	0.10	0.00	32.00	32.00	3.30	10.92

KOAH grubu nicel değişkenlerin tanımlayıcı istatistiklerine göre en düşük ortalama PEF olup en yüksek yine TIFF değeridir. Acil başvuru sayısı ortalaması poliklinik başvuru sayısı ortalamasına göre daha düşüktür.

Tablo 4-9 Pnömoni grubu için nicel değişkenlerin tanımlayıcı istatistikleri

	Ortalama	Standart Hata	Minimum	Maksimum	Değişim Aralığı	Standart Sapma	Varyans
FEV1	72.58	0.53	29.70	140.70	111.00	17.01	289.54
FVC	76.43	0.53	59.90	152.50	92.60	16.76	280.89
PEF	58.77	0.64	15.71	124.80	109.09	20.54	422.20
MEF75	55.44	0.70	8.43	134.80	126.37	22.27	496.34
MEF50	51.88	0.76	7.70	164.20	156.50	24.31	591.07
MEF25	53.33	1.08	5.53	296.80	291.27	34.45	1186.82
MMEF	50.38	0.74	7.70	158.60	150.90	23.65	559.74
FEV3	92.14	0.21	69.06	100.00	30.94	6.95	48.36
PIF	90.32	1.14	18.34	211.82	193.48	36.18	1309,37
TIFF	95.43	0.46	45.31	135.90	90.59	14.67	215,24
Acil	4.51	0.03	1.00	6.00	5.00	1.03	1.06
Poliklinik	5.81	0.22	0.00	51.00	51.00	6.95	48.41

Pnömoni grubu nicel değişkenlerin tanımlayıcı istatistik tablosuna göre en düşük MMEF değeri olup en yüksek değer diğer gruplarda da olduğu gibi TIFF değeridir. Yine diğer tablolarda da olduğu gibi acil başvuru sayısı ortalaması poliklinik başvuru sayısı ortalamasına göre daha düşüktür.

Tablo 4-5, 4-6, 4-7, 4-8 incelendiğinde dört grupta da solunum fonksiyon test değerlerinden en yüksek ortalamaya sahip TIFF değeri bulunmuştur. En düşük ortalamaya sahip olan solunum fonksiyon testi ise gruplara göre farklılık göstermektedir. Acil ve poliklinik başvuru sayıları göz önüne alındığında acil başvuru sayısı, poliklinik başvuru sayısına göre daha az olup ayrıca pnömoni grubunda en yüksek ortalama değerlerine sahip bulunmuşlardır.

4.3. Grupların karşılaştırılması ve Korelasyon analizi

4.3.1. Nitel değişkenlerin karşılaştırılması

Sağlıklı grup ile astım, KOAH, pnömoni hastalık gruplarının cinsiyet ve yaş grupları bakımından karşılaştırılması aşağıdaki tablolarda gösterilmiştir.

Tablo 4-10 Sağlıklı grup ile astım grubu karşılaştırılması

		Sağlıklı		Astım		Toplam		
		Sayı	Yüzde	Sayı	Yüzde	Sayı	Yüzde	p
Cinsiyet	Erkek	538	53.8	595	59.5	1133	56.7	0.010
	Kadın	462	46.2	405	40.5	867	43.4	
Yaş	0-14	1	0.1	0	0.0	1	0.1	<0.001
	15-29	14	1.4	11	1.1	25	1.3	
	30-44	108	10.8	79	7.9	187	9.4	
	45-59	362	36.2	327	32.7	689	34.5	
	60-74	372	37.2	473	47.3	845	42.3	
	75+	143	14.3	110	11.0	243	12.7	
Toplam		1000	100.0	1000	100.0	2000	100.0	

Sağlıklı grup ile astım hastaları arasında hem cinsiyet hem de yaş bakımından istatistiksel açıdan anlamlı fark görülmektedir. Astım görülme sıklığı erkeklerde daha fazladır. Ayrıca 60-74 yaş aralığında da astım daha fazla görülmektedir.

Tablo 4-11 Sağlıklı grup ile KOAH grubu karşılaştırılması

		Sağlıklı		KOAH		Toplam		
		Sayı	Yüzde	Sayı	Yüzde	Sayı	Yüzde	p
Cinsiyet	Erkek	538	53.8	689	68.9	1227	61.4	<0.001
	Kadın	452	46.2	311	31.1	773	38.7	
Yaş	0-14	1	0.1	0	0.0	1	0.1	<0.001
	15-29	14	1.4	6	0.6	20	1.0	
	30-44	108	10.8	45	4.5	153	7.7	
	45-59	362	36.2	300	30.0	662	33.1	
	60-74	372	37.2	497	49.7	869	43.5	
	75+	143	14.3	152	15.2	295	14.8	
Toplam		1000	100.0	1000	100.0	2000	100.0	

Sağlıklı grup ile KOAH grubu kıyaslandığında yine her iki grup arasında hem cinsiyet hem de yaş bakımından anlamlı fark bulunmuştur. KOAH sıklığı erkeklerde daha fazla görülmektedir. Ayrıca 45-59, 60-74, 75+ yaş aralığında KOAH sıklığı daha fazladır.

Tablo 4-12 Sağlıklı grup ile pnömoni grubu karşılaştırılması

		Sağlıklı		Pnömoni		Toplam		p
		Sayı	Yüzde	Sayı	Yüzde	Sayı	Yüzde	
Cinsiyet	Erkek	538	53.8	579	57.9	1117	55.9	<0.065
	Kadın	462	46.2	421	42.1	8833	44.2	
Yaş	0-14	1	0.1	3	0.3	4	0.2	0.001
	15-29	14	1.4	40	4.0	54	2.7	
	30-44	108	10.8	108	10.8	216	10.8	
	45-59	362	36.2	298	29.8	660	33.0	
	60-74	372	37.2	388	38.8	760	38.0	
	75+	143	14.3	163	16.3	306	15.3	
Toplam		1000	100.0	1000	100.0	2000	100.0	

Sağlıklı grup ile pnömoni hastaları arasında yaş bakımından anlamlı fark vardır. 15-29 aralığında pnömoni daha fazladır. Ayrıca sağlıklı grup ile pnömoni hastaları cinsiyet bakımından anlamlı fark bulunamamıştır.

4.3.2. Nicel değişkenlerin karşılaştırılması

4.3.2.1. Solunum fonksiyon testlerinin karşılaştırılması

Tıbbi nicel değişkenlere dayanarak her grubun istatistiksel olarak anlamlı olup olmadığına karar vermek için Welch t testi kullanılmıştır. Gruplar arasındaki korelasyon analizi için Spearman korelasyon katsayısı kullanılmıştır. Sonuçlar Tablo 4-13' te verilmiştir. En önemli değişkenlerin belirlenmesinde doğruluk sağlamak ve bazı kanıtlar elde etmek için hem korelasyon analizi hem de istatistiksel anlamlılık testi yapılmıştır.

Tablo 4-13 Sağlıklı grup ile astım grubu karşılaştırılması

	Welch t test		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
FEV1	✓	p < 0.01	✓	p < 0.01
FVC	✓	p < 0.01	✓	p < 0.01
PEF	✓	p < 0.01	✓	p < 0.01
MEF75	✓	p < 0.01	✓	p < 0.01
MEF50	✓	p < 0.01	✓	p < 0.01
MEF25	✓	p < 0.01	✓	p < 0.01
MMEF	✗	p = 0.169	✗	p = 0.119
FEV3	✓	p = 0.012	✓	p < 0.01
PIF	✓	p < 0.01	✓	p < 0.01
TIFF	✓	p < 0.01	✓	p < 0.01

Sağlıklı grup ile astım grubunun solunum fonksiyon testleri bakımından karşılaştırılmasında MMEF hariç diğer değişkenler arasında anlamlı fark bulunmuştur.

Ayrıca korelasyon katsayıları da MMEF değişkeni hariç diğer değişkenler istatistiksel açıdan anlamlıdır.

Tablo 4-14 Sağlıklı grup ile KOAH grubu karşılaştırılması

	Welch t test		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
FEV1	✓	$p < 0.01$	✓	$p < 0.01$
FVC	✓	$p < 0.01$	✓	$p < 0.01$
PEF	✓	$p < 0.01$	✓	$p < 0.01$
MEF75	✓	$p < 0.01$	✓	$p < 0.01$
MEF50	✓	$p < 0.01$	✓	$p < 0.01$
MEF25	✓	$p < 0.01$	✓	$p < 0.01$
MMEF	✓	$p < 0.01$	✓	$p < 0.01$
FEV3	✓	$p < 0.01$	✓	$p < 0.01$
PIF	✓	$p < 0.01$	✓	$p < 0.01$
TIFF	✓	$p < 0.01$	✓	$p < 0.01$

Sağlıklı grup ile KOAH grubu arasındaki karşılaştırmada tüm değişkenler açısından anlamlı fark olup korelasyon katsayıları da anlamlı bulunmuştur.

Tablo 4-15 Sağlıklı grup ile pnömoni grubun karşılaştırılması

	Welch t test		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
FEV1	✓	$p < 0.01$	✓	$p < 0.01$
FVC	✓	$p < 0.01$	✓	$p < 0.01$
PEF	✓	$p < 0.01$	✓	$p < 0.01$
MEF75	✓	$p < 0.01$	✗	$p = 0.301$
MEF50	✓	$p < 0.01$	✗	$p = 0.932$
MEF25	✓	$p < 0.01$	✓	$p < 0.01$
MMEF	✓	$p < 0.01$	✗	$p = 0.118$
FEV3	✓	$p < 0.01$	✓	$p < 0.01$
PIF	✓	$p < 0.01$	✓	$p < 0.01$
TIFF	✓	$p < 0.01$	✓	$p < 0.01$

Sağlıklı grup ile pnömoni grubu karşılaştırılmasında yine tüm değişkenler anlamlı bulunmuştur. Ancak korelasyon katsayısı bakımından MEF75, MEF50 ve MMEF değişkenleri hariç diğer değişkenler anlamlıdır.

4.3.2.2. Acil ve poliklinik başvuru sayılarının karşılaştırılması

Tablo 4-16 Sağlıklı grup ile astım grubu karşılaştırılması

	Mann Whitney U		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
Acil	✗	p = 0.816	✗	p = 0.650
Poliklinik	✗	p = 0.065	✗	p = 0.407

Sağlıklı grup ile astımın acil ve poliklinik başvuru sayısı bakımından karşılaştırılmasında anlamlı fark yoktur. Korelasyon katsayıları da anlamlı bulunmamıştır.

Tablo 4-17 Sağlıklı grup ile KOAH grubu karşılaştırılması

	Mann Whitney U		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
Acil	✓	p < 0.01	✗	p = 0.549
Poliklinik	✓	p < 0.01	✗	p = 0.264

Sağlıklı grup ve KOAH grubu karşılaştırılmasında hem acil hem de poliklinik başvuru sayısı bakımından anlamlı fark vardır. Ancak korelasyon katsayısı anlamlı değildir.

Tablo 4-18 Sağlıklı grup ile pnömoni grubu karşılaştırılması

	Mann Whitney U		Spearman Korelasyon	
	Anlamlılık	p	Anlamlılık	p
Acil	✓	p < 0.01	✗	p = 0.588
Poliklinik	✓	p = 0.01	✗	p = 0.142

Sağlıklı grup ile pnömoni arasında acil ve poliklinik başvuru sayısı bakımından karşılaştırmada anlamlı fark olup korelasyon katsayısı bakımından anlamlılık bulunmamıştır.

4.4. Makine öğrenmesi algoritmalarının astım, KOAH, pnömoni hastalık grupları ve sağlıklı grup sınıflandırılmasında performanslarının karşılaştırmaları

Tablo 4-19 Sağlıklı grup ve astım grubu sınıflandırması için performans karşılaştırmaları

Algoritmalar	Veri	Doğruluk	95% G.A.	Kappa	Kesinlik	Duyarlılık	F1
Naif Bayes	Eğitim	0.780	(0.758 - 0.800)	0.560	0.765	0.788	0.776
	Test	0.747	(0.701 - 0.789)	0.495	0.720	0.761	0.740
Lojistik Regresyon	Eğitim	0.799	(0.778 - 0.818)	0.598	0.775	0.814	0.794
	Test	0.797	(0.754 - 0.835)	0.595	0.780	0.808	0.793
YSA	Eğitim	0.919	(0.904 - 0.932)	0.838	0.897	0.938	0.917
	Test	0.922	(0.891 - 0.946)	0.845	0.890	0.951	0.919
SVM (Linear)	Eğitim	0.805	(0.784 - 0.824)	0.610	0.766	0.830	0.797
	Test	0.810	(0.768 - 0.847)	0.620	0.780	0.829	0.804
SVM (Non Linear)	Eğitim	0.950	(0.938 - 0.960)	0.901	0.910	0.990	0.948
	Test	0.907	(0.874 - 0.934)	0.815	0.860	0.950	0.902
KNN	Eğitim	0.889	(0.873 - 0.904)	0.777	0.855	0.918	0.885
	Test	0.820	(0.778 - 0.859)	0.640	0.795	0.836	0.815
Karar ağaçları	Eğitim	0.940	(0.927 - 0.951)	0.880	0.933	0.945	0.939
	Test	0.875	(0.838 - 0.905)	0.750	0.890	0.864	0.876
Bagging	Eğitim	0.999	(0.996 - 1.000)	0.997	0.998	1.000	0.999
	Test	0.932	(0.903 - 0.955)	0.865	0.925	0.939	0.932
Random Forest (CART)	Eğitim	1.000	(0.997 - 1.000)	1.000	1.000	1.000	1.000
	Test	0.947	(0.920 - 0.967)	0.895	0.930	0.963	0.946
Random Forest (CIT)	Eğitim	0.988	(0.981 - 0.992)	0.976	0.982	0.993	0.988
	Test	0.922	(0.891 - 0.946)	0.845	0.875	0.966	0.918
Gradient Boosting Makinesi	Eğitim	0.997	(0.993 - 0.999)	0.995	0.997	0.997	0.997
	Test	0.927	(0.897 - 0.950)	0.855	0.910	0.943	0.926
C5.0	Eğitim	1.000	(0.997 - 1.000)	1.000	1.000	1.000	1.000
	Test	0.947	(0.920 - 0.967)	0.895	0.940	0.954	0.947

Tablo 4-19'daki sağlıklı grup ve astım grubu verilerinin sonuçlarına göre; doğruluk, kappa, kesinlik, duyarlılık ve F1 ölçüm değerleri dahil olmak üzere performans kriterleri açısından en iyi eğitim performansı, Bagging, Random Forest (CART), Random Forest (CIT), GBM ve C5.0 algoritmaları ile sağlanmıştır. Modeli, mümkün olduğu kadar az hataya zorlamak için tüm algoritmaların eğitim verileriyle daha iyi performans gösterme eğiliminde olduğu unutulmamalıdır. En kötü eğitim performansları sırasıyla Naif Bayes, Lojistik Regresyon ve SVM(linear) yani Destek Vektör Makinesi (doğrusal) algoritmaları

kullanılarak elde edilmiştir. Test performansına dayalı algoritmalar incelendiğinde, Random Forest (CART) ve C5.0 en yüksek doğruluk, kappa, kesinlik, duyarlılık ve F1 ölçüm değerlerini veren en iyi algoritmalar olarak bulunmuştur. Bu algoritmaları takiben benzer ve iyi performansa sahip algoritmalar Random Forest (CIT), Bagging, GBM ve Yapay Sinir Ağlarıdır. Öte yandan, Naif Bayes, Lojistik Regresyon ve Destek Vektör Makinesi (doğrusal) algoritmaları, diğer algoritmalarından daha zayıf test performansı göstermiştir.

Tablo 4-20 Sağlıklı grup ve KOAH grubu sınıflandırması için performans karşılaştırmaları

Algoritmalar	Veri	Doğruluk	%95 G.A.	Kappa	Kesinlik	Duyarlılık	F1
Naif Bayes	Eğitim	0.859	(0.841 - 0.876)	0.718	0.858	0.859	0.859
	Test	0.845	(0.805 - 0.879)	0.690	0.815	0.867	0.840
Lojistik Regresyon	Eğitim	0.895	(0.879 - 0.910)	0.791	0.903	0.889	0.896
	Test	0.892	(0.857 - 0.921)	0.785	0.900	0.886	0.893
YSA	Eğitim	0.900	(0.889 - 0.918)	0.810	0.918	0.894	0.906
	Test	0.897	(0.863 - 0.925)	0.795	0.900	0.895	0.897
SVM (Linear)	Eğitim	0.901	(0.885 - 0.915)	0.802	0.900	0.902	0.901
	Test	0.897	(0.863 - 0.925)	0.795	0.895	0.899	0.897
SVM (Non-Linear)	Eğitim	0.961	(0.950 - 0.970)	0.922	0.973	0.950	0.961
	Test	0.960	(0.935 - 0.977)	0.920	0.970	0.951	0.960
KNN	Eğitim	0.920	(0.906 - 0.933)	0.841	0.918	0.922	0.920
	Test	0.905	(0.871 - 0.931)	0.810	0.880	0.926	0.902
Karar ağaçları	Eğitim	0.943	(0.931 - 0.954)	0.887	0.946	0.941	0.943
	Test	0.920	(0.888 - 0.944)	0.840	0.920	0.920	0.920
Bagging	Eğitim	0.993	(0.988 - 0.997)	0.987	0.997	0.990	0.993
	Test	0.940	(0.912 - 0.961)	0.880	0.955	0.927	0.940
Random Forest (CART)	Eğitim	0.994	(0.989 - 0.997)	0.988	1.000	0.988	0.994
	Test	0.952	(0.926 - 0.971)	0.905	0.965	0.941	0.953
Random Forest (CIT)	Eğitim	0.948	(0.936 - 0.988)	0.896	0.957	0.939	0.948
	Test	0.922	(0.891 - 0.946)	0.845	0.925	0.920	0.922
Gradient Boosting Makinesi	Eğitim	0.970	(0.961 - 0.978)	0.941	0.983	0.958	0.971
	Test	0.955	(0.926 - 0.971)	0.905	0.970	0.937	0.953
C5.0	Eğitim	0.994	(0.989 - 0.997)	0.988	1.000	0.998	0.994
	Test	0.955	(0.929 - 0.973)	0.910	0.970	0.941	0.955

Tablo 4-20’ deki sağlıklı grup ve KOAH verilerinin sonuçlarına göre, doğruluk, kappa, kesinlik, duyarlılık ve F1 ölçüm değerleri dahil olmak üzere performans kriterleri açısından en iyi eğitim performansı C5.0 ve Random Forest (CART) algoritmaları ile

sağlanmıştır. En kötü eğitim performansları sırasıyla Naive Bayes, Lojistik Regresyon ve Yapay Sinir Ağları algoritmaları kullanılarak elde edilmiştir. Test performansına dayalı algoritmalar incelendiğinde, SVM (Non-Linear) yani Destek Vektör Makinesi (doğrusal olmayan) en iyi algoritma olarak bulunmuş ve en yüksek doğruluk, kappa, precision, recall ve F1 ölçüm değerlerini vermiştir. DVM (doğrusal olmayan) algoritmasının ardından, GBM ve C5.0, DVM (doğrusal olmayan) ile oldukça yakın bir performansa sahiptir. Öte yandan, Naive Bayes, Lojistik Regresyon, Yapay Sinir Ağları ve Destek Vektör Makinesi (doğrusal) algoritmaları, diğer algoritmalarından daha zayıf test performansı göstermiştir.

Tablo 4-21 Sağlıklı grup ve pnömoni grubu sınıflandırması için performans karşılaştırmaları

Algoritmalar	Veri	Doğruluk	%95 G.A.	Kappa	Kesinlik	Duyarlılık	F1
Naif Bayes	Eğitim	0.764	(0.742 - 0.785)	0.528	0.847	0.726	0.782
	Test	0.750	(0.704 - 0.791)	0.500	0.800	0.727	0.761
Lojistik Regresyon	Eğitim	0.723	(0.701 - 0.744)	0.446	0.787	0.697	0.739
	Test	0.722	(0.675 - 0.765)	0.445	0.780	0.699	0.737
YSA	Eğitim	0.756	(0.735 - 0.777)	0.513	0.831	0.723	0.773
	Test	0.690	(0.642 - 0.735)	0.380	0.740	0.672	0.704
SVM (Linear)	Eğitim	0.723	(0.701 - 0.745)	0.447	0.817	0.688	0.747
	Test	0.730	(0.683 - 0.772)	0.460	0.805	0.700	0.748
SVM (Non-Linear)	Eğitim	0.782	(0.761 - 0.802)	0.565	0.790	0.778	0.784
	Test	0.772	(0.728 - 0.812)	0.545	0.745	0.788	0.766
KNN	Eğitim	0.745	(0.723 - 0.766)	0.491	0.851	0.702	0.769
	Test	0.695	(0.649 - 0.742)	0.395	0.785	0.668	0.721
Karar ağaçları	Eğitim	0.796	(0.775 - 0.815)	0.592	0.922	0.736	0.819
	Test	0.805	(0.762 - 0.842)	0.610	0.920	0.748	0.825
Bagging	Eğitim	0.878	(0.861 - 0.891)	0.756	0.876	0.879	0.877
	Test	0.712	(0.665 - 0.756)	0.425	0.690	0.722	0.705
Random Forest (CART)	Eğitim	0.877	(0.861 - 0.893)	0.755	0.976	0.815	0.888
	Test	0.715	(0.668 - 0.758)	0.430	0.700	0.721	0.710
Random Forest (CIT)	Eğitim	0.812	(0.792 - 0.831)	0.625	0.891	0.770	0.826
	Test	0.750	(0.704 - 0.791)	0.500	0.765	0.742	0.753
Gradient Boosting Makinesi	Eğitim	0.810	(0.790 - 0.829)	0.621	0.946	0.744	0.833
	Test	0.827	(0.786 - 0.863)	0.655	0.940	0.767	0.844
C5.0	Eğitim	0.818	(0.799 - 0.837)	0.637	0.851	0.799	0.824
	Test	0.782	(0.738 - 0.822)	0.565	0.780	0.783	0.782

Tablo 4-21'deki sağlıklı grup ve pnömoni verilerinin sonuçları incelendiğinde doğruluk, kappa, kesinlik, duyarlılık ve F1 ölçüm değerlerine göre performans açısından en iyi

eğitim performansı Bagging ve Random Forest (CART) algortimaları göstermiştir. En kötü eğitim performansı ise Lojistik Regresyon ve KNN göstermiştir. Test performansına dayalı algoritmalar incelendiğinde GBM en iyi algoritma olarak bulunmuştur ve buna en yakın Karar Ağaçları görülmektedir. Ayrıca Yapay Sinir Ağları, KNN en zayıf test performansı göstermiştir.

Tablo 4-22 Astım, KOAH, pnömoni grupları ve sağlıklı grup sınıflandırması için performans karşılaştırmaları

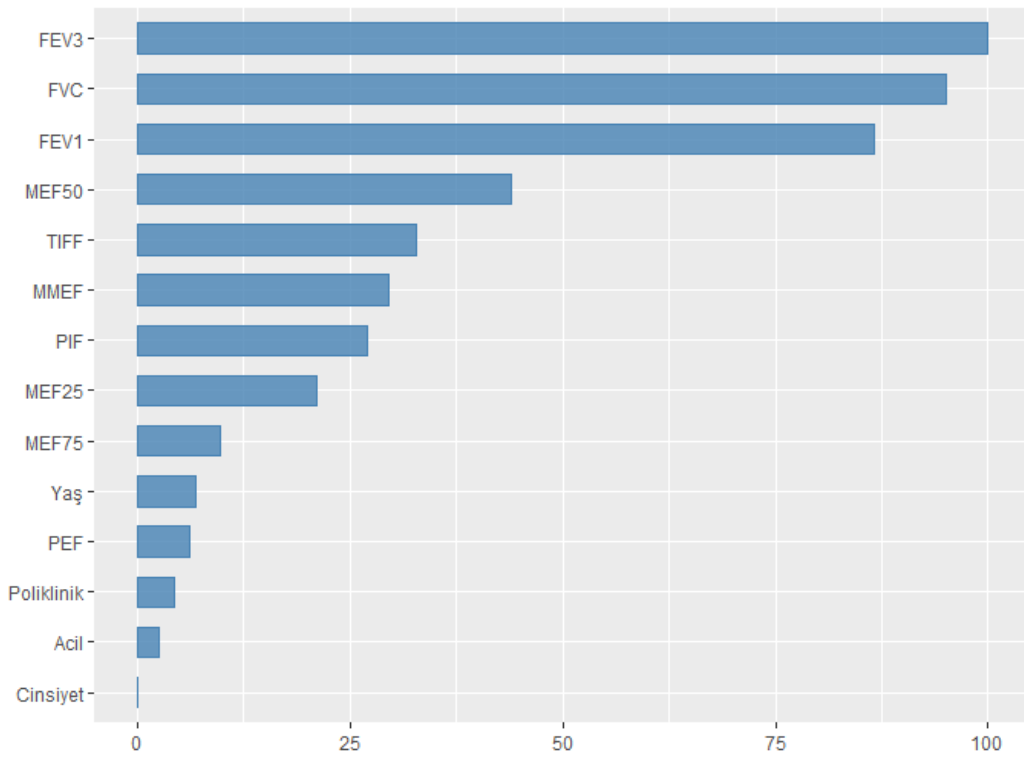
Algoritmalar	Veri	Doğruluk	%95 G.A.	Kappa
Naif Bayes	Eğitim	0.728	(0.712 - 0.743)	0.637
	Test	0.707	(0.674 - 0.738)	0.610
Lojistik Regresyon	Eğitim	0.713	(0.695 - 0.728)	0.448
	Test	0.708	(0.667 - 0.727)	0.443
YSA	Eğitim	0.817	(0.803 - 0.830)	0.756
	Test	0.766	(0.735 - 0.795)	0.688
SVM (Linear)	Eğitim	0.644	(0.627 - 0.661)	0.526
	Test	0.655	(0.620 - 0.687)	0.540
SVM (Non-Linear)	Eğitim	0.741	(0.726 - 0.757)	0.655
	Test	0.712	(0.679 - 0.743)	0.616
KNN	Eğitim	0.798	(0.784 - 0.812)	0.731
	Test	0.756	(0.725 - 0.785)	0.675
Karar ağaçları	Eğitim	0.877	(0.865 - 0.888)	0.836
	Test	0.770	(0.739 - 0.798)	0.693
Bagging	Eğitim	0.939	(0.932 - 0.947)	0.918
	Test	0.788	(0.758 - 0.816)	0.718
Random Forest (CART)	Eğitim	0.939	(0.930 - 0.947)	0.919
	Test	0.801	(0.771 - 0.828)	0.735
Random Forest (CIT)	Eğitim	0.908	(0.897 - 0.917)	0.877
	Test	0.803	(0.774 - 0.830)	0.738
Gradient Boosting Makinesi	Eğitim	0.878	(0.838 - 0.899)	0.829
	Test	0.831	(0.809 - 0.867)	0.697
C5.0	Eğitim	0.898	(0.887 - 0.904)	0.864
	Test	0.845	(0.818 - 0.869)	0.793

Tablo 4-22’de dört grubun sınıflandırma sonuçlarına göre doğruluk ve Kappa değerleri ele alındığında en iyi eğitim performansı Rasgele Orman, Bagging, C5.0, GBM ve Karar Ağaçları bulunmuştur. En zayıf eğitim performansını ise Destek Vektör Makinesi (doğrusal) göstermiştir. Test performansına gelindiğinde en yüksek değerler C5.0 ve

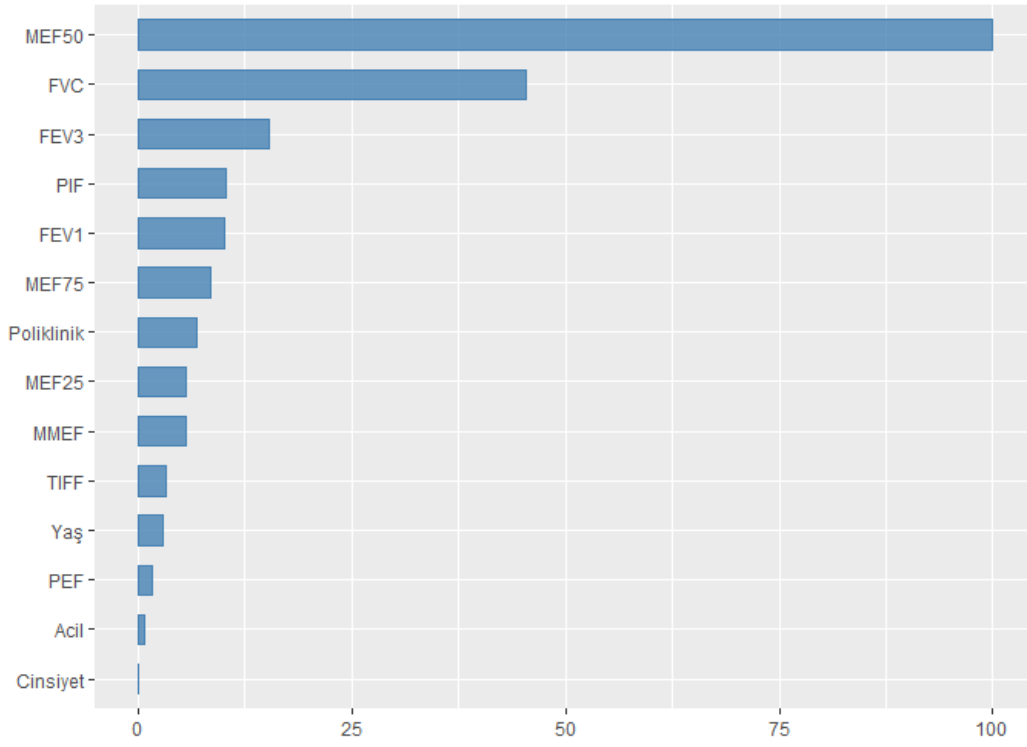
GBM’de olduğu görülmektedir. En zayıf yine Destek Vektör Makinesi(doğrusal) bulunmuştur.

4.5. Astım, KOAH, pnömoni grupları ve sağlıklı grup sınıflandırmasında değişken önemlilikleri

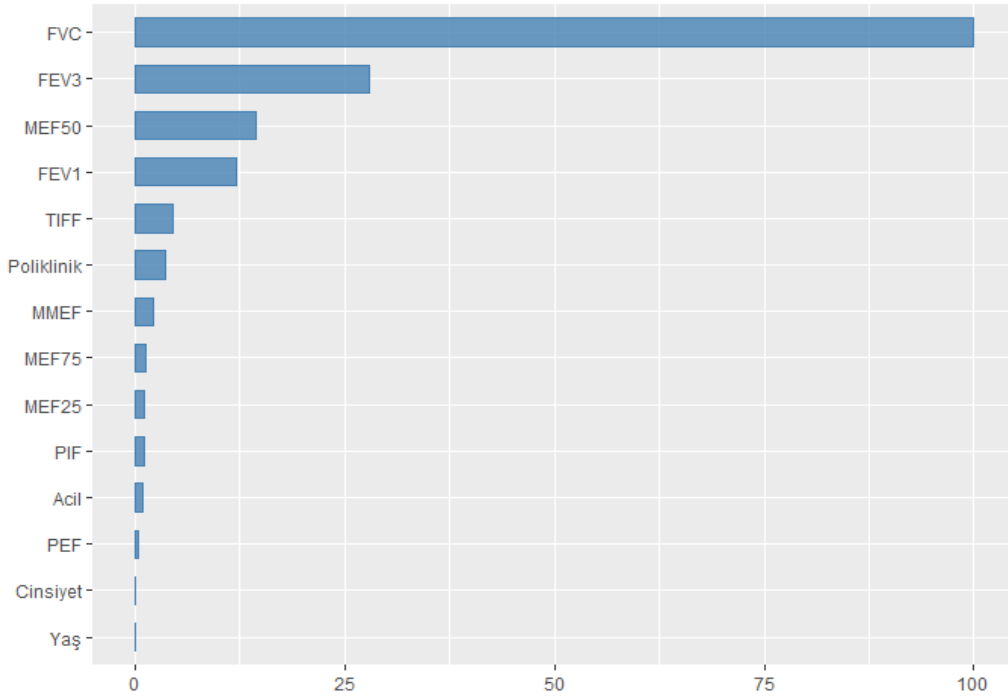
Grupların sınıflandırılmasında değişkenlerin önemlilikleri grafikler halinde gösterilmiştir. Astım ve sağlıklı grup için Şekil 4-1, KOAH ve sağlıklı grup için Şekil 4-2, pnömoni ve sağlıklı grup için Şekil 4-3 ve tüm grupların sınıflandırması için Şekil 4-4 değişken önemlilik sıralamasını göstermektedir.



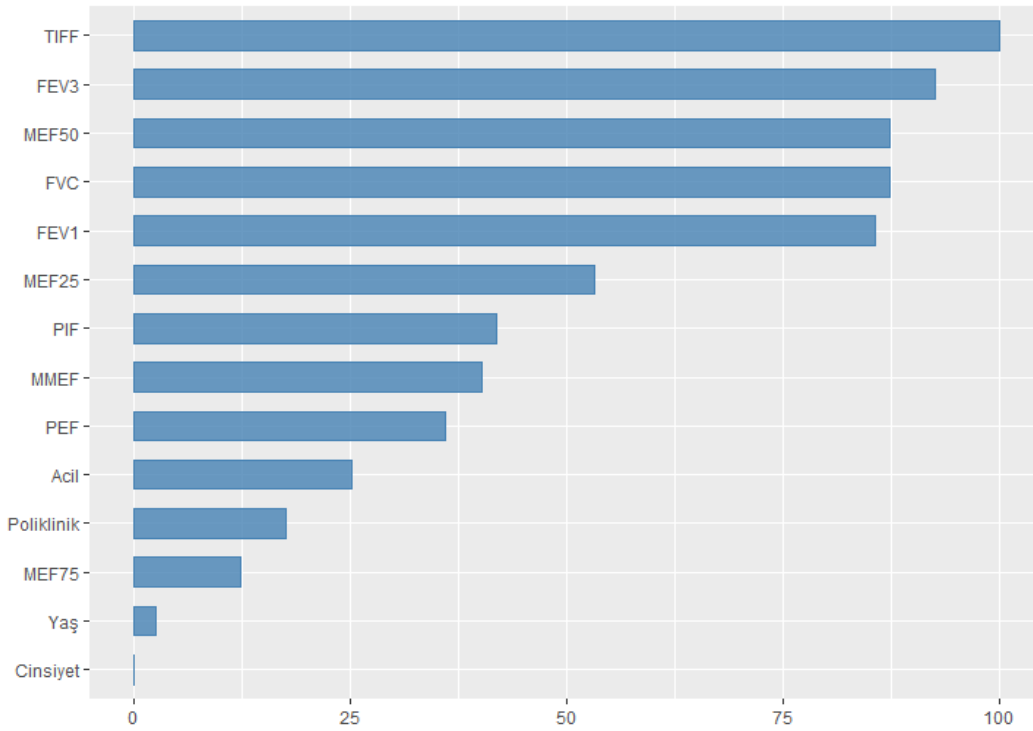
Şekil 4-1 Sağlıklı grup ve astım grubu sınıflandırması için Rastgele Orman algoritmasına dayalı değişkenlerin önemi



Şekil 4-2 Sağlıklı grup ve KOAH grubu sınıflandırması için GBM algoritmasına dayalı değişkenlerin önemi



Şekil 4-3 Sağlıklı grup ve pnömoni grubu sınıflandırması için Rastgele Orman algoritmasına dayalı değişkenlerin önemi



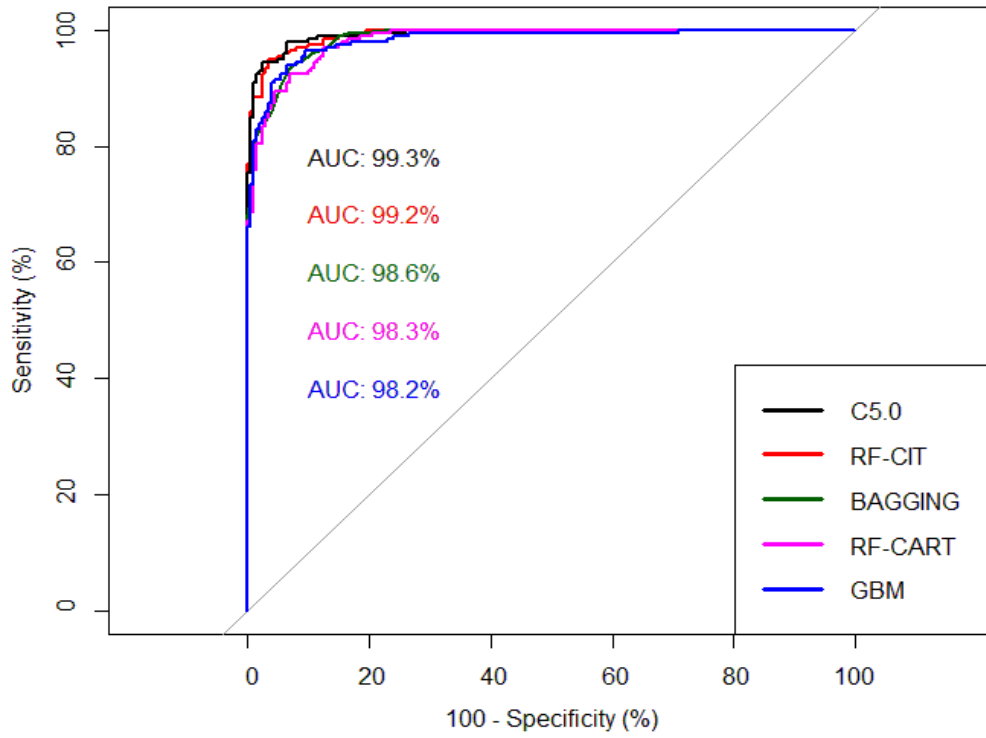
Şekil 4-4 Astım, KOAH, pnömoni grupları ve sağlıklı grup sınıflandırması için C5.0 algoritmasına dayalı değişkenlerin önemi

Rastgele Orman (CART) ve C5.0 algoritmaları astım vakası için test performansı açısından en iyi algoritmalar olarak bulunurken, Destek Vektör Makinesi (doğrusal olmayan) KOAH vakası için ve GBM pnömoni vakası için en iyi sonuçlar vermiştir. Ayrıca, tüm grupların sınıflandırılmasında C5.0 en iyi performansı göstermiştir. Astım, KOAH ve pnömoni hastalıklarının teşhisinde en anlamlı faktörlerin (değişkenler) belirlenmesi için değişken önem derecesinin değerlendirilmesi yapılmıştır. Astım ve pnömoni durumunda, Random Forest (CART) ve GBM algoritmasının değişkenlerin önemi üzerine kendi iç ölçüleri vardır. Ancak KOAH olgularında Destek Vektör Makinesi için benzer hesaplamalar yapılamaz. Her durumda da karşılaştırmalı grafik sunarak değişken önemini değerlendirmek için Destek Vektör Makinesi (doğrusal olmayan)'den biraz daha düşük performansa sahip GBM algoritması dikkate alınmıştır. Astım vakası sınıflandırma sonuçlarına göre, en önemli değişkenler FEV3, FVC, FEV1, MEF50 olarak bulunmuştur. En az önemli değişkenler PEF, acil durum ve cinsiyettir. KOAH hastalığının teşhisinde Şekil 4-2'de görüldüğü gibi MEF50, FVC, FEV3, PIF ve FEV1, en önemli değişkenler olarak bulunmuştur. PEF, acil durum ve cinsiyet en az öneme sahip değişkenlerdir. Pnömoni teşhisinde ise özellikle FVC ön planda olup, FEV3, MEF50 ve FEV1 dikkate değer değişkenlerdir. Diğer değişkenler ise önemli sayılacak

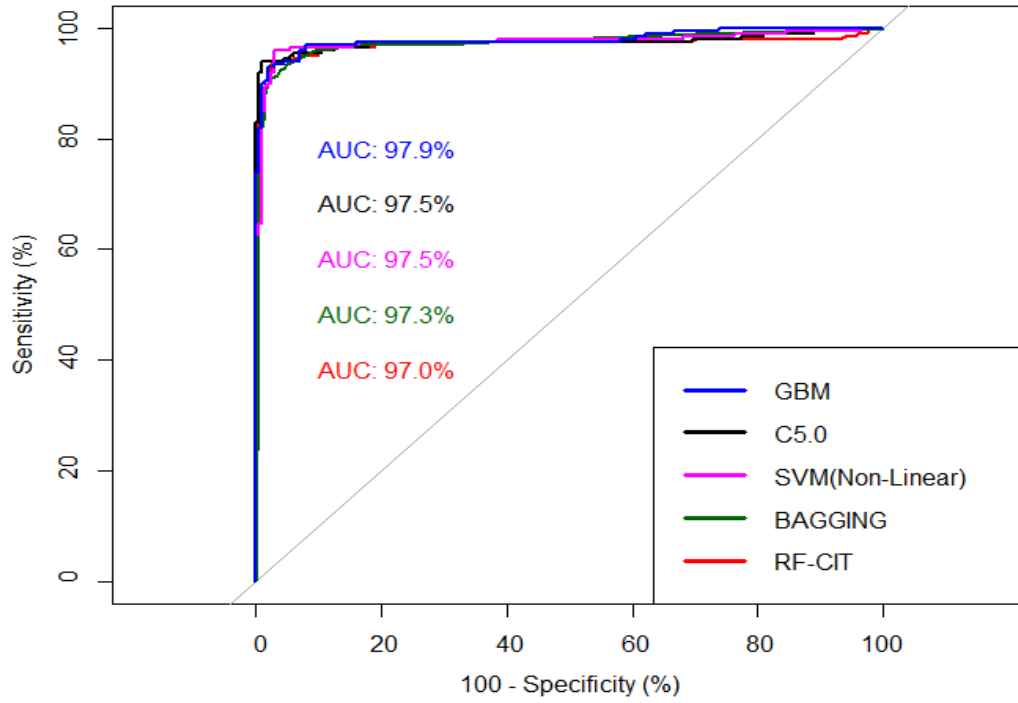
düzeyde değildir. Tüm grupların sınıflandırmasında değişken önemliliğine değinildiğinde ise neredeyse tıbbi değişkenlerin hepsi büyük öneme sahiptir. Tıbbi ölçümler dışında poliklinikte KOAH için astımdan daha önemli olduğu saptanmıştır. Rastgele Ormanlarda, değişken önemi bina ağacı için kullanılmayan veri seti olan torba dışı veri bölümü kullanılarak tahmin doğruluğuna göre hesaplanır. Her değişkenin karıştırılmış değerleri kullanılarak elde edilen doğruluk, karıştırmadan önce temel doğruluk değeri ile karşılaştırılır. Doğruluk arasındaki farkların ortalaması, standart hata ile ortalanan ve normalleştirilir. En yüksek değer, daha önemli değişkenler anlamına gelir. GBM'de, her değişkenin önem seviyesi, hatadaki iyileşme, diğer bir deyişle sınıflandırma durumu için yanlış sınıflandırma oranındaki azalma miktarı dikkate alınarak hesaplanmaktadır. Bir değişken bölmek için kullanıldığında, hatadaki değişiklik hesaplanır ve diğer değişken performanslarıyla karşılaştırılır. Son olarak, her değişken için göreceli önem seviyesi, maksimum değer yüzde olacak şekilde ölçeklendirilerek elde edilir.

4.6. ROC Eğrileri

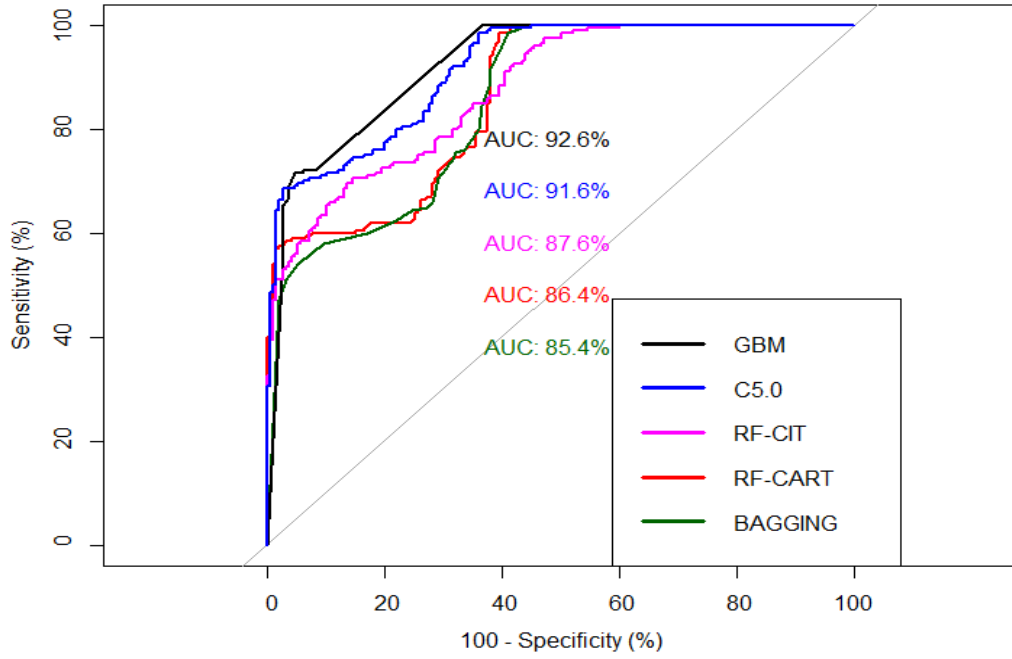
Daha önce kullanılan performans kriterlerine ek olarak, sınıflandırma durumlarında ROC (alıcı işlem karakteristiği) eğrisi ve AUC (eğri altındaki alan) değeri kullanılabilir. Bu kriterler, sınıf değerlerini ayırt edici algoritma yeteneğini ölçmek için kullanılır. ROC eğrisi görsel bir sunumdur ve her cut-off değeri için duyarlılık ve özgüllük arasındaki değişimi gözlemlememizi sağlar. Daha yüksek AUC değeri, algoritmanın verilere daha iyi uyması anlamına gelir. Test performansına dayanan ilk beş algoritma için ROC eğrisi ve AUC değerleri sırasıyla Şekil 4-5, Şekil 4-6 Şekil ve Şekil 4-7'de astım, KOAH ve pnömoni hastalıkları için verilmiştir.



Şekil 4-5 Sağlıklı grup ve astım grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri



Şekil 4-6 Sağlıklı grup ve KOAH grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri



Şekil 4-7 Sağlıklı grup ve pnömoni grubu için ilk beş algoritmanın ROC eğrileri ve AUC değerleri

Şekil 4-5, 4-6, 4-7 'deki sonuçlara göre, tüm algoritmaların ROC eğrileri ve AUC değerleri benzerdir ve oldukça başarılı uyumu temsil etmiştir. Aralarında hafif performans farklılıkları vardır. Astım vakası için en iyi algoritma olarak bulunan rastgele orman (CART) ve C5.0, KOAH vakası için en iyi bulunan SVM (doğrusal olmayan) ve GBM olup pnömoni vakası için GBM ve C5.0 ROC eğrisi ve AUC değeri incelendiğinde yine oldukça yüksek değerlerde görülmektedir.

5. TARTIŞMA

Bu tezde kapsamlı bir makine öğrenmesi çalışması yapılmıştır. Bilinen ve oldukça etkili makine öğrenme algoritmaları, astım, KOAH, pnömoni hastalık gruplarını ve sağlıklı grubu büyük ve ayrı veri setlerinde k-kat çapraz doğrulama ve hiperparametre tarama uzayı ile teşhis etmek veya sınıflandırmak için göz önünde bulundurulmuştur. Doğruluk, kappa istatistiği, duyarlılık, kesinlik, F1 ölçümü ve AUC değerleri dahil olmak üzere çeşitli performans kriterleri analitik olarak değerlendirilmiş ve ROC eğrisi görsel olarak incelenmiştir. Test performans sonuçlarına göre, astım vakası için Random Forest (CART) ve C5.0, KOAH vakası için DVM (doğrusal olmayan) ve GBM, pnömoni vakası için Bagging ve Random Forest (CART) ayrıca tüm grupların birlikte sınıflandırılması için C5.0 ve GBM algoritması en iyi algoritmalar olarak bulunmuştur.

Değişken önemlilikleri dikkate alındığında sağlıklı grup ile hastalıklar arasında astım vakası sınıflandırma sonuçlarına göre, en önemli değişkenler FEV3, FVC, FEV1, MEF50'dır. KOAH'da MEF50, FVC, FEV3, PIF ve FEV1, pnömonide ise FVC, FEV3, MEF50 ve FEV1 dikkate değer değişkenlerdir. Tüm grupların birlikte sınıflandırmasında ise neredeyse tıbbi değişkenlerin hepsi büyük öneme sahiptir. Benzer şekilde, her grubun sınıflandırması için FVC ve FEV1'in önem seviyelerine dikkat edilmelidir.

Acil başvuru sayısı, poliklinik başvuru sayısı, cinsiyet, yaş genel olarak tüm sınıflandırmalarda fazla önem arz etmemektedir.

Değişkenlerin önem sırası, korelasyon katsayısı ve istatistiksel anlamlılık testleriyle desteklenmiştir.

Son zamanlarda yapılan bir araştırmada, KOAH'ın durumunda, Random Forest'in sınırlı bir çalışma popülasyonunda klinik geçmişine göre yüzde 97,7 duyarlılıkla diğer teknikleri geride bıraktığını göstermektedir (Spathis ve Valomos, 2017). Diğer yandan; verilerimizi hastanın sigara içme durumu gibi tarihsel değişkenleriyle analiz etsek daha iyi sonuçlar elde edebilirdik.

Diğer bir çalışmada astım, KOAH ve sağlıklı arasında ayrım yapabilen bir uzman tanı sistemi geliştirilmiştir (Badnjevic ve ark., 2018). Bu çalışmada, Spathis ve Valamos'un yöntemlerine ek olarak Gradyan Boosting algoritmasını kullanmışlardır. Gradyan Boosting, 98.33 ile en yüksek doğruluk yüzdesine sahip bulunmuştur. Özellikle bu çalışmaların ve diğer çalışmaların sonuçlarının çalışmamızdaki sonuçlara benzerlik gösterdiği anlaşılmaktadır. Aynı tür algoritmalarla benzer sonuçlar elde edilmiştir.

Başka bir çalışmada, ilk defa, uyarlanabilir destek vektör makinesi algoritması astım ve KOAH'ı teşhis etmek için incelenmiştir (Yahyaoui ve Yumuşak, 2018). Sınıflama doğruluğu astım için % 98.45 ve KOAH için % 92.63 bulunmuştur. Bu çalışmadaki yüzdelik oranları çalışmamızda incelediğimiz destek vektör makinesi oranlarıyla örtüşmektedir.

Ayrıca, çeşitli yapay sinir ağ mimarileri ve bulanık mantık sınıflandırıcılarının yanı sıra her iki metoda dayanan uzman sistem yapıları kullanılarak solunum yolu hastalıklarının teşhisi için çeşitli çalışmalar yapılmıştır. Bu çalışmalar, çeşitli özelliklere sahip farklı veri setleri kullanılarak, astım ve KOAH da dahil olmak üzere çeşitli solunum yolu hastalıklarının teşhisi için farklı sinir ağı yapıları uygulamıştır (Price ve ark, 2001; Hanania ve ark, 2010; Badnjevic ve ark, 2018; Granulo ve ark., 2016; Gurbeta ve ark, 2018). Bu çalışmalarda doğruluk oranları yeterli derecede yüksektir.

Yapay zeka ve makine öğrenmesinin gelecekte hastalıkların tahmininde kullanımının artacağına ve sağlık alanında birçok noktada yapay zeka yöntemlerinden yararlanılacağına dair öngörüler bulunmaktadır.

Makine öğrenmesi yöntemlerinde hastalıkların sınıflandırılması ve tahmin edilmesinde başarılı tahminlerde bulunmasının farklı unsurları bulunmaktadır. Veri tabanlarında çok sayıda hastanın verilerinin bulunması, bu hastalıkların makine öğrenme yöntemleri ile sınıflandırılmasına olanak tanımaktadır (Doğaner, 2020).

Toplumda prevalansı düşük olan hastalık türlerine ilişkin makine öğrenme yöntemi kullanılarak yapılan sınıflandırma ve tahmin araştırmalarının az sayıda olduğu gözlemlenmiştir. Bu durum, yeterli hasta sayısının olmaması ve veriye ulaşamamaktan kaynaklanmaktadır (Doğaner, 2020).

Makine öğrenme yöntemleri az sayıda örnekte, eğitim veri setindeki örüntüyü yeterince öğrenememesinden dolayı, düşük sınıflandırma performansı gösterebilmektedir.

Astım, KOAH, zatürre gibi göğüs hastalıkları solunum yollarını tehdit eden hastalıklar arasındadır. Günümüzde bir çok araştırmacı, bu tür hastalıkları saptamak ve teşhis etmek için kullanılabilecek etkili yöntemleri araştırmaktadır. Yapılan çalışmalar ve araştırmalar doktorların daha kolay ve doğru teşhis koymalarını yardımcı olabilecek etkili karar destek sistemleri üzerine yoğunlaşmaktadır. Bu kapsamda, uzman sistemler ve farklı yapay zeka teknikleri tıp alanı dahil olmak üzere farklı alanlar üzerindeki etkinliklerini ve doğruluklarını kanıtlamaya devam etmektedir (Amani, 2017).

Topluluk öğrenme yöntemlerinin geliştirilmesi ve daha yüksek performans ile sınıflandırma gerçekleştirebilecek yöntemler üzerine yapılacak araştırmalarda, farklı sınıflayıcıları modele dahil eden, birleştirme tekniğinde sınıflayıcıların entropilerini dikkate alan ve ağırlıklı birleştirme tekniğini esas alan, modelde karar ağacı tabanlı sınıflayıcılara yer veren yeni yöntemler üzerinde çalışılması önerilmektedir (Doğaner, 2020).

Çalışmada astım, KOAH ve pnömoni hastalıklarının teşhisinde ve sınıflandırılmasında değişkenlerin sayısı bir sınırlama olarak görülmektedir. Gelecekte, daha fazla demografik özellik, sağlık alışkanlığı ve tıbbi ölçüm gibi her hastalık grubu için de etkili olabilecek yeni değişkenler daha fazla örneklem ve farklı makine öğrenmesi yöntemleri eklenebilir. Sonuç olarak, bu çalışma sağlık bilimleri ve makine öğrenimi uygulamaları konusunda uzman sistemler üzerinde çalışan uygulayıcılar ve araştırmacılar için değerli olacaktır. Bu çalışmadaki tüm sonuçların, sağlık uzmanlarına, araştırmacılara ve uzman sistem geliştiricilere, belirli bir durum için daha iyi bir görüş sağlamasına yardımcı olacağı ümit edilmektedir.

KAYNAKLAR

- Afonso, A.SM., Verhamme, K.MC., Sturkenboom, M. CJM., Brusselle. G.G.O. (2011). COPD in the general population: prevalence, incidence and survival, *Respiratory medicine*, 105(12): 1872-1884.
- Akar, Ö. ve Güngör, G. (2012). Rastgele orman algoritması kullanarak çok bantlı görüntülerin sınıflandırılması, *Jeodezi ve Jeoinformasyon Dergisi*, 139-146.
- Alpaydın, E. (2004). *Introduction to machine learning*, MIT press.
- Alwan, A. (2010). *Global Status Report on Non-Communicable Diseases*, WHO.
- Amani, Y. (2017). Göğüs hastalıklarının teşhis edilmesinde makine öğrenmesi algoritmalarının kullanılması, *Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Doktora tezi*.
- Amaral, J.L.M., Lopez, J.A., Jansen, J.M., Faria, A.C.D, Melo, P. L.(2011). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease, *computer methods and programs in biomedicine*.
- Amasyalı, M.F. (2008). Yeni makine öğrenmesi metotları ve ilaç tasarımına uygulamalar, *YTÜ, Fen Bilimleri Enstitüsü*.
- Archer, K.J. ve Kimes, R.V. (2008). Empirical characterization of random forest variable importance measures, *Computational Statistics & Data analysis* 52.4:2249:2260.
- Asaithambi, M., Manoharan, S., C., Subramanian, S. (2012). Classification of Respiratory Abnormalities Using Adaptive Neuro Fuzzy Inference System. *Int. Inf. and Database SysLecture Notes in Computer Science*. 7198, 65–73.
- Badnjevic, A., Gurbeta, L., Cifrek, M., & Marjanovic, D. (2016). Classification of Asthma Using Artificial Neural Network, *IEEE 39th International convention on information and communication technology, electronics and microelectronics (MIPRO)*.
- Badnjevic, A., Gurbeta, L., Custovic E. (2018). An Expert Diagnostic System to Automatically Identify Asthma and Chronic Obstructive Pulmonary Disease in Clinical Settings, *Scientific Reports*.
- Bilgin, M. (2018). *Makine Öğrenmesi Teorisi ve Algoritmaları*, Papatya Yayıncılık.
- Breiman, L. (1996). “Bagging Predictors” *Machine Learning*, 24(2) 123–140.
- Breiman, L. (1998). Arcing Classifiers, *The Annals of Statistics*, 26, 123–140.

- Breiman, L. (2001). Random Forests, Machine learning,5-32.
- Carbonell, J.G., Ryszard S.M. ve Tom M. M.(1983). Machine learning: A historical and methodological analysis.
- Chao, W. L .(2011). Machine learning Tutorial.
- Çakır, M. (2005). Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makine Öğrenmesi Teknikleri: Ampirik Uygulamalar ve Karşılaştırmalı Analiz, Uzmanlık Tezi, TC Merkez Bankası.
- Çentik, G.(2013). Makine Öğrenmesi Yöntemlerinin Polisomnografik Verilere Uyarlanması, Tez, Trakya Üniversitesi Fen Bilimleri Enstitüsü.
- Djeffal, A. (1999).Utilisation de methode support vector machines dans l'analyse des bases de donnees. Mohamed Khider-Biskra University, Biskra, Algeria, PhD thesis.
- Doğan, S. (2006).Türkçe Dökümanlar için N-gram tabanlı sınıflandırma: Yazar, Tür ve Cinsiyet, Tez, YTÜ, Fen Bilimleri Enstitüsü.
- Doğaner, A.(2020). Topluluk Öğrenme Yöntemleri ile Renal Hücreli Karsinom'un Tahmin Edilmesi, Tez, İnönü Üniversitesi Sağlık Bilimleri Enstitüsü.
- Eisner, M.D., et al. (2011). An official American Thoracic Society public policy statement: novel risk factors and the global burden of chronic obstructive pulmonary disease. American Journal of Respiratory and Critical Care Medicine, 182: 693–718.
- Fawcett, T. (2006). An introduction to ROC analysis, Pattern recognition letters:861-874.
- Finkelstein, J., Jeong, I. (2018). Machine learning approaches to personalize early prediction of asthma exacerbations., Ann N Y Acad Sci, 1387(1): 153–165.
- Freund, Y. (1996). Schapire R., Experiments with a New Boosting Algorithm, Machine Learning: Proceedings of the Thirteenth International Conference, pp. 148–156.
- Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics,29(5), 1189–1232.
- Friedman, J. (2002). Stochastic Gradient Boosting, Computational Statistics and Data Analysis, 38(4), 367–378.
- Global Initiative for Asthma (GINA): Global strategy for asthma management and prevention. Update 2018 and Online Appendix. Available at <http://www.ginasthma.org>. (Accessed 18 Nov 2018)
- Granulo,E., Becar,L., Gurbeta,L., Badnjevic,A.(2016).Telemetry System for Diagnosis

- of Asthma and Chronical Obstructive Pulmonary Disease (COPD), Internet of Things Technologies for HealthCare. Springer, International Publishing, pp. 113–118.
- Gurbeta, L., Badnjevic, A., Maksimovic, M., Omanovic-Miklicanin, E., Sejdic, E. (2018). A telehealth system for automated diagnosis of asthma and chronical obstructive pulmonary disease, Journal of the American Medical Informatics Association.
- Guyton, C., Hall, J. (2010). Textbook of Medical Physiology, Saunders Elsevier
- Gürsakaç, N. (2018). Makine Öğrenmesi, Dora Yayınevi
- Haider, N.S., Singh, B. K., Periyasamy, R., Behera, A.K. (2019). Respiratory Sound Based Classification of Chronic Obstructive Pulmonary Disease: a Risk Stratification Approach in Machine Learning Paradigm, Journal of Medical Systems.
- Han, J., ve Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufman, San Francisco
- Hanania, N. A., Mannino, D. M & Yawn, B. P. (2010). Predicting risk of airflow obstruction in primary care: validation of the lung function questionnaire (LFQ). Respiratory Medicine. 104, 1160–1170.
- Harrington, P. (2012). Machine learning in action, CT: Manning.
- James, G., Witten D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning, Vol. 112 New York: Springer.
- Kabacoff, R. (2015). R in action: Data analysis and graphics with R.
- Kaushik, B.K., Nagamalai, D. (2011). Advances in computer science and information technology. Berlin: Springer, pp. 570–576.
- Khalid, M.S., Haque, M.A., Hossain, M.S., et al. A belief rule-based (BRB) decision support system for assessing clinical asthma suspicion. In: Proceedings of Scandinavian conference on health informatics, Grimstad, August (2014), pp. 83–89. Linköping: Linköping University Electronic Press.
- Küçükbay, S.R., (2015). Metaloproteinlerin Biyoenformatik Analizi, Tez, Başkent Üniversitesi Fen Bilimleri Enstitüsü.
- Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling (Vol. 26), New York: Springer.
- Lantz, B. (2015). Machine learning with R. 2nd ed. Birmingham: Packt Publishing.
- Larose, D.T. (2005). Discovering Knowledge in Data: An introduction Mining John Wiley & Sons., New Jersey.
- Masoli, M., Fabian, D., Holt, S., Beasley, R. (2004). Global initiative for asthma (GINA)

- Program. The global burden of asthma: executive summary of the GINA dissemination committee report. *Allergy*, 59:469–78.
- Medhekar, D.S., Bote, M. P., Deshmukh, S. D. (2013). Heart disease prediction system using naive bayes. *International journal of enhanced research in science technology & engineering.*, 2(3).
- Michalski, R.S., Carbonell, J.G. ve Mitchell, T.M. (1986). *Machine learning: Artificial intelligence approach*, Morgan Kaufmann, Springer.
- Mohktar M.S., Redmond, S.J., Antoniadis, N.C., et al. (2015). Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. *Artif Intell Med*; 63(1): 51–59.
- Musen, M.A., Middleton, B., and Greenes, R.A. (2014). Clinical decision-support systems In: Shortliffe EH and Cimino JJ (eds) *Biomedical informatics*. London: Springer, pp. 643–674.
- Nilsson, N. (1996). *Introduction to machine learning*, An early draft of a proposed textbook.
- Öğücü, M.N. (2006). *Yapay Sinir Ağları ile Sistem Tanıma*, Tez, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.
- Özgür, A. (2004). *Supervised and unsupervised machine learning techniques for text document categorization*, Tez, Boğaziçi Üniversitesi Fen bilimleri Enstitüsü.
- Öztemel, E., *Yapay Sinir Ağları*, Papatya Bilim Yayınevi,
- Pöyhönen, S. (2004). *Support vector machine based classification in condition monitoring of induction motors*. Helsinki University of Technology.
- Prasad, BDCN., Prasad, PK. and. Sagar, Y.A. (2011). A comparative study of machine learning algorithms as expert systems in medical diagnosis (Asthma). In: Meghanathan N, Kaushik BK and Nagamalai D (eds) *Advances in computer science and information technology*. Berlin: Springer, pp. 570–576.
- Prosperi, M.C., Marinho, S., Simpson, A., et al. (2014). Predicting phenotypes of asthma and eczema with machine learning. *BMC Med Genom*; 7(Suppl. 1): S7.
- Price, D. B., et al. (2006). Symptom-based questionnaire for identifying COPD in smokers. *Respiratio*. 73, 285–294.
- Quinlan, J. R., *Learning with continuous classes*. In 5th Australian joint conference on artificial intelligence, November 1992, Vol. 92, pp. 343-348.
- Samuel, A.L. (1999). *Some studies in machine learning using the game of checkers*, IBM

- Journal of research and development: 206-226.
- Sebastiani, F. (2002). Machine learning in automated text categorization, ACM computing surveys (CSUR), 1-47.
- Spathis, D., Vlamos, P. (2017). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics Journal.
- Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N.G., Bazaz, G, Kappel, G., Gerber, A.N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease, Plos One, 12(11).
- Witten, I.H., Frank, E., Hall, M.A. ve Pal, C.J. (2016). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.
- World Health Organization. Global surveillance, prevention and control of chronic respiratory diseases: a comprehensive approach. (2007).
- Yahyaoui, A., Yumusak, N. (2018). Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems., Biomedical Research.

İNTİHAL RAPORU İLK SAYFASI

SOLUNUM SİSTEMİ HASTALIKLARININ SINIFLANDIRILMASINDA MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KULLANIMI

ORIJINALLIK RAPORU

% 3	%	% 3	%
BENZERLİK ENDEKSİ	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCİ ÖDEVLERİ

BİRİNCİL KAYNAKLAR

- 1** Alvaro Camilo Dias Faria, Karla Kristine Dames da Silva, Gerusa Martmo da Costa, Agnaldo Jos, Pedro Lopes de Melo. "Chapter 13 Forced Oscillation Technique in the Detection of Smoking-Induced Respiratory Changes", IntechOpen, 2012

Yayın

%**1**
- 2** Hasan Tetiker, Ceren Uğuz Gençer. "3 Boyutlu Bilgisayarlı Tomografi ile Volüm Rendering Tekniği Kullanarak Skapula Ölçümlerinden Anadolu Popülasyonunda Cinsiyet Tahmini", The Bulletin of Legal Medicine, 2020

Yayın

<%**1**
- 3** YILMAZ, Pakize Özçiftçi and GÜLEN, Şule Taş. "KOAHA'da Nutrisyonel Durumun Değerlendirmesinde Mini ", Kocaeli Derince Eğitim ve Araştırma Hastanesi, 2017.

Yayın

<%**1**