T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE MEME KANSERİNİN ERKEN TEŞHİSİ

MELİHA NUR DURAK

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANABİLİM DALI

DANIŞMAN
DOÇ. DR. İBRAHİM DEMİR

İSTANBUL, 2017

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE MEME KANSERİNİN ERKEN TEŞHİSİ

Meliha Nur DURAK tarafından hazırlanan tez çalışması 12/10/2017 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri İstatistik Anabilim Dalı'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Doç. Dr. İbrahim DEMİR

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Doç. Dr. İbrahim DEMİR

Yıldız Teknik Üniversitesi

Doç. Dr. Ersoy ÖZ

Yıldız Teknik Üniversitesi

Doç. Dr. Barış AŞIKGİL

Mimar Sinan Güzel Sanatlar Üniversitesi

ÖNSÖZ

Tez çalışmamın başından sonuna kadar her aşamasında deneyimlerini ve bilgisini benden esirgemeyen doğru yönlendirmeler neticesinde oluşturduğum tezimde öğrencisi olmaktan onur duyduğum danışman hocam Doç. Dr. İbrahim DEMİR'e sonsuz teşekkürlerimi bir borç bilirim.

Hayatımın her alanında olduğu gibi, tez çalışmalarım sırasında da, destekleri, anlayışları ve sevgileriyle her an yanımda bulunan, destek veren aileme, kardeşim Beyza'ya, dostum Elif'e ve bana verdiği bilgilerle, her türlü yardımı ile, bu tezi yazmamda büyük katkısı olan canım Gülbüke'ye teşekkür ederim.

Mayıs, 2017

Meliha Nur DURAK

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

p_i	i'nin olasılıkları
e_i	Hata terimi
x_i	Öğrenecek sınıfın üyesi
y_i	Eğitim örneği
Ø	Çekirdek işlevleri
f(x)	Optimizasyon fonksiyonu
b	Sapma değeri
d	İki değişken arasındaki uzaklık
$\hat{\mathbf{p}}(x)$	Olasılık yoğunluk tahmincisi
k	En yakın komşu değeri
s	Orijinal yığının entropisi
Ci	Her sınıfın sonraki olasılığı
Wi	Ağırlık

KISALTMA LİSTESİ

WBCD	Meme Kanseri Veri Seti
RBF	Radyal Taban Ağ
GRNN	Genel Regresyon Sinir Ağı
PNN	Olasılıksal Sinir Ağı
MLP	Çok Katmanlı Algılama
NB	Naive Bayes
SMO	Sıralı Minimum Optimizasyon
k-NN	k-En Yakın Komşu
DVM	Destek Vektör Makineleri
YSA	Yapay Sinir Ağları
LSSVM	En Küçük Kareler Destek Vektör Makinesi
GELM	Grafik Düzenlenmiş Aşırı Öğrenme Makinesi
ROC	İşlem Karakteristik Eğrisi
AUC	İşlem Karakteristik Eğrisi Altında Kalan Alan
DSÖ	Dünya Sağlık Örgütü
IARC	Uluslararası Kanser Ajans

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE MEME KANSERİNİN ERKEN TEŞHİSİ

Meliha Nur DURAK

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç. Dr. İbrahim Demir

Tıp alanında veri madenciliği yaklaşımı, gereksiz yöntemlerden kaçınarak hastalık teşhisinde daha doğru tahmin yapması, tıp uygulayıcılarına daha hızlı karar vermelerinde oldukça yardım sağlaması bakımından son yıllarda yaygın olarak kullanılmaktadır. Hastalığın teşhisinin kolaylaşması, daha doğru tahmin yapılması, gereksiz biyopsiden kaçınmak hasta sonuçlarını iyileştirmenin yanı sıra kullanılan maliyeti düşürür ve klinik çalışmaların artmasına olanak sağlar.

Akciğer kanserinden sonra kadınlarda en yüksek ölüm oranı meme kanseridir. Bu çalışmada amaç gereksiz biyopsiden kaçınarak meme kanserinin erken teşhisini makine öğrenmesi yöntemleriyle araştırmaktır. Önceki çalışmalarda teşhis için tümör bilgilerini içeren değişkenler kullanılırken yapılan bu çalışmada daha çok kültürel ve fiziksel etkisi olan değişkenlerle çalışıldı. Erken teşhiste bu değişkenlerin ne kadar önemli olup olmadığı araştırıldığı gibi farklı değişkenlerle yapılan çalışmalarla da karşılaştırma imkânı sağlandı.

Meme kanseri teşhisi için makine öğrenmesi sınıflandırma yöntemleri kullanılmıştır. Her yöntem kendi içinde yöntemin uygunluğuna göre değişkenlerde boyut azaltması yapmıştır. Her yönteme göre en etkili değişken farklılık göstermiştir.

Aynı veri seti üzerinde iki farklı uygulama yapılmıştır. İlki veri setinin eğitim ve test seti olarak ayrılarak SPSS.18 Moduler programında yapılmış uygulamadır. İkincisi ise k-kat çapraz doğrulama ile Weka'da yapılan uygulamadır.

Kullanılan makine öğrenmesi yöntemlerinden Lojistik Regresyon (%78) ve Naive Bayes (%78) en iyi sonucu vermiştir. Bunlardan sonra sırasıyla Destek Vektör Makineleri (%76), Karar Ağacı C5.0 algoritması (%76), ve Yapay Sinir Ağları (%74) modelleri gelmektedir.

3 kat çapraz doğrulama yapıldığı durumda en iyi sınıflandırma doğrulununu veren yöntemler sırasıyla, Karar Ağacı C5.0 algoritması (%76), Lojistik Regresyon (%73), Destek Vektör Makineleri (%73), Yapay Sinir Ağları (%73) ve Naive Bayes (%73) olarak bulunmuştur.

Anahtar Kelimeler: Makine öğrenmesi, lojistik regresyon, destek dektör makineleri, karar ağaçları, yapay zeka

ABSTRACT

EARLY DIAGNOSIS OF BREAST CANCER WITH MACHINE LEARNING CLASSIFICATION METHODS

Meliha Nur DURAK

Department of Statistics

MSc. Thesis

Adviser: Assoc.Prof. İbrahim Demir

In the medical fields, the data mining approach has been widely used in recent years in order to make more accurate prediction of disease diagnosis by avoiding unnecessary methods and to help medical practitioners make decisions more quickly. Being able to make easy diagnosis, more accurate prediction and avoiding unnecessary biopsys would enable to have better outcomes of diseases, as well as reducing the cost of unnecessary methods and allows for increased clinical trials.

The breast N is the highest mortality rate in women after lung N. The purpose of this study is to investigate early detection of breast N by machine learning methods, avoiding unnecessary biopsy. In previous studies, while variables including tumor information were used for diagnosis, in this study, variables that were mostly cultural and physical influences were used. In early diagnosis, it was investigated whether these variables were important or not, and it was also compared with studies done with different variables.

Machine learning classification methods for breast N diagnosis have been used. Each method has its own size reduction in the variables according to the suitability of the method. The most effective variable according to each method varied.

Two different applications were made on the same data set. The data set is divided into training and test set and applied in SPSS.18 Modular program. The second is the application in Weka with k-fold cross validation.

Logistic regression and Bayesian network are the best result of the machine learning methods used. After that, they are followed by support vector machines, CRT and C5.0 algorithm in decision trees and neural network.

Logistic Regression (78%) and Naive Bayes (78%) were the best results of machine learning methods used. These are followed by Support Vector Machines (76%), Decision Tree C5.0 algorithm (76%), and Artificial Neural Networks (74%) respectively. 3 fold cross validation performed in the case, the methods that give the best classification accuracy are Decision Tree C5.0 algorithm (76%), Logistic Regression (73%), Support Vector Machines (73%), Artificial Neural Networks (73%) and Naive Bayes (73%).

Key Words: Machine learning, logistic regression, support vector machines, decision trees, artificial neural networks

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Makine öğrenmesi yöntemleri son yıllarda tıp alanında oldukça yaygın kullanılmakta ve gelecek araştırmalara kapı açma konusunda oldukça önemli sonuçlar vermektedir.

Kıyan ve Yıldırım [1], Wisconsin meme kanseri verileri (WBCD) üzerinde istatistiksel sinir ağı ağlarını, radyal temel ağ(RBF), genel regresyon sinir ağı (GRNN) ve olasılıksal sinir ağı (PNN) performansını incelemişlerdir. Wisconsin meme kanseri veri setinde hastalığa ait tümör bilgileri bulunmaktadır ve bu değişkenler üzerinde tahmin yapılmıştır. Genel sınıflandırma performansları RBF için %96,18, PNN için %97, GRNN için %98,8 bulunmuştur.

Maglogiannis vd. [2], meme kanseri hastalığının prognozu ve teşhisi için bayesian sınıflandırıcıları ve yapay sinir ağlarını destek vektör makineleri (DVM) ile karşılaştıran bir sistem önermişlerdir. Optimize edilmiş DVM algoritması, tüm sınıflar için alternatif yaklaşımlardan üstün, yüksek doğruluk değerleri (%96,91'e kadar), özgüllük (%97,67 kadar) ve duyarlılığı (%97,84'e kadar) sergileyerek çok iyi performans göstermiştir.

Bradley [3], makine öğrenme algoritmaları için bir performans ölçütü olarak alıcı çalışma karakteristiği (ROC) eğrisi (AUC) altında kullanım alanlarını incelemiştir. Bir vaka çalışması olarak, altı tıbbi teşhis veri setinde altı makine öğrenme algoritmasını (C4.5, Çok Katlı Sınıflandırıcı, k-En Yakın Komşular ve Bir İkinci Dereceli Ayırt Edici Fonksiyon) değerlendirmiştir. Genel olarak, tüm veri setleri üzerinde öğrenme algoritmalarının her birinden elde edilen doğruluklarda büyük bir fark bulunmamıştır. Genellikle, bayes, çok

katlı sınıflandırıcı ve k-NN tabanlı yöntemler, genel doğruluk ve AUC açısından karar ağaçlarına göre daha iyi sonuç vermiştir.

Salama vd. [4], üç farklı veri seti kullanarak karar ağacı algoritmalarından J48, Çok Katmanlı Algılama (MLP), Naive Bayes (NB), Sıralı Minimum Optimizasyon (SMO) ve K-En Yakın Komşu (k-NN) yöntemlerini değerlendirmişlerdir. Sınıflandırma başarıları için, karışıklık matrisi ve 10 katlı çapraz doğrulama yöntemi kullanılmıştır. Ayrıca, her veri kümesinde en uygun çok sınıflandırıcı yaklaşımı elde etmek için bu sınıflandırıcılar arasında sınıflandırma düzeyinde bir füzyon oluşturulmuştur. Deneysel sonuçlar MLP ve J48'in temel bileşenler analiz ile füzyonunu kullanan sınıflandırmada WBCD veri setini kullanan diğer sınıflandırıcılardan daha üstün olduğunu göstermiştir.

Sawarkar vd. [5], gereksiz biyopsiden kaçınmak amaçlı meme kanserini destek vektör makinesi ile teşhis etmek istemiştir. Meme kanserinin eldeki imkanlarla tespit edilmesinin etkinliği %85'tir ve elde edilen destek vektörü makinelerinin verimliliği %97'dir. Bu yüksek doğruluk oranı, doktorun biyopsiden kaçınma kararını desteklemek için kullanılabilir olduğunu göstermiştir.

Polat ve Güneş [6], meme kanseri teşhisini en küçük kareler destek vektör makineleri algoritması kullanılarak gerçekleştirmiştir. En küçük kareler destek vektör makinelerinin tanısal performansını göstermek için; sınıflandırma doğruluğu, duyarlılık ve özgüllük, k-kat çapraz doğrulama yöntemi ve hata (karışıklık) matrisi kullanılmıştır. Elde edilen sınıflandırma doğruluğu %98,53 bulunmuştur.

Decruyenaere vd. [7], böbrek naklinden sonra gelişen gecikmiş graft fonksiyonun lojistik regresyon ve makine öğrenmesi yöntemleriyle tahmin modelleri değerlendirmişlerdir. Doğrusal diskriminant analizi, kuadratik diskriminant analizi, destek vektör makineleri, karar ağacı, rasgele orman ve stokastik gradient boosting olmak üzere 6 tahmin yöntemi kullanılmıştır. Modellerin performansı, 10 kat çapraz doğrulama yönteminden sonra duyarlılık, pozitif tahmin değeri ve AUC hesaplanarak değerlendirilmiştir. Destek vektör makinelerinin duyarlılığı diğer yöntemlerden iyi ayırt edici kapasitede (%84,3 AUC) ve lojistik regresyondan üstün olan tek yöntem olduğu görülmüştür.

Hurtado vd. [8], insan elinin ölçümleri ile bir takım demografik özellikler arasındaki ilişkileri araştırmışlardır. El özelliklerinden demografik bilgileri tahmin edebilmek için

doğrusal lojistik regresyon ve makine öğrenmesi sınıflandırma yöntemlerini değerlendirmişler ve hem ilişkinin gücü hem de bu ilişkiyi destekleyen temel özellikler için kanıtlar sağlanmaya çalışmışlardır. Makine öğrenmesi sınıflandırma yöntemleri, cinsiyet (%89), ağırlık (%65), boy (%70) ve ayak numarasını (%80) tahmin etmede lineer lojistik regresyondan daha iyi sonuçlar vermiştir. Sağ el görüntülerini kullanarak en iyi performansı destek vektör makineleri (DVM) sınıflandırıcısı %88,7 başarı oranı ile sol el görüntülerini kullanarak en iyi performansı ise Naive Bayes (NB) sınıflandırıcısı %87,7 doğruluk yüzdesiyle vermiştir.

Saraoğlu vd. [9], diyabet, pre diyabet ve sağlıklı kişilerin avuç içi terleme ölçümlerinden diyabet hastalığına etki eden parametreleri değerlendirmişlerdir. Veri kümelerinde boyut azaltmak için temel bileşenler analizi kullanılırken, sınıflandırma için ise destek vektör makineleri yöntemi kullanılmıştır. Terlemeye etki eden parametrelerden en yüksek sınıflandırma başarısında glukoz %82 ve HbA1c %84 bulunmuştur.

Şatır vd. [10], çalışmalarında Glukom hastalığı teşhisi için karar ağaçları ve yapay sinir ağları sınıflandırma yöntemlerini kullanılmışlar ve çapraz doğrulama ile performans değerlendirme gerçekleştirmişlerdir. En yüksek başarı oranı C4.5 algoritmasına ait olup %93,45 olduğu görülmüştür.

Ulgen vd. [11], sağlıklı ve kontrol grubunu içeren hastaların bir klinik çalışmada kullanılan EMG tarama yöntemiyle Juvenil Myoklonik Epileps hastalarının ve normal kontrol gruplarının sınıflandırılmasını ileri sinir ağları, destek vektör makineleri, karar ağaçları ve naive bayes yöntemlerini kullanarak gerçekleştirmişlerdir. Modelleri eğitmek ve test etmek için çapraz doğrulama uygulanmıştır. ROC eğrileri, 4, 6, 8 ve 10 kat çapraz doğrulama değerleri için çizilmiştir. Hem duyarlılık hem de yanlış pozitif değerleri dikkate alındığında, sinir ağı tabanlı modelin sağlıklı ve hasta grupları sınıflandırma performansının Naive Bayes, destek vektör makineleri ve karar ağaçlarından daha iyi sonuç verdiği görülmüştür.

Çomak vd. [12], en küçük kareler destek vektör makinesi (LSSVM) ile bulanık ağırlıklandırma ön-işleme yöntemlerini melezleştirerek karaciğer bozukluğu teşhis problemini çözmek istemişlerdir. Önerilen sistemin doğruluğunu istatistiksel olarak test etmek için ROC eğrileri kullanılmıştır. LSSVM' de %60 sınıflandırma doğruluğu elde

edilirken, LSSVM ile bulanık ağırlıklandırma ön işleme birleşiminde %94,29 sınıflandırma doğruluğu elde edilmiştir.

Dreiseitl vd. [13], ortak nevüs, diplastik nevüs veya melonom olarak pigmentli deri lezyonlarını sınıflama üzerine destek vektör makineleri (DVM), karar ağaçları, yapay sinir ağları, lojistik regresyon ve k en yakın komşuluk yöntemlerinin ayırıcı güçlerini analiz etmişlerdir. Yöntemlerin ayırıcı gücünü ölçmek için ROC analizi kullanılmıştır. En doğru sınıflandırma yüzdesi %97 ile vektör destek makinelerinde görülmüştür.

Verplancke vd. [14], yoğun bakım ünitesinde yatan hematolojik kötü huylu tümörleri olan hastalarda hastane mortalitesinin tahmin edilmesinde, çoklu lojistik regresyon ile destek vektör makinesi yöntemlerini karşılaştırmışlardır. Ayrım için alıcı çalışma karakteristiği (ROC) eğrisi altındaki alan kullanılmıştır. Çoklu lojistik regresyon ve destek vektör makinesi için ROC eğrisi altındaki alan sırasıyla 0.768 (±0.05) ile 0.802 (±0.04) bulunmuştur.

Yu vd. [15], akciğer adenokarsinomu ve skuamöz hücreli karsinoma hastalarının histopatoloji görüntülerinden distile edilen objektif özellikler vasıtasıyla prognostik tahmini geliştirmeyi amaçlamışlardır. Naive Bayes, Gauss çekirdeği ile destek vektör makineleri (DVM), doğrusal çekirdekli DVM, polinom çekirdeği olan DVM, sınıflandırma ağaçları için bagging, koşullu çıkarım ağaçlarını kullanan rastgele orman ve Breiman'ın rastgele ormanı olmak üzere yedi sınıflandırıcı kullanılmıştır. Niceliksel görüntü özellikleri akciğer adenokarsinomunun histopatoloji görüntülerini akciğer skuamöz hücreli karsinomun görüntülerinden ayırma başarısında en iyi destek vektör makinesi; Bagging (AUC=0.75) Naive bayes (AUC=0.73), rasgele orman (AUC=0.76), Gaussian tabanlı DVM (AUC=0.85), Lineer kernel tabanlı DVM (AUC=0.82) Polinom tabanlı DVM (AUC=0.78). En iyi sınıflandırma yönteminin DVM olduğu görülmüştür.

Rampun vd. [16], prostat kanser teşhisi için Bayes ağları, rasgele orman ve k-NN yöntemlerini kullanmışlardır. Değerlendirme sonuçları en iyi sınıflandırıcıların Bayes ağları (%92,8 ±%5,9), rasgele orman (%89,5 ± 7.1) ve k-NN (%86,9 ±%7,5) olduğunu göstermiştir.

Salah vd. [17], cilt kanseri görüntülerindeki tümör yerinin tespiti ve cilt kanseri türünü belirlemek için yapay sinir ağı ağı ve bulanık mantık kullanmışlardır. Hiyerarşik sinir ağı

kullanılarak cilt kanseri tanısının doğruluk oranı %90,67 iken bulanık mantık kullanıldığında cilt kanseri tanısının doğruluk oranının %91,26 olduğu görülmüştür.

Soni vd. [18], 3 farklı kalp hastalıkları veri setinde naive bayes, k-NN, karar ağaçları, kümeleme yoluyla sınıflandırma, sinir ağı, kaba kurallar yöntemleriyle en iyi teşhis yüzdeliğinin bulunmasını amaçlamışlardır. Her veri setinde karar ağaçları ve Naive Bayes yöntemi diğer yöntemlere göre daha iyi sonuçlar vermiştir.

Wang vd. [19], ulusal uyku araştırmalarını baz alarak 8 saat uyku, 6 saat uyku ve 4 saat uyku olmak üzere üç farklı uyku koşulunda sekiz kişiden uyanık EEG sinyalleri toplamak için bir uyku deneyi tasarlamışlardır. Farklı uyku kalitelerini EEG özellikleriyle doğru bir şekilde ayırt edebilen makine öğrenmesi sınıflandırma yöntemleri kullanılmıştır. Bu yöntemler k-en yakın komşu (k-NN), destek vektör makinesi (DVM), ayırt edici grafik düzenlenmiş aşırı öğrenme makinesi (GELM) dir. GELM'nin sınıflandırma doğruluğu ortalama olarak %62,16, ki bu da k-NN ve DVM'den daha yüksek olduğu görülmüştür.

Aljaaf vd. [20], yürüme sırasında ayak bileği, diz, kalça ve pelvisin 3D Euler açılarından frontal düzlem diz abdüksiyon momentini öngörmek için C4.5 karar ağacı, rasgele orman, doğrusal regresyon ve çok katmanlı sinir ağı makine öğrenme algoritmalarının performansını karşılaştırmışlardır. Doğrulama yöntemi olarak Holdout yöntemi kullanılmıştır. Bu çalışmada, veri seti eğitim için %70'e ve test için %30'a bölünmüştür. Tahmini modellerin performansını değerlendirmek için R^2, RMSE ve geri çağırma eğrisi altındaki alan yöntemleri kullanılmıştır. Üç tip deney yapılmıştır. İlk iki deneyde en iyi performansı rasgele orman yöntemi gösterirken son deneyde çok katmanlı sinir ağı en iyi sonucu vermiştir.

Parthiban ve Srivatsa [21], diyabet hastalarındaki kalp rahatsızlıkları teşhisini Naive Bayes ve Destek Vektör Makineleri sınıflandırma yöntemlerini kullanarak tahmin etmişlerdir. Sınıflandırıcının doğruluğunu test etmek için 10 kat çapraz doğrulama yöntemi kullanılmıştır. Destek vektör makineleri, %94,60, Naive Bayes ise %74 doğruluk yüzdesine sahip bulunmuştur.

Krawczyk vd. [22], 3 baş ağrısı türünün sınıflandırılması için otomatik bir tıbbi karar destek sistemi önerisi sunmuşlardır. Makine öğrenme tekniklerinden naive bayes, C4.5 karar ağacı, destek vektör makineleri, boosting, bagging, random forest sınıflandırma

yöntemleri kullanılarak performansları karşılaştırılmıştır**.** Tüm testler, on kat çapraz doğrulama kullanılarak sonuçlandırılmıştır. ReliefF Greedy özellik seçimi ile Naive Bayes %76,17, SVM %78,41, C4.5 %80,14, rasgele orman %81, adaBoost %77,89 ve bagging %80,14 doğruluk yüzdelerini vermiştir.

Çınar vd [23], prostat kanseri teşhisinde gereksiz biyopsiden kaçınmak için bir sistem tasarlamışlardır. Prostat kanseri, prostat hacmi, yoğunluğu ve benzeri özellikleri kullanarak erken teşhis için üç farklı yapay sinir ağı medtodu ve destek vektör makineleri kullanılmıştır. Özellik seçimi için bağımsız örnek t-testi uygulanmıştır. Performans değerlendirme için çapraz doğrulama yöntemi kullanılmıştır. Sınıflandırıcının başarısını tahmin etmek için, duyarlılık, özgüllük ve doğruluk hesaplanmıştır. Destek vektör makinesi algoritmasının yapay sinir ağı öğrenme algoritmalarından daha iyi sonuç verdiği görülmüştür. DVM'nin polinom temelli çekirdek fonksiyonu en iyi sonucu vermiştir (doğruluk: %79, özgüllük: %78,8).

Demšar vd. [24], prostat kanseri nüksü üzerinde etkili olan değişkenleri tahmin etmede makine öğrenme yöntemlerinin modern istatistiksel yöntemlerin yanında daha iyi performans sergilediğini ve sembolik yinelenme modelleri oluşturarak model verileri içindeki ilişkilere daha fazla bilgi sağlayabileceğini göstermek istemişlerdir. Sınıflandırma yöntemleri olarak lojistik regresyon, karar ağaçları, birliktelik kuralı, Naive Bayes ve yapay sinir ağları kullanılmıştır. İstatistiksel sağkalım analiz yöntemleri olarak Cox orantısal risk modeli ve regresyon yöntemleri (yapay sinir ağları) kullanılmıştır. Farklı modelleme tekniklerinin prostat kanseri verilerine uygulanması durumunda genel olarak Naive Bayes ve Cox orantısal risk modeli en iyi performansı göstermiştir. Lojistik regresyon en yüksek uyum endeksini elde etmiştir. Cox modeli için 0,74, YSA için 0,76 doğruluk elde edilmiştir.

1.2 Tezin Amacı

Veri madenciliği yöntemlerinden makine öğrenmesi tıpta son yıllarda yaygın bir şekilde kullanılmaktadır. Meme kanseri teşhisinde makine öğrenmesi kullanımı ile ilgili birçok çalışma yapılmıştır. Fakat bu çalışmaların hepsi meme kanseri tümörüne ilişkin değişkenlerin etkilerini tahminlemeye ve modelin bu değişkenler üzerindeki doğruluklarını bulmaya yöneliktir.

Yapılan bu çalışmada ise meme kanseri olanların ırk, yaş, mamagrom görüntüleri gibi fiziksel ve kültürel özelliklere sahip değişkenler kullanılarak teşhis edilmeye çalışılmıştır.

1.3 Hipotez

Bu çalışmada hipotezimiz; “Makine öğrenmesi yöntemleri meme kanserine etki eden fizyolojik ve kültürel değişkenler ile meme kanseri riskini daha iyi açıklamaktadır ve bu sonuçlarda makine öğrenmesi yöntemleri arasında farklılık vardır.”

BÖLÜM 2

MAKİNE ÖĞRENMESİ ve TARİHİ

Makine öğrenmesi kavramı 1946 yılında ilk elektronik bilgisayar olan ENIAC'ın geliştirilmesiyle başlamıştır. Takip eden yıllarda farklı elektronik bilgisayarların gelişmesi ile daha önceden hesap makineleriyle yapılamayan problemler çözülebilir olmaya başlamış ve bu gelişmeler bilim insanlarını bir makinenin insana özgü davranışları nasıl öğreneceği düşüncesine sürüklemiştir.

1950 yılında Alan Turing makinelerin düşünebilir olacağı ile ilgili bir fikir ortaya atmış ve bunun için bir test önermiştir. Bu test bugün "Turing Testi" ismiyle bilinmektedir. Bu test, bir insan ile bir makine arasında sözlü bir konuşmadan oluşur. Eğer insan, artık başka bir insanla mı yoksa makineyle mi konuştuğunu söyleyemeyecek duruma gelirse, makine bilinçlenmiş demektir [25].

1950'lerde Shannon ve Turing ilk satranç oyun programını geliştirmişlerdir. 1951' de ilk yapay sinir ağı temelli bilgisayar SNARC, Minsky ve Edmonds tarafından yapılmıştır. Yapay zekâ bir bilim dalı olarak 1956 yılında C. Shannon, M. Minsky ve J. McCarthy tarafından başlatılmıştır.

Yapay zekâ insanlar gibi düşünebilen, mantık yürüten makineler geliştirebilmekle ilgilenen bir bilim dalıdır. Yapay zekâ; bilişsel modelleme yaklaşımı, makine öğrenmesi (öğrenme ve bilgi tabanlı sistemler) ve sembolik düşünmeye dayanan kavram olarak üç alanda incelenir.

2.1 Makine Öğrenmesi Nedir?

Günümüzde teknolojinin de gelişmesiyle bilgisayar ortamından sürekli bir akış içerisinde çok büyük veri tabanı elde edilmektedir. Bu verinin işlenebilmesi insanın sınırlı kaynaklara sahip olması nedeni ile insanlar tarafından uygulanması oldukça zordur. Karmaşık yapıda çok veri ve çok fazla değişken elde edilmesi veri madenciliği kavramını doğurmuştur. Veri madenciliği büyük ve karmaşık verilerin işlenebilir hale gelmesini ve anlamlı çıkarımlar yapmaya olanak sağlar. Veri üzerinde makine öğrenmesi metodlarının uygulanması veri madenciliği olarak adlandırılır [26]. Veri madenciliği alanında makine öğrenmesi finansal işlemlerde, sağlık kayıtlarında, kredi başvurularında ekipman bakım kayıtları gibi büyük veri tabanına sahip alanlarda kullanılabilir [27].

2.2 Öğrenme Kavramı

Öğrenme kavramını açıklayacak olursak;

Öğrenme, deneyimi bilgiye ve uzmanlığa dönüştürme sürecidir [28]. Öğrenme, karşılaştığımız olayları ya da durumları değerlendirerek kazandığımız tecrübeleri mevcut durumda açığa çıkarma olarak da tanımlanabilir.

Makineler de insan zekâsı gibi öğrenebilir mi sorusunun cevabı bizi makine öğrenmesi kavramına getirmiştir. Makine öğrenmesi geçmiş verileri istatistik ve bilgisayar bilimi ile harmanlayarak mevcut durum hakkında çıkarımlar, tahminler yapmayı amaçlamaktadır.

Literatürde makine öğrenmesinin tanımları şöyle belirtilmiştir;

Makine öğrenmesi deneyim ile otomatik olarak gelişen bilgisayar programlarının nasıl yapılandırılması gerektiği sorusu ile ilgilenen bir alandır [27]. Makine öğrenmesi yapay zekâ ile ilgili görevleri uygulayan sistem değişikliklerini ifade eder [29]. Makine öğrenmesi, geçmiş deneyimleri ve örnek verileri kullanarak bir performans ölçütünü en iyi hale getirebilen bilgisayar programlamadır [26]. Makine öğrenmesi otomatik olarak veri modellerinin belirlenmesi ve gelecek veri tahmini için karmaşık modellerin kullanılması ya da belirsizlik altında karar verme yöntemlerinin uygulanmasıdır [30]. Makine öğrenmesi bir yapay bilimdir ve deneyimler ile insan yapımı şeyler (yapay)

özellikle de algoritmalar ile performans geliştiren bir çalışma alanıdır [21]. Makine öğrenmesi sistemleri veriden programları otomatik bir şekilde öğrenmektir [22].

Mitchell makine öğrenmesinin davranışını şu şekilde ifade etmiştir:

Bir bilgisayar programı, P ile ölçülen G'deki performansı D deneyimi ile gelişirse görev G ve performans ölçüsü P ile ilgili olarak D deneyimlerinden öğrenildiği söylenir [27].

Dama oynamayı öğrenen bir bilgisayar programı kendine karşı oynadığı oyunla bir deneyim elde ederek performansını geliştirebilir. Bu performans dama oynamanın dahil olduğu görev sınıflarında kazanma yeteneği ile ölçülür. İyi tanımlanmış bir öğrenme probleminde üç özellik değerlendirilmelidir: görev sınıfı, geliştirilmiş performans ölçüsü ve deneyimin kaynağı.

Görev G: dama oynama

Performans ölçümü P: rakiplerine karşı kazandığı oyunların yüzdesi

Deneyim D: kendine karşı oynadığı oyunlar,

Mitchell'in yaptığı tanıma örnek olarak gösterilebilir.

2.3 Makine Öğrenmesi Kullanım Alanları

Teknoloji makine öğrenmesine dayalı bir tabanda ilerlemiştir: arama motorlarının en iyi sonuçları getirmesinde, spam e-postaları ve diğer e-postların birbirinden ayırma işleminde, dijital kameralarda yüz tanıma işlemlerinde, akıllı telefonlarda ses komutu algılama uygulamalarında, online alışverişlerde öneri sekmelerinde, otomobillerde kaza önleme sistemlerinde, biyoinformatik alanında, hastalık teşhisinde ve astronomi gibi bilimsel uygulamalarda da makine öğrenmesi algoritmaları kullanılmaktadır.
Bu uygulamaların ortak bir özelliği, bilgisayarların daha geleneksel kullanımlarından farklı olarak, algılanması gereken kalıpların karmaşıklığı nedeniyle, bir insan programcısı bu tür görevlerin nasıl yürütüleceğine ilişkin açık ve ayrıntılı bir spesifikasyon sunamaz. Akıllı varlıklardan örnek alınarak, birçok yeteneklerimiz tecrübelerimizden öğrenme yoluyla edinilmiştir.

Makine öğrenme araçları, öğrenme ve uyarlama yeteneği olan programlar sağlamakla ilgilidir. Makine öğrenmesinin işleyişine bir örnek verecek olursak: spam e-postları

filtreleyen bir makine programlamak istediğimizde makine daha önce kullanıcı tarafından etiketlenmiş spam mesajlarını ezberleyerek yeni bir mesaj geldiğinde makine önceki spam e-postalarını tarayacak ve eşleşirse spama aksi takdire gelen klasörüne taşıyacaktır.

Kısaca makine öğrenmesi günlük hayatımızda bize yardımcı olacak en ufak teknolojiden uzay görevlerini yerine getirecek kadar geniş bir alana sahiptir ve önümüzdeki yıllarda teknolojinin daha da ilerlemesiyle her alanda kullanılacak ve geliştirilecek bir bilim dalıdır.

Makine öğrenmesi ile ilgili birçok örnek verilebilir.

- Optik karakter tanıma: el yazısı karakter görüntülerini kategorize etmek ve sonrasında tanıma
- Yüz algılama: görüntülerde yüz bulma
- Spam filtreleme: e posta mesajlarını spam veya spam olmayan olarak belirleme
- Konu seçme: haber makalelerini siyaset, spor, eğlence olarak ayırma.
- Konuşulan dili anlama
- Tıbbi teşhis
- Müşteri segmentasyonu: belli müşterilere özel hizmet sağlamak
- Dolandırıcılık tespiti
- Hava tahmini

2.4 Öğrenme Türleri

Öğrenme çok geniş bir alandır. Makine öğrenmesinin alanı farklı tiplerdeki öğrenme görevlerini ele alan çeşitli alt alanlara ayrılmıştır. Bunlar;

2.4.1 Denetimli Öğrenme

Öğrenici eğitim verisi olarak bir örnekler dizisi alır ve görünmeyen tüm noktalar için öngörüler yapar. Sınıflandırma ve regresyon problemi olarak ikiye ayrılır.

Önceki bölümde ele alınan spam algılama sorunu denetimli öğrenme örneğidir.

Sınıflandırma: Geçmiş bilgilerin hangi sınıftan olduğu bilgisine sahip olduğu biliniyorsa mevcut verinin hangi sınıfa ait olduğunun belirlenmesidir. Birçok sınıflandırma yöntemi vardır. [31]

Regresyon: Regresyon, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır.

Tahmin: Öğrenilen verinin nicel olması durumunda kullanılan yöntemlerin ürettiği değerlerdir.

Kestirim: Tahmini yapmaya yardımcı olan eğrinin parametrelerinin belirlenmesinde kullanılır.

2.4.2 Denetimsiz Öğrenme

Sınıf bilgisi bulunmayan veri içerisindeki grupları keşfetmeyi amaçlar. Geçmiş verileri kullanmaz ve mevcut verilerden çıkarım yapmayı hedefler. Günümüzde de uygulanabilirlik ve daha doğru sonuçlar vermesi dolayısı ile denetimli öğrenme yöntemleri tercih edilmektedir. Kümeleme ve boyut azaltma, denetimsiz öğrenme problemlerine örnektir [32].

Kümeleme: Geçmişteki verilerin sınıf ya da küme bilgisi olmadığı takdirde mevcut verilerin birbirlerine yakın benzerlikleri ile belirlenmesidir.

Örneğin, bir şirket geçmiş verilerine bakarak müşterilerinin profilini çıkarmak isteyebilir. Bu durumda kümeleme algoritmaları benzer davranış gösteren müşterileri aynı küme içine toplayacaktır. Bu sayede şirket farklı müşteri grupları için farklı hizmetler sunmuş olacaktır.

2.4.3 Yarı Denetlenmiş Öğrenme

Öğrenici hem etiketli hem de etiketsiz veriden oluşan bir eğitim örneği alır ve görünmeyen tüm noktalar için tahminler yapar.

Etiket, tahmin etmeye çalışılan kategoridir. Eğitim sırasında, öğrenme algoritmasına etiketli örnekler verilirken, test sırasında sadece etiketsiz örnekler sağlanır [33].

Etiketlenmemiş verilerin öğreniciye ulaşılabilir şekilde dağıtılmasının, denetlenen ortamdan daha iyi bir performans elde etmesine yardımcı olabilir. Bunun gerçekleşebileceği koşulların analizi, çok modern teorik ve uygulamalı makine öğrenimi araştırmasının konusudur.

2.4.4 Çevrimiçi Öğrenme

Çevrimiçi öğrenme bir dizi ardışık turda gerçekleşir. Her turda, öğrenici etiketsiz bir eğitim noktası alır, bir tahminde bulunur. Tahmin edilen cevap ile doğru cevap arasında fark bir kayıptır. Çevrimiçi öğrenmenin amacı, tüm turlarda biriken kayıpları en aza indirmektir. Daha önce bahsedilen öğrenmelerin aksine, çevrimiçi öğrenmede dağılımsal varsayım yapılmaz [31].

2.4.5 Takviyeli Öğrenme

Takviyeli öğrenme, bulunduğu ortamı algılayan ve kendi başına kararlar alabilen bir sistemin, hedefine ulaşabilmesinde doğru kararlar almayı nasıl öğrenebileceğini gösterir [31].

2.4.6 Aktif Öğrenme

Öğrenici, genellikle bir oracle'a sorarak yeni puan için etiket talep ederek eğitim örnekleri toplar. Aktif öğrenmede amaç, daha az sayıda etiketlenmiş örnekle standart gözetim altında öğrenme senaryosuyla karşılaştırılabilir bir performans elde etmektir [32].

BÖLÜM 3

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ

Makine öğrenmesi yöntemleri sınıflandırma ve kümeleme problemleri olarak iki başlıkta incelenecektir. Sınıflandırma yöntemleri, geçmiş bilgilerin sınıf bilgisinin bilindiği durumda mevcut verinin hangi sınıfa ait olduğunu belirlemek amacı ile kullanılır. Birçok sınıflandırma yöntemi vardır. Bu çalışmada en çok kullanılan sınıflandırma yöntemlerinden bazıları anlatılacaktır

1) Lojistik regresyon
2) Destek vektör makineleri
3) En yakın k komşu algoritması
4) Karar ağaçları
5) Naive bayes sınıflandırma algoritması
6) Yapay sinir ağları

3.1 Lojistik Regresyon

Lojistik regresyon, lineer regresyondan farklı olarak bağımlı değişkenini çoklu olarak ele alarak hem sınıflandırma hem tahmin gerçekleştirir. Lojistik regresyon modeli aşağıdaki gibi yazılmaktadır.

$$L = \ln \frac{p_i}{1-p_i} = b0 + b_1 X_i + e_i \quad (3.1)$$

Lojistik regresyon modelinin parametreleri, analitik olarak elde edilemediğinden, iteratif bir yöntem olan maksimum olabilirlik tekniği ile tahmin edilmektedir. Tahmini parametrelerin başlangıç değerleri kullanılır ve örneklemin bu parametrelerle bir

popülasyondan çıkma ihtimali hesaplanır. Tahmini parametrelerin değerleri, tahmini parametreler için maksimum olasılık değeri elde edilinceye kadar tekrar tekrar ayarlanır. Yani, maksimum olasılık yaklaşımları, verileri gerçekte "en olası" gözlemlenen parametrelerin tahminlerini bulmaya çalışmaktadır [34].

3.2 Destek Vektör Makineleri

Destek vektör makinesi (DVM) Vapnik ve grubunun AT & T Bell Laboratuarlarında sınıflandırma ve regresyon için kullandığı bir kavramdır [35].

DVM sınıfları birbirinden ayıran, her iki sınıfa da en uzak olan hiperdüzlem adı verilen özel bir çizginin bulunmasını amaçlar. Eğitim verileri kullanılarak hiperdüzlem bulunduktan sonra, test verileri sınırın hangi tarafında kalmışsa o sınıfa dahil edilir. Hiperdüzlem doğrusal olmayan örnek uzayını doğrusal olarak ayrıştırır ve farklı örnekler arasında maksimum ayrımı yapmayı sağlar [36].

DVM'ler, ayrımcılık gücünü artırmak için veri modellerini daha yüksek boyutlu bir alanda temsil etme yaklaşımını doğrusal olmayan haritalama işlevi ile kullanır [37]. DVM, hata oranını en aza indirmeyi amaçlayan risk azaltma ilkesine dayalıdır [38].

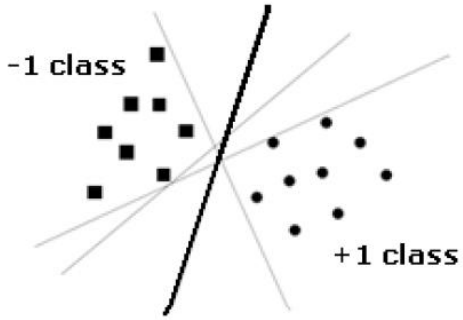
Şekil 3.1 de görüldüğü gibi, veri gruplarında birbirinden ayrılabilen çok sayıda hiperdüzlem ihtimali bulunmaktadır. Minimum düzeyde hata toleransı elde etmek için gruplardan her birine uzak bir hiper düzlem bulunmalıdır.

İki taraflı ve çok taraflı sınıflandırma yöntemleri göz önüne alındığında, veri kümelerindeki sınıfların aşağıdaki gibi uygun sonuçlara sahip olduğu görülmektedir:

$x = x_1 \ldots . x_i$ bir dizi eğitim örneği olsun ve $y = y_1 \ldots y_i$, karşılık gelen olsun

Sınıflamaların seti, burada x_i öğrenecek sınıfın bir üyesiyse y_i= 1 ve aksi halde y_i=-1'dir

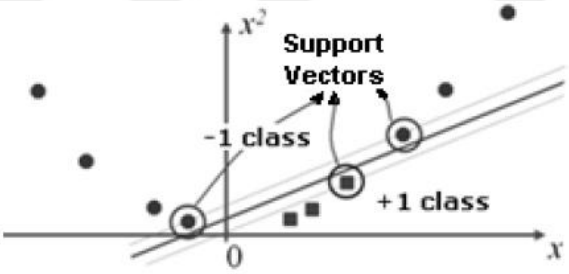
$$D = (x_1, y_1) \ldots, (x_i, y_i), x \epsilon Rn, y \epsilon \{1, -1\} \quad (3.2)$$



Şekil 3. 1 Optimum ayırıcı düzlemler

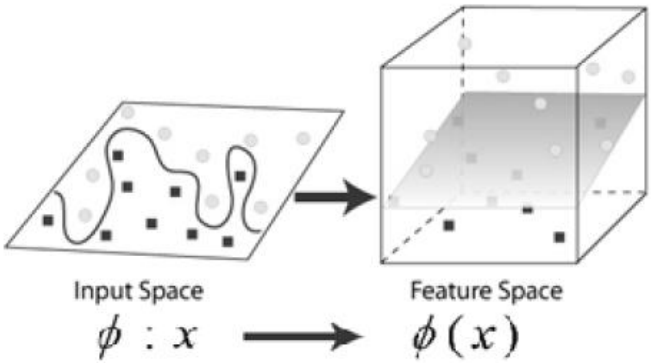
Verilerin doğrusal yöntemlerle ayrıştırılamadığı durumlarda doğrusal olmayan sınıflandırmalar kullanılabilir. Doğrusal olmayan sınıflandırmalar, verileri çok boyutlu bir alana taşıyarak sınıflandırmaları gerçekleştirir.

Şekil 3. 2 Tek boyutlu uzayda doğrusal yöntemle ayrılmayan veri kümesi



Şekil 3. 3 İki boyutlu uzaya taşınarak doğrusal ayrılma

Doğrusal olmayan problemlerin çözümü, çekirdek işlevleri (Ø) ile örnekleri çok boyutlu ve doğrusal yöntemlerle ayrılabilen bir alana taşıyarak gerçekleşir.



Şekil 3. 4 Çok boyutlu uzayda bir veri kümesinin sınıflandırılması

Şekil 3.4 'te, Ø fonksiyonları ile x girdileri çok boyutlu bir boşluğa Ø (x) olarak taşınmaktadır. Böylece optimum uzay hizası, örneklem alanındaki optimizasyon problemlerinin çözümü ile elde edilir.

Bu sınıflandırma f (x) optimizasyon fonksiyonu ile yapılır [39]. Destek vektör makinelerinin optimizasyon formülü aşağıdaki gibidir,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, y_i) + b \quad (3.3)$$

Bu durumda hiper düzlem olmadığından b bertaraf edilir.

Söz konusu b sapma değeri, çekirdek işlevlerinde yer alır. Destek Vektörü sınıflandırmalarında çeşitli çekirdek işlevleri kullanılır. En çok kullanılanlar şunlardır [39].

Lineer Kernel:

$$K(x_i, y_j) = x_i^T x_j \quad (3.4)$$

Polinom Kernel:

$$K(x_i, y_j, c, d) = (c + x_i^T x_j)^d \quad (3.5)$$

Sigmoid Kernel:

$$K(x_i, y_j) = \tanh(\gamma x_i^T x_j + \Phi) \quad (3.6)$$

Radyal Tabanlı Kernel:

$$K(x_i, y_j, \sigma) = e^{\frac{|xi-xj|^2}{2\sigma^2}} \quad (3.7)$$

Bu çalışmada lineer kernele dayalı destek vektör makinesi kullanılmıştır.

3.3 En yakın k komşu algoritması

K-en yakın komşu algoritması (k-NN), özellik uzayında en yakın eğitim verilerine dayalı nesneleri sınıflandırmak için kullanılan bir yöntemdir. K-NN, örnek tabanlı bir öğrenme türüdür. K-en yakın komşu algoritması, tüm makine öğrenme algoritmalarının en basitleri arasındadır. Fakat k-NN algoritmasının doğruluğu, gürültülü veya alakasız özelliklerin bulunması durumunda ciddi oranda bozulabilir [40].

K-en yakın komşularında (k-NN) yeni bir örnek, en yakın komşularının sınıf değerine göre sınıflandırılır. Bir test örneğinin sınıfı belirlenirken eğitim kümesinde o örneğe en yakın k adet örnek seçilir. Her örnek birbirine en yakın diğer örneklere uzaklığını hesaplar ve en yakın uzaklığa sahip örnekler birleştirilir. Bu işlem istenilen küme sayıları elde edilinceye kadar devam eder.

x bir örnek ise iki örnek arasındaki uzaklık

$$d(x) = | x - x^t | \quad (3.8)$$

olarak tanımlanırsa ;

k en yakın komşuluk yönteminin olasılık yoğunluk tahminicisi;

$$\hat{p}(x) = \frac{k}{2Nd_k(x)} \quad (3.9)$$

Test örneği ile eğitim örnekleri arasındaki mesafe çeşitli şekillerde hesaplanabilir. Nümerik değişkenler için en çok kullanılan yöntem Öklid uzaklıklarını kullanmasıdır. Nominal değişkenler için ise basit eşleme uzaklığı ve Jaccard uzaklığı kullanılabilir.

3.4 Karar Ağaçları

Karar ağacı öğrenmesi, otomatik öğrenme için en yaygın kullanılan algoritmalardan biridir. Alan bilgisi veya parametre ayarı gerektirmez ve yüksek boyutlu verileri

işleyebilir. Karar ağaçlarından elde edilen sonuçların okunması ve yorumlanması kolaydır.

Karar ağacı, düğümlerden (bir öznitelik değerini test eden), dallardan (öznitelik değerine dayanan yol) ve yapraklardan (sınıflandırmayı sağlar) oluşur [30]. Bir karar ağacı, verileri (ana düğüm) en iyi bölme özelliği ile iki alt kümeye (alt düğümler) ayırır. Ortaya çıkan iki alt küme, daha sonra iki alt düğüme ayrılmış yeni ana düğüm olur. Bu prosedür, tüm gözlemler sınıflandırılana kadar devam eder [11].

Karar ağacı, bir ağaç yapısı şeklinde sınıflandırma veya regresyon modelleri oluşturur. Bir veri kümesini daha küçük ve daha küçük alt kümelere bölerken, aynı zamanda ilişkili bir karar ağacı kademeli olarak geliştirilir.

Karar ağaçları hem kategorik hem sayısal verileri işleyebilir. En çok kullanılan algoritma olan ID3 ün entropi ve bilgi kazancı formülasyonu aşağıdaki gibidir. Entropi, bir süreç için rassal bir olayın gerçekleşmesi durumunun beklenen değeridir. Eşit olasılıklı durumlar yüksek belirsizliği temsil eder. Örnek tamamen homojen ise, entropi sıfırdır ve örnek eşit olarak bölünmüşse, entropi bir'e eşittir. Entropi, örnek kümesinin homojenliğinin bir ölçüsüdür.

$$\mathrm{E(S)} = -\mathrm{p(P)log2\ p(P)} - \mathrm{p(N)log2\ p(N)} \tag{3.10}$$

Bilgi kazancı, entropi veya belirsizlikte beklenen azalmayı ölçer.

$$\mathrm{Kazanç(S,A)} = \mathrm{Entropi(S)} - \sum_{\mathrm{v}\epsilon\mathrm{A}} \frac{|\mathrm{S_v}|}{|\mathrm{S}|} \mathrm{Entropi(S_v)} \tag{3.11}$$

Değerler (A), A özniteliği için olası tüm değerler kümesidir ve Sv özniteliğinin A değeri olan S altkümesidir.

$$\mathrm{Sv} = \{\mathrm{s\ in\ S} \mid \mathrm{A\ (s)} = \mathrm{v}\} \tag{3.12}$$

Kazanç denklemindeki ilk terim, orijinal yığının entropisi S, ikinci terim, S, A niteliğini kullanarak bölümlendirildikten sonra entropinin beklenen değeridir. En yüksek bilgi kazancına sahip özellik, ağaçta dallanma yapmak için tercih edilir [41].

3.5 Naive Bayes

Naïve Bayes sınıflandırıcısı, olasılıksal Bayes kuralı üzerine kuruludur ve özellikle girdilerin boyutsallığı yüksek olduğunda uygundur [42].

Belirli bir sınıftaki bir özellik değerinin etkisinin diğer özelliklerin değerlerinden bağımsız olduğunu varsaymaktadır. Bu varsayıma sınıf koşullu bağımsızlık denir. Naive Bayes algoritması koşullu olasılıklara dayanır. Bayes Teoremi, daha önce meydana gelen başka bir olay olasılığı göz önüne alındığında, bir olayın oluşma ihtimalini bulur. Her sınıf değeri için belirli bir örneğin o sınıfa ait olma ihtimalini tahmin ederler.

B, bağımlı olayı temsil ediyorsa ve A, önceki olayı temsil ediyorsa, Bayes teoremi şu şekilde ifade edilebilir.

$$\mathrm{P(B/A) = P(A/B)/P(A)} \tag{3.13}$$

A sı bilinen B'nin olasılığını hesaplamak için, algoritma, A ve B'nin birlikte oluştuğu vakaların sayısını sayar ve A'nın tek başına ortaya çıktığı vakaların sayısıyla böler.

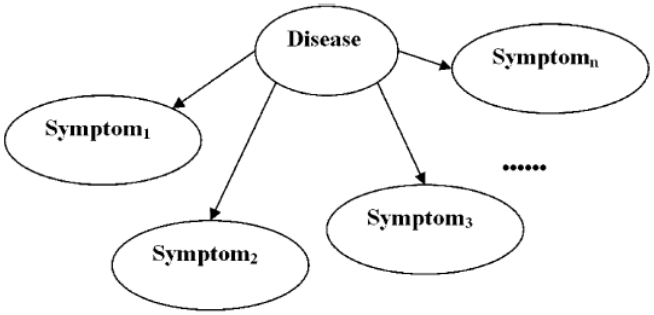
Her sınıfın sonraki olasılığı, $\mathrm{C_i}$, naive bayes sınıflandırıcısı tarafından bayes kuralı ile elde edilir.

$$\mathrm{P(C_i/A_1 \dots\dots, A_n) \;=\; P(C_i)\; P(A_1/\; C_i\;) \dots\dots P(A_n/\; C_i)/P(A)} \tag{3.14}$$

Sınıflandırıcı, sınıf bilgisi verildiğinde A özniteliklerinin bağımsız olduğu varsayımını yapar. Bu nedenle olasılık, sınıfı verilen her niteliğin bireysel koşullu olasılıklarının çarpımı tarafından elde edilebilir.

Naive Bayes sınıflandırıcısının bir avantajı, sınıflandırma için gerekli olan parametreleri (değişkenlerin varyanslarını) tahmin etmek için az miktarda eğitim verisinin yeterli olmasıdır. Hem ikili hem çok sınıflı sınıflandırma problemleri için kullanılabilir [39].

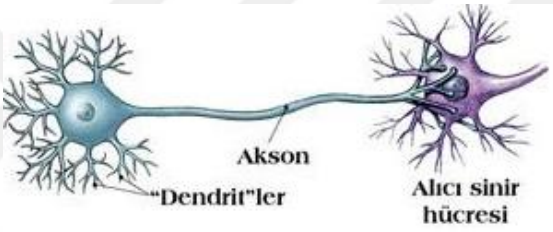
Naive Bayes sınıflandırıcısı, olasılık dağılımlarını grafiksel olarak kısa ve anlaşılabilir bir şekilde temsil eder [43].



Şekil 3. 5 Bir teşhis problemi için Naive Bayes

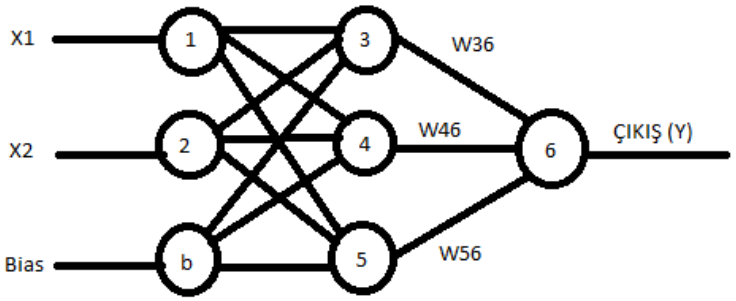
3.6 Yapay Sinir Ağları

Yapay sinir ağları (YSA'lar) beynin ve bilgiyi öğrenme ve işleme biçiminden esinlenen tekniklerdir. Canlılardaki sinir hücreleri ve ağları modellenerek oluşturulmuştur.



Şekil 3. 6 Sinir hücresi

YSA'lar, gerçek dünya uygulamalarında sınıflandırma ve regresyon problemlerini çözmek için sıklıkla kullanılır. Yapay sinir ağları düğümler ve bağlantılardan oluşur. Düğümlerin genellikle sınırlı hesaplama gücü vardır. Tıpkı yeterli nörotransmitter biriktikçe nöronlar aktive edileceği gibi nöronları bir geçiş gibi davranarak simüle eder. Bağlantının yoğunluğu ve karmaşıklığı, sinir ağının hesaplama gücünün gerçek kaynağıdır [44]. Yapay sinir ağları, kendi girdileri aracılığıyla bilgi sinyalleri alan yapay sinir hücrelerinin iki veya daha fazla tabakasından oluşur. Giriş bilgisinin değeri, giriş değişkenlerinin ağırlıklı toplamına sapma (b) değerinin eklenmesi ile hesaplanır.



Şekil 3. 7 Yapay sinir ağı matematiksel gösterimi

$$y = f(b + \sum_i w_i x_i) \quad (3.15)$$

Bağlı nöronların birçok katmanını kullanarak karmaşık karar süreçleri modellenebilir. Nöron ağırlıklarını öğrenmek veya aktivasyon işlevlerini ayarlamak için farklı yaklaşımlar geliştirilmiştir [45].

BÖLÜM 4

MODEL PERFORMANS DEĞERLENDİRME YÖNTEMLERİ

Model değerlendirme yöntemleri, veriyi temsil eden en iyi modeli bulmaya ve seçilen modelin gelecekte ne kadar iyi çalışacağına karar vermeye yardımcı olur. Verinin eğitim ve test seti olarak ayrılmasını eğitimin test setini ne doğru tahmin edeceğini gösteren çeşitli yöntemlerden oluşur.

4.1 Holdout Yöntemi

Holdout yöntemi çapraz doğrulamanın en basit türüdür. Veri seti, eğitim seti ve test seti olarak adlandırılan iki gruba ayrılır. Genellikle test seti için verinin üçte biri, geri kalanlar ise eğitim seti için kullanılır. İşlev yaklaşımcısı, yalnızca eğitim setini kullanarak bir işleve uyar. Ardından, fonksiyon yaklaşımından test setindeki verilerin çıktı değerlerini tahmin etmesi istenir. Holdout yönteminde az sayıda veri seti ile çalışmak test seti sayısını kısıtlayacağından kullanışlı değildir [46].

4.2 K-Kat Çapraz Doğrulama

Holdout yöntemini geliştirmenin bir yoludur. Veri kümesi k alt küme ye bölünür ve kısıtlama yöntemi k kez tekrarlanır. Her defasında, k alt kümelerinden biri test kümesi olarak kullanılırken diğer k-1 alt kümeleri bir eğitim kümesi oluşturmak üzere bir araya getirilir. Ardından, tüm k denemelerindeki ortalama hata hesaplanır Her veri noktası tam olarak bir kez test kümesine girer ve k-1 defa eğitim setine girer [46].

4.3 Sınıflandırma Doğruluğu, Duyarlılık ve Belirleyicilik

Herhangi bir makine öğrenme modelinin etkinliği, kontenjans matrisinden elde edilen doğru oran, yanlış pozitif oran, doğru negatif oran ve yanlış negatif oran, f ölçütü ile saptanmaktadır. Karışıklık matrisi gerçek ve tahmin edilen değerleri gösterir.

Duyarlılık ve özgüllük ölçütleri, klinik tanı testini açıklamak ve tanısal testin ne kadar iyi ve tutarlı olduğunu tahmin etmek için yaygın olarak kullanılmaktadır. Duyarlılık ölçümleri gerçek pozitif oran veya pozitif sınıf doğruluğu, özgüllük ise gerçek negatif oran veya negatif sınıf doğruluğu olarak belirtilir. Doğru olarak sınıflandırılan test setindeki hastaların yüzdesiyle ifade edilir [47].

Çizelge 4. 1 Kontenjans matrisi

Tahmin	**Gerçek**	
	Pozitif	Negatif
Pozitif	a (doğru pozitif) dp	b (yanlış pozitif) yp
Negatif	c (yanlış negatif) yn	d (doğru negatif) dn

- a , bir örneğin pozitif olduğu doğru tahminlerin sayısıdır.
- b, bir örneğin negatif olduğu tahminlerin yanlış sayısıdır.
- c, bir örneğin pozitif olduğu yanlış tahminlerin sayısıdır.
- d, bir örneğin negatif olduğu doğru tahminlerin sayısıdır.

Kesinlik(Precision): Doğru pozizif olarak sınıflandırılan örneklerin tahmin edilen tüm pozitif örneklere oranı olarak ifade edilir.

$$\text{Kesinlik} = \frac{\text{dp}}{\text{dp+yp}} \tag{4.1}$$

Doğruluk(Accuracy): Doğru olarak sınıflandırılan test setindeki hastaların yüzdesiyle ifade edilir.

$$\text{Doğruluk} = \frac{\text{dp+dn}}{\text{toplam}} \tag{4.2}$$

$$\text{Hata oranı (error rate)} = 1 - \text{doğruluk} \tag{4.3}$$

Duyarlılık(Sensitivity): Sınıflandırıcının doğru olarak tahmin ettiği pozitif değerlerin gerçek doğrulara oranıdır. Hasta olanları hasta olarak tahmin ettiği doğruluk oranıdır.

$$\text{Duyarlılık} = \frac{\text{dp}}{\text{dp+yn}} \tag{4.4}$$

Özgüllük (Specificity): Sınıflandırıcının doğru olarak tahmin ettiği negatif değerlerin gerçekte negatif olan durumlara oranıdır. Sağlıklıları sağlıklı olarak tahmin ettiği doğruluk oranıdır.

$$\text{Özgüllük} = \frac{\text{dn}}{\text{dn+yp}} \tag{4.5}$$

Yanlış pozitif (YP): Gerçekte negatif olan ancak sınıflandırıcının pozitif olarak tahmin ettiği değerlerin tüm negatif değerlere oranıdır. Gerçekte hastalığa sahip olmayanların testin yanlışlıkla pozitif sonuç vermesidir.

$$\text{Yp} = 1 - \text{Özgüllük} \tag{4.6}$$

Yanlış negatif (YN): Gerçekte pozitif olan ancak sınıflandırıcının negatif olarak tahmin ettiği değerlerin tüm pozitif değerlere oranıdır. Gerçekte hasta olan bireylerin testin yanlışlıkla negatif sonuç vermesidir.

$$\text{Yn} = 1 - \text{Duyarlılık} \tag{4.7}$$

F Ölçütü: Duyarlılık ve özgüllük ölçütlerinin harmonik ortalamasıdır.

$$\text{F Ölçütü} = \frac{2*\text{Duyarlılık}*\text{Kesinlik}}{\text{Duyarlılık+Kesinlik}} \tag{4.8}$$

4.4 Alıcı İşlem Karakteristikleri (Receiver Operating Characteristic-ROC) Eğrileri:

ROC eğrileri, gerçek pozitif ve yanlış pozitif oranlarını kullanan istatistiksel bir karşılaştırma yöntemidir. Oluşturulan her bir modelin sağlıklılardan hastaları ayırt etme kabiliyetini test etmek için (ayrımcılık), alıcı çalışma karakteristiği eğrisinin (AUC) altındaki alan hesaplanır.

ROC eğrisi altındaki alan (AUC) bir modelin öngörücü kalitesini değerlendirmek için kullanılan standart bir yöntemdir. Farklı sınıflandırıcıların performansını değerlendirir. AUC hem duyarlılığı hem de özgüllüğü dikkate alır ve sınıflar arasında dengesiz olan

veriyle başa çıkmak için nesnel bir yol gösterir. 1'e yakın bir AUC değeri, iyi tahmin niteliğindeki bir modeli tanımlarken 0.5'e yakın bir değer, modelin rasgele bir karardan daha iyi olmadığını gösterir. Sıfıra yakın bir AUC değeri, tüm örnekleri yanlış bir etiketle sınıflandıran bir modeli tanımlar [45].

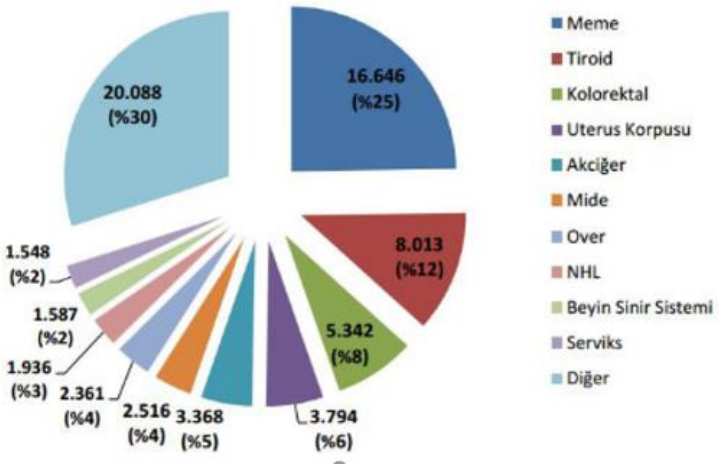
BÖLÜM 5

VERİ YAPISI ve UYGULAMA

Meme kanseri, kadınlarda dünya genelinde en sık rastlanan kanser olup, 2012 yılında yaklaşık 1,7 milyon yeni vaka tespit edilmiştir [48]. Mamografi, meme kanserinin erken teşhisi için tercih edilen yöntemdir. Mamografi, kanseri bulmakta çok hassas olmasına rağmen birçok yanlış pozitif sonuç vermektedir.

Halihazırdaki biyopsi vakalarının yalnızca %20'si kanser ortaya koymaktadır. Son yıllarda meme kanseri teşhisinde sadece biyopsi değil çevresel ve genetik faktörlerin incelenmesi konusunda düşünceler ortaya atılmaktadır.

Dünya'da genel olarak en çok tanı konulan kanserler akciğer (%13,0), meme (%11,9) ve kolon (%9,7) kanseri olarak görülmüştür. Kanserden meydana gelen ölümlerin en çok akciğer (%19,4), karaciğer (%9,1) ve mideden (%8,8) gerçekleştiği belirtilmiştir. Kanser artış hızının devam etmesi durumunda 2025 yılında toplam 19,3 milyon yeni kanser vakası olacağı belirtilmiştir. Uluslararası Kanser Ajansı özellikle meme kanserindeki artışa dikkat çekmiştir. Kadınlarda meme kanseri insidiansının önceki yıllara göre %20 arttığını, meme kanserinden kaynaklanan ölümlerin ise %14 arttığını belirtmiştir. Kanser teşhisi konulan her 4 kadından 1 inin meme kanseri olduğu görülmüştür [49].



Şekil 5. 1 2014 yılı meme kanseri istatistikleri

2015 yılı şubat ayı, Dünya Sağlık Örgütü'nün (DSÖ) resmi internet sitesinde yayınlanan verilere göre, kadınlarda ise en sık rastlanan kanser türünün meme kanseri olduğu ifade edilirken, meme kanserini sırasıyla kalın bağırsak, akciğer, rahim ağzı ve mide kanser türleri takip etmektedir.

Çizelge 5. 1 Uluslararası Kanser Ajansı (IARC) Tarafından Yayınlanan Verilere Göre Kadınlarda En Sık Görülen İlk Beş Kanser Türünün Dağılımı

	Türkiye*	**Dünya**	**IARC'a üye 24**	**AB (28 ülke)**	**ABD**
1	Meme	Meme	Meme	Meme	Meme
2	Tiroid	Kolorektal	Kolorektal	Kolorektal	Akciğer
3	Kolorektal	Uterus serviksi	Akciğer	Akciğer	Kolorektal
4	Uterus korpusu	Akciğer	Uterus serviksi	Uterus korpusu	Tiroid
5	Akciğer	Uterus korpusu	Uterus korpusu	Uterus serviksi	Uterus

* Türkiye Birleşik Veri Tabanı, 2014

Amerikan Kanser Derneği, insidians, ölüm, hayatta kalma ve tarama verileri dahil olmak üzere Birleşik Devletler 'deki kadın meme kanseri istatistiklerine genel bir bakış sunmaktadır. 2015 yılında ABD'li kadınlarda meme kanseri insidians oranları hispanik olmayan siyah (siyah) ve Asya / Pasifik Adalı kadınlarda artmıştır. 2008'den 2012'ye kadar beyaz (beyaz), Hispanik ve Amerikalı Hintli / Alaska Yerli kadınlarda meme kanseri görülme oranı beyaz kadınlarda fazla olmasına rağmen 2012 de siyah kadınlarda görülme oranı ile aynı seviyeye gelmiştir [50].

2016’daki Dünya Sağlık Kongresinde diyet ve yaşam biçiminin meme kanseri riskini ve sağkalımı nasıl etkilediğine ilişkin durumlar tartışılmıştır. Dünya Kanser Araştırma Fonu'nun Sürekli Güncelleme Projesindeki dünyadaki araştırmalarının analizi, ağırlığın, alkol alımının ve fiziksel aktivitenin meme kanseri riskini etkileyebileceğini ortaya çıkarmıştır. Birleşik Krallık'ta yapılan bir çalışma her 5 meme kanseri vakasının yaklaşık 2'sinin sağlıklı bir kiloyu korumak, fiziksel olarak aktif olmak ve alkol almamak suretiyle önlenebileceğini belirtmiştir. Bu kesinlikle, çevrenin ve beslenmenin kanser riskinin arttırabileceğini düşündürmektedir.

Oturum Başkanlığı, İskoç Kanseri Önleme Ağı'ndan Annie Anderson, "Meme kanserini önlemek için farkındalık yaratmanın, kilo yönetimini desteklemenin, fiziksel aktiviteyi artırmanın ve alkol alımını azaltmanın gerekli olduğu" yorumunu yapmıştır. Tartışma daha sonra meme kanseri için genetik risk faktörlerini araştırmıştır. Breast N Now, nesiller üzerinde 40 yıl boyunca takip edilen 100.000'den fazla kadının kanının incelendiğini ve genetik altyapısını değerlendirmek ve genlerinde meydana gelen değişikliklere bakmak için alınan kanları finanse etmiştir. Bu çalışma, genlerin metabolizmayı ve meme kanseri riskini nasıl etkilediği konusunda değerli bilgiler sağlamıştır.

Gelecekte, yaşam tarzı faktörlerinin ve genetik faktörlerin nasıl bir etkileşime girdiğine ilişkin daha fazla araştırma yapmak istenmektedir. Böylece meme kanserine neden olan değişkenler daha doğru belirlenir ve riski tahmin etmede daha doğru sonuçlara varılabilir.

Yaptığımız bu çalışma ise meme kanseri riskinin kişinin fiziksel ve kültürel özelliklerden ne kadar etkilendiğini hangi değişkenlerin bu riski arttırdığını bize gösterecektir.

5.1 Veri Seti

Kullanılan veri seti Amerika’da Meme Kanseri Gözlem Konsorsiyuma gelen kadınlardan alınan 280.660 adet tarama mamogramı ("indeks mamografisi" olarak anılacaktır) kaydını içermektedir. Bu veriler 1996-2002 yılları arasında kaydedilmiştir. Tüm kadınlar daha önce meme kanseri tanısı almamışlardır. Kanser kayıtları ve patoloji verileri mamografi verileriyle bağlantılıdır.

Mevcut araştırmanın Meme Kanseri Gözlem Konsorsiyumunda korunması da gereklidir. Bu sebeplerden dolayı, veriler aşağıdaki değişkenlerle sınırlıdır.

280.660 kayıt bulunan veri setinde eksik değerler çıkarılmış ve kullanılabilir 19.715 kayıt elde edilmiştir. 19.715 kayıttan 2146(%10,88) i hastalıklı grup iken 17.569(%90,13) si sağlıklı grubu oluşturmaktadır. Bu oran sağlıklı sonuçlar vermeyeceği için veri seti yarısı kanser yarısı kanser olmayan olarak bölünmüş ve 847 (%49,88) si sağlıklı 851(%50,12) i meme kanseri hastası olmak üzere 1697 kayıta indirgenmiştir.

Veri setinde 12 değişken kullanılmaktadır ve değişkenlerin hepsi kategoriktir.

Çizelge 5. 2 Değişken tanımları

Yaş grupları (agegrp)	0=35-39; 1=40-44;2=45-49;3=50-54;4=55-59;5=60-64;6=65-69;7=70-74;8=70-74;9=75-79
Yoğunluk (density)	1 = Hemen hemen tamamen yağlı; 2 = Dağılmış fibroglandüler yoğunluklar; 3 = Heterojen yoğun; 4 = Aşırı yoğun 5=Diğer
Irk (race)	1=Beyaz; 2=Asya/Pasifik Adalı; 3=Siyah; 4=Amerikalı;5= Diğer
İspanyol köken (Hispanic)	0=İspanyol kökenli değil ;1=İspanyol kökenli
Beden kütle indeksi (bmi)	1=10-24;2=25-29;3=30-34;4=35+
İlk yaş (Agefirst)	İlk doğum yaşı 0=<30;1=>30;2=Hiç doğum yapmamış
Ailede geçmiş kanser öyküsü (nrelbc)	0=Ailede kanser geçiren kişi yok;1=Ailede 1 kişi kanser geçirmiş;2=Ailede 2 ya da daha fazla kişi kanser geçirmiş
Önceki meme işlemi (brstproc)	0=Yok; 1=Var
Son mamografi (lastmamm)	İndeks mamografiden önce son mamografi sonucunun negatif (0) ya da yanlış poziti (1) olma durum
Cerrahi menepoz (surgmeno)	0=Doğal;1= Cerrahi
Hormon tedavisi (hrt)	0=Yok; 1=Var
Kanser (N)	İndeks taramasından bir yıl sonra invaziv veya duktal karsinom meme kanseri tanısı 0=Kanser yok;1=Kanser

(http://www.bcsc-research.org/rfdataset/dataset.html) veri seti halka açıktır.

5.2 Uygulama

Çalışmada Naive Bayes, lineer destek vektör makineleri (DVM), lojistik regresyon, radyal tabanlı yapay sinir ağları, k en yakın komşuluk ve karar ağacı algoritmalarından C.5.0 olmak üzere 6 sınıflandırma yöntemi kullanılmıştır.

Uygulama iki bölümden oluşmaktadır. İlk uygulamada veri setinin %80'ni eğitim ve %20'si test seti olarak kullanılmış ve SPSS.18 Moduler ile yapılmıştır. İkinci uygulamada ise veri setinde k kat çapraz doğrulama kullanılmış, farklı k değerleri için sonuçlar karşılaştırılmış ve bu analizler WEKA ile yapılmıştır.

Modellerin tanısal performansını göstermek için sınıflandırma doğruluğu, ROC eğrisi altındaki alan ve kontenjans matrisinden elde edilen duyarlılık, özgüllük, hata oranı ölçütleri kullanılmıştır.

5.2.1 Eğitim ve Test Seti Kullanılarak Yapılan Uygulama Sonuçlar

5.2.1.1 Lojistik Regresyon Yöntemi ile Elde Edilen Sonuçlar

Veri seti test ve eğitim seti olarak ayrılmıştır. Eğitim veri seti katılımcıların %80'ini (1697'den 1347'si), test veri seti %20 (351) sini oluşturmaktadır.

Çizelge 5. 3 Lojistik regresyon değişkenlerin modeldeki anlamlılığı

N[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
1	Intercept	-10.861	.714	231.266	1	.000	
	[density=1]	0[b]	.	.	0	.	.
	[density=2]	-.528	.272	3.779	1	.052	.590
	[density=3]	.907	.207	19.164	1	.000	2.478
	[density=4]	1.133	.203	31.136	1	.000	3.106
	[race=1]	0[b]	.	.	0	.	.
	[race=2]	2.345	.326	51.898	1	.000	10.436
	[race=3]	.428	.384	1.247	1	.264	1.535
	[race=4]	-.013	.414	.001	1	.974	.987
	[race=5]	-1.195	.619	3.730	1	.053	.303
	[hispanic=0]	0[b]	.	.	0	.	.
	[hispanic=1]	2.448	.261	88.293	1	.000	11.568
	[bmi=1]	0[b]	.	.	0	.	.

	[bmi=2]	.581	.197	8.675	1	.003	1.789
	[bmi=3]	.652	.197	10.914	1	.001	1.920
	[bmi=4]	.204	.210	.944	1	.331	1.226
	[agefirst=0]	0[b]	.	.	0	.	.
	[agefirst=1]	-.273	.190	2.063	1	.151	.761
	[agefirst=2]	1.052	.148	50.419	1	.000	2.864
	[nrelbc=0]	0[b]	.	.	0	.	.
	[nrelbc=1]	1.253	.275	20.798	1	.000	3.499
	[nrelbc=2]	2.089	.266	61.561	1	.000	8.078
	[brstproc=0]	0[b]	.	.	0	.	.
	[brstproc=1]	.358	.124	8.345	1	.004	1.430
	[lastmam=0]						
	[lastmam=1]	2.655	.296	80.700	1	.000	14.227
	[surgmeno=0]	0[b]	.	.	0	.	.
	[surgmeno=1]	.580	.126	21.318	1	.000	1.786

Meme yoğunluğu aşırı yoğun (4) olan kişilerin meme kanserine yakalanma olasılığı yoğunluk derecesi hemen hemen tamamen yağlı (1) kişilere göre 3,1 kat daha fazladır. Heterojen meme yoğunluğuna (3) sahip kişilerin meme kanserine yakalanma olasılığı yoğunluk derecesi hemen hemen tamamen yağlı olan kişilere göre 2,4 kat daha fazladır. Meme yoğunluğu dağılmış fibroglandüler olan kişilerin, yoğunluk derecesi hemen hemen tamamen yağlı olan kişilere göre meme kanserine yakalanma olasılığı ise yaklaşık 0,6 kat daha fazladır. Irkı Asya/Pasifik Adalı olanların diğer ırklara göre kanser riski 10,4 kat fazladır.

Hiç doğum yapmamış kadınların ilk doğum yaşı 30'dan küçük olan kadınlara göre meme kanserine yakalanma riski yaklaşık 3,4 kat daha fazladır. İspanyol kökenli olan kişilerin, İspanyol kökenli olmayan kişilere göre meme kanserine yakalanma olasılığı 11,5 kat daha fazladır. Ailede kanser geçirmeyen kişilerin ailede kanser geçiren 2 kişi ya da daha fazla kişi olmasına göre meme kanserine yakalanma olasılığı 8 kat daha fazladır. Önceden meme işlemi geçirmemiş kişilerin meme işlemi geçiren kişilere göre kansere yakalanma riski 14 kat fazladır. Son mamografi sonucu yanlış pozitif (1) çıkan kişilerin negatif (0) çıkan kişilere göre kanser riski 14 kat fazladır. Cerrahi menapozu cerrahi yolla geçiren kadınların doğal geçiren kadınlara göre kanser riski 1.7 kat daha fazladır.

Çizelge 5. 4 Lojistik regresyon test ve eğitim seti hata matrisi

	Eğitim		Test	
Doğru	1054	78,31%	269	76,64%
Yanlış	292	21,69%	82	23,36%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde sağlıklı kişileri sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %78,32 modelin hata yapma yüzdesi, %21,68 olarak görülmüştür.

Test setinde sağlıklı kişilere sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %76,64 modelin hata yapma yüzdesi %23,36 olarak görülmüştür.

Çizelge 5. 5 Lojistik regresyon kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	640	207	847
	%	75,56	24,43	100
Hasta	N	167	683	850
	%	19,62	80,35	100
Toplam	N	807	890	1697
	%	47,55	52,44	100

Lojistik regresyon hastaları tahmin etmede %80,35 başarı gösterirken sağlıklı bireyleri tahmin etmede %75,56 başarı göstermiştir.

Çizelge 5. 6 Lojistik regresyon model performans değerlendirme ölçütleri

Doğruluk	0,78
Duyarlılık	0,77
Hata oranı	0,22
Yanlış negatif oranı	0,23
Özgüllük	0,79
Yanlış pozitif oranı	0,21

Modelin genel anlamda doğru sınıflandırma (hasta bireye hasta ve sağlıklı bireye sağlıklı deme) oranı 0,78, modelin hata oranı (hasta bireye sağlıklı ve sağlıklı bireye hasta deme)

0,22, duyarlılık (hastaları doğru tahmin etme) oranı 0,77, özgüllük (sağlıklı bireyleri tahmin etme) oranı 0,79 bulunmuştur.

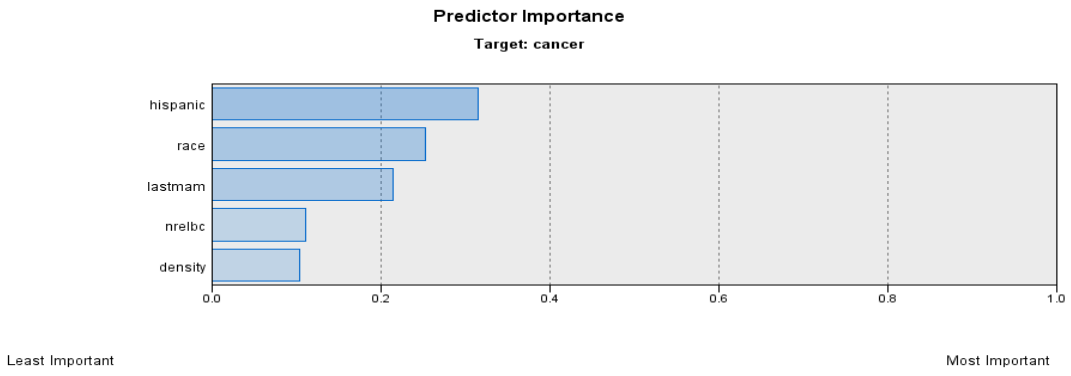
Çizelge 5. 7 Lojistik Regresyon ROC eğrisi altında kalan alan (AUC)

	Eğitim	Test
Model	AUC	AUC
Kanser	0,841	0,836

ROC eğrisi altındaki alan eğitim setinde 0,84, test setinde 0,83 bulunmuştur, bu sonuçlar modelin tanısal performansının iyi olduğunu göstermektedir.

5.2.1.2 C5.0 Algoritması Sonuçları

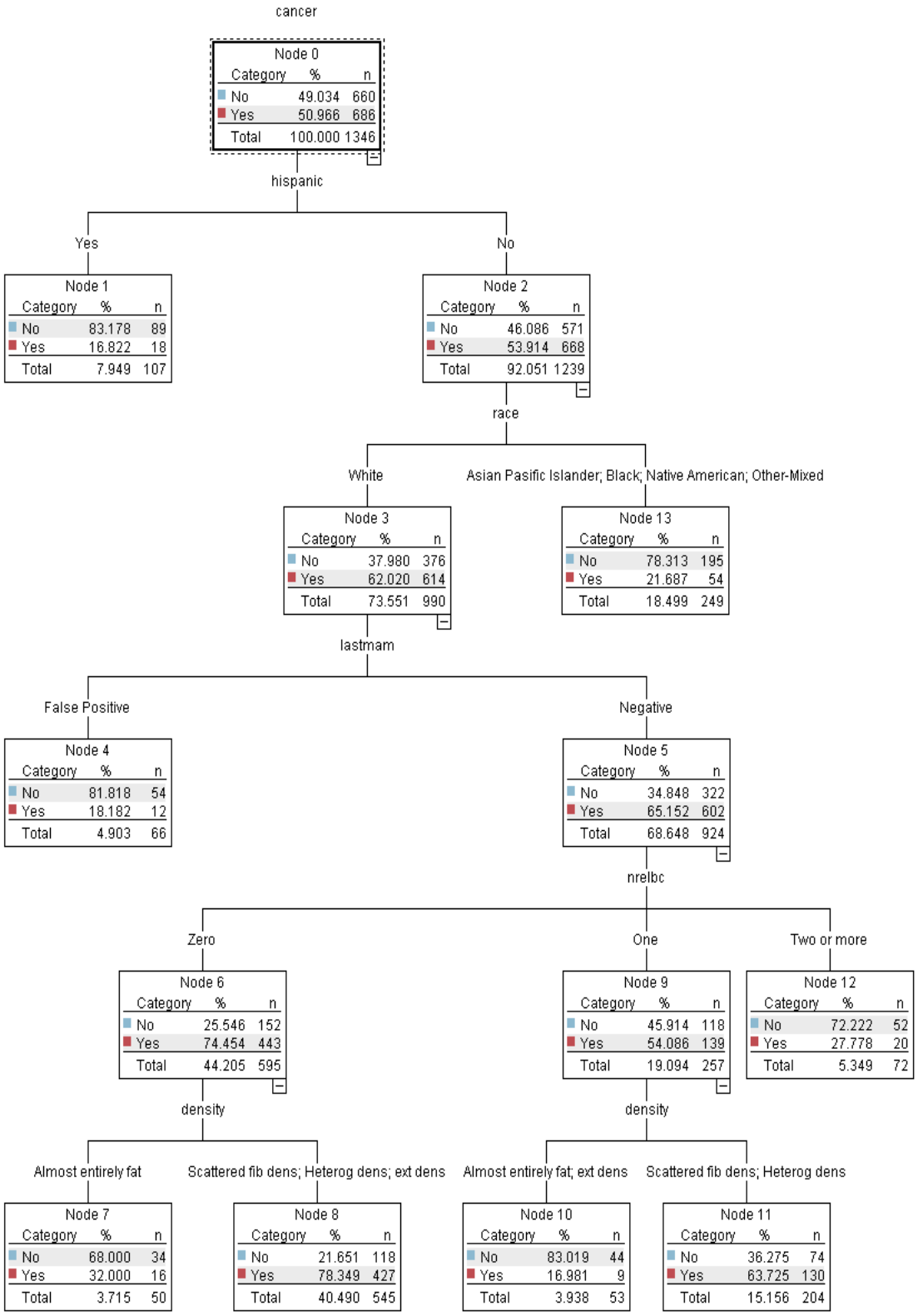
Veri seti test ve eğitim seti olarak ayrılmıştır. Eğitim veri seti katılımcıların %80'ini (1697'den 1347'si), test veri seti %20 (351) sini oluşturmaktadır.



Şekil 5. 2 C5.0 algoritması değişken tahmincisi önem göstergesi

C.5.0 algoritması 11 değişkenden 5 değişkeni anlamlı bulmuştur. En etkili değişkenin İspanyol kökenli olup olmama durumu olduğu, daha sonra gelen değişkenlerin ırk, önceki mamografi sonucu, ailesinde kanser öyküsü olup olmama durumu ve meme yoğunluğu değişkenleri olmuştur.

Aşağıdaki karar ağacı şeklinde hangi değişkenlerin hangi gruplarının etkili olduğu göstermektedir.



Şekil 5. 3 C5.0 algoritması karar ağacı

Veri setinin tamamında 1346 kişi bulunmaktadır. Bu kişilerin 686'sı (%50,9) kanserli, 660'ı (%49.03) sağlıklı kişilerden oluşmaktadır. Karar ağacında bakıldığında C5.0 algoritması kanser olmakta en etkili değişkeni İspanyol kökenli olup olmama durumu olarak bulmuştur. İspanyol kökenli olan kişilerin meme kanserine yakalanma olasılığı %17 olarak bulunmuştur. İspanyol kökenli olmayan kişilerde meme kanseri görülme olasılığı %53'e yükselmiştir ve ağaç buradan kırılmaya devam etmiştir. Bu kişilerde kansere yakalanmada en etkili değişkeni ırk olarak bulmuştur. Irkı beyaz olan kişilerde kansere yakalanma olasılığı %62 iken, diğer ırklarda kansere yakalanma olasılığı %22'dir. Beyaz olan kişilerde kansere yakalanmada en etkili değişkenin son mamografi sonucu olduğu görülmektedir. Yanlış pozitif sonucu çıkan kişilerde kansere yakalanma olasılığı %19'a düşer.

Karar ağacına genel bir bakıldığında; örneğin, meme yoğunluğu 'dağılmış fibroglandüler yoğunluk' ve 'heterojen yoğun' olan kişilerde, aile öyküsünde kansere sahip bulunan kişi sayısı 1 olanlarda, son mamografi sonucu negatif çıkan İspanyol kökenli olmayan beyaz kadınlarda meme kanserine yakalanma olasılığı %63 olarak bulunmuştur.

Çizelge 5. 8 C5.0 algoritması eğitim ve test seti hata matrisi

	Eğitim		Test	
Doğru	1025	76,15%	267	76,07%
Yanlış	321	23,85%	84	23,93%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde sağlıklı kişilere sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %76,15, modelin hata yapma yüzdesi %23,85'tir.

Test setinde sağlıklı kişilere sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %76,07, modelin hata yapma yüzdesi %23,93 tür.

Çizelge 5. 9 C5.0 algoritması kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	604	243	847
	%	71,31	28,689	100
Hasta	N	162	688	850
	%	19,059	80,941	100
Toplam	N	766	931	1697
	%	45,138	54,862	100

Model test ve eğitim seti olmaksızın sağlıklı kişilerin %71,31'ini, hasta olan kişilerin de %80,94 ünü doğru tahmin etmiştir.

Çizelge 5. 10 C5.0 algoritması model performans değerlendirme ölçütleri

Doğruluk	0,76
Duyarlılık	0,74
Hata oranı	0,24
Yanlış negatif oranı	0,26
Özgüllük	0,79
Yanlış pozitif oranı	0,21

Modelin genel anlamda doğru sınıflandırma oranı 0,76, modelin hata oranı 0,24, duyarlılık (hastaları doğru tahmin etme) oranı 0,74, özgüllük (sağlıklı kişileri tahmin etme) oranı 0,79 bulunmuştur.

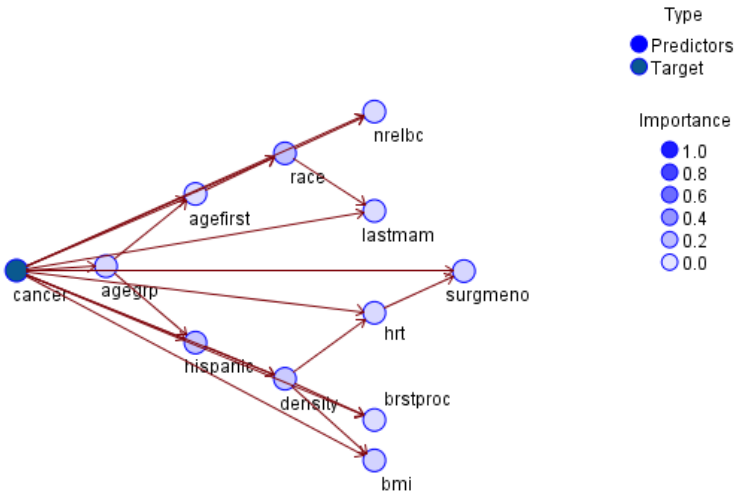
Çizelge 5. 11 C5.0 algoritması ROC eğrisi altında kalan alan (AUC)

	Eğitim	Test
Model	AUC	AUC
Kanser	0,788	0,79

C5.0 algoritmasının model performansı ROC eğrisi altında kalan alan AUC ile hesaplanmıştır. AUC'un altında kalan alanın eğitim setinde 0,78 ve test setinde 0,79 olduğu görülmüştür. Model performansının iyi olduğu söylenebilir.

5.2.1.3 Naive Bayes Yöntemi Sonuçları

Veri seti test ve eğitim seti olarak ayrılmıştır. Eğitim veri seti katılımcıların %80'ini (1697'den 1347'si), test veri seti %20 (351) sini oluşturmaktadır.



Şekil 5. 4 Bayesyan ağ değişken önem düzeyleri

Şekil 5.4'de bayes ağı modelinde kanser riskine etki eden en önemli değişkenler sırasıyla görülmektedir. Kanser riski için en iyi tahminci yaş grup değişkeni iken sonraki değişken İspanyol kökenli olma durumu ve onu takip eden değişkenler ilk doğum yaşı, ırk, meme yoğunluğudur.

Çizelge 5. 12 Bayes ağlarının eğitim ve test seti hata matrisi

	Eğitim		Test	
Doğru	1052	78,16%	273	77,78%
Yanlış	294	21,68%	78	22,22%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde sağlıklı kişilere sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %78,16, modelin hata yapma yüzdesi %21,68'dir.

Test veri setinde modelin doğruluk yüzdesi %77,78, modelin hata yapma yüzdesi %22,22'dir.

Çizelge 5. 13 Bayes ağları kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	637	210	847
	%	75,21	24,79	100
Hasta	N	162	688	850
	%	19,059	80,941	100
Toplam	N	799	898	1697
	%	47,083	52,917	100

Model test ve eğitim seti olmaksızın sağlıklı kişilerin %75,21'sını, hasta olan kişilerin de %80,95'ini doğru tahmin etmiştir.

Çizelge 5. 14 Bayes ağları model performans değerlendirme ölçütleri

Doğruluk	0,78
Duyarlılık	0,77
Hata oranı	0,22
Yanlış negatif oranı	0,23
Özgüllük	0,80
Yanlış pozitif oranı	0,20

Modelin genel anlamda doğru sınıflandırma oranı 0,78, modelin hata oran oranı 0,22, duyarlılık (hastaları doğru tahmin etme) oranı 0,77, özgüllük (sağlıklı bireyleri tahmin etme) oranı 0,80 bulunmuştur.

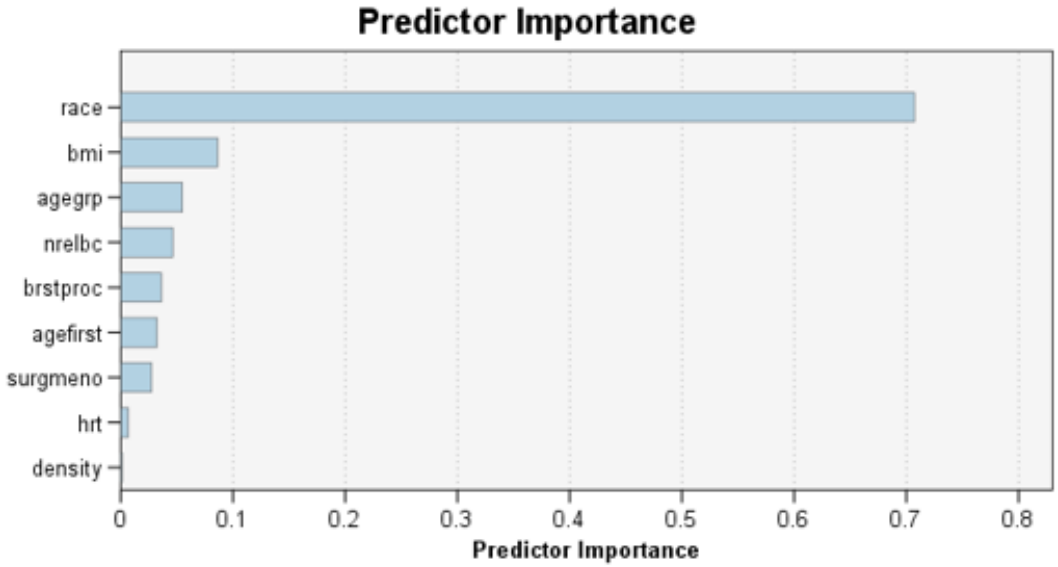
Çizelge 5. 15 Bayesyan ağ ROC eğrisinin altındaki alan (AUC)

	Eğitim	Test
Model	AUC	AUC
Kanser	0,865	0,822

Bayes ağlarında model performansı ROC eğrisi altında kalan alan AUC ile hesaplanmıştır. AUC'un altında kalan alanın 0,822 olduğu görülmüştür. Model performansının oldukça iyi olduğu söylenebilir.

5.2.1.4 Destek Vektör Makineleri Yöntemi Sonuçları

Destek vektör makineleri yönteminde lineer yöntem kullanılmıştır. Veri setinin %80 i eğitim ve %20 si test olmak üzere ayrılmıştır.



Şekil 5. 5 DVM değişken tahmini önem düzeyi

11 değişkenden en önemlileri sırasıyla ırk, beden kütle indeksi, yaş grupları, ailede kanser öyküsü, önceki meme işlemi, ilk doğum yaşı, cerrahi menopoz, hormon tedavisi ve meme yoğunluğudur.

Çizelge 5. 16 DVM eğitim ve test seti hata matrisi

Partition	Eğitim		Test	
Doğru	1018	75,63%	265	75,5%
Yanlış	328	24,37%	86	24,5%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde sağlıklı kişilere sağlıklı, hasta kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %75,63, modelin hata yapma oranı %24,37 dir.

Test setinde modelin doğruluk yüzdesi %75,5, modelin hata yapma yüzdesi %24,5 tir.

Çizelge 5. 17 DVM kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	579	268	847
	%	68,359	31,641	100
Hasta	N	146	704	851
	%	17,156	82,824	100
Toplam	N	725	972	1697
	%	42,722	57,278	100

Model test ve eğitim seti olmaksızın hasta olmayan kişilerin %68,35 ini, hasta olan kişilerin de %82,82 sini doğru tahmin etmiştir.

Çizelge 5. 18 DVM model performans değerlendirme ölçütleri

Doğruluk	0,76
Duyarlılık	0,72
Hata oranı	0,24
Yanlış negatif oranı	0,28
Özgüllük	0,80
Yanlış pozitif oranı	0,20

Modelin genel anlamda doğru sınıflandırma (hastaya hasta ve hasta olmayana hasta değil deme) oranı 0,76, modelin hata oranı (hasta ya hasta değil ve hasta değile hasta deme) 0,24, duyarlılık yani hastaları doğru tahmin etme oranı 0,72, özgüllük yani sağlıklı bireyleri tahmin etme oranı 0,80 bulunmuştur.

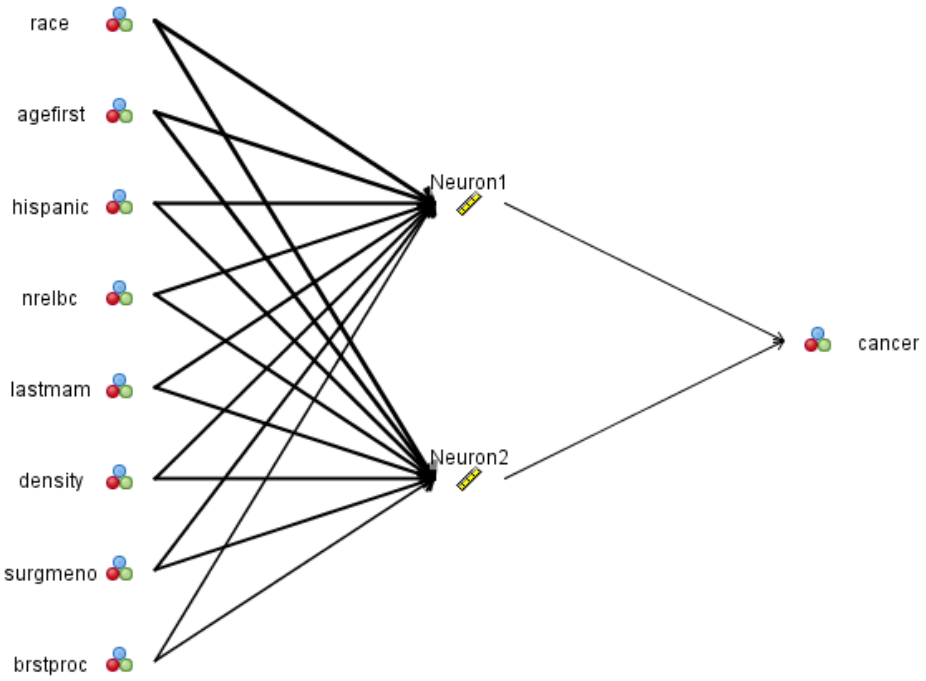
Çizelge 5. 19 DVM ROC eğrisi altında kalan alan (AUC) değerlendirmesi

	Eğitim	Test
Model	AUC	AUC
Kanser	0,843	0,829

Model değerlendirme performansı olarak AUC'a bakıldığında eğitim setinin 0,843 ve test setinin 0,829 olduğu görülmektedir. Model performansı oldukça iyidir.

5.2.1.5 Yapay Sinir Ağları ile İlgili Bulgular

Yapay sinir ağında radyal tabanlı sinir ağı kullanılmıştır. Veri seti %80 eğitim ve %20 test seti olarak ayrılmıştır.



Şekil 5. 6 Yapay sinir ağları

Yukarıdaki şekil değişkenlerin önem düzeylerini göstermektedir. Kalın ve koyu renkliler daha etkili değişken olarak belirtilirken ince çizgiler de önem düzeyi azalan değişkenler olarak belirtilmektedir. 11 değişkenden en önemlileri ırk, ilk doğum yaşı, İspanyol kökenli olup olmama durumu, ailede meme kanseri öyküsü, son mamografiden önceki tarama sonucu, meme yoğunluğu, beden kütle indeksi, ailede kanser öyküsü, son mamografiden önceki tarama sonucu, meme yoğunluğu cerrahi menapoz ve önceki meme işlemi olmak üzere 8'i anlamlı bulunmuştur.

Çizelge 5. 20 YSA eğitim ve test seti hata matrisi

	Eğitim		Test	
Doğru	1011	75,11%	253	72,08%
Yanlış	335	24,89%	98	27,92%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde sağlıklı kişilere sağlıklı, hasta olan kişilere de hasta deme oranı yani modelin doğruluk yüzdesi %75,11, modelin hata yapma yüzdesi %24,89 dur.

Test setinde modelin doğruluk yüzdesi %72,08, modelin hata yapma yüzdesi %27,92 dir.

Çizelge 5. 21 YSA kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	649	198	847
	%	76,623	23,377	100
Hasta	N	235	615	851
	%	27,647	72,353	100
Toplam	N	885	813	1697
	%	52,092	47,908	100

Model test ve eğitim seti olmaksızın hasta olmayan kişilerin %76,62 sini hasta olan kişilerin de %72,35 ini doğru tahmin etmiştir.

Çizelge 5. 22 YSA ROC eğirişi altındaki alan (AUC) değerlendirmesi

	Eğitim	Test
Model	AUC	AUC
Kanser	0,81	0,791

Yapay sinir ağı algoritmasının model performansı ROC eğrisi altında kalan alan (AUC) ile hesaplanmıştır.

AUC'un altında kalan alanın eğitim setinde 0,81 ve test setinde 0,791 olduğu görülmüştür. Model performansının iyi olduğu söylenebilir.

Çizelge 5. 23 YSA Model performans değerlendirme ölçütleri

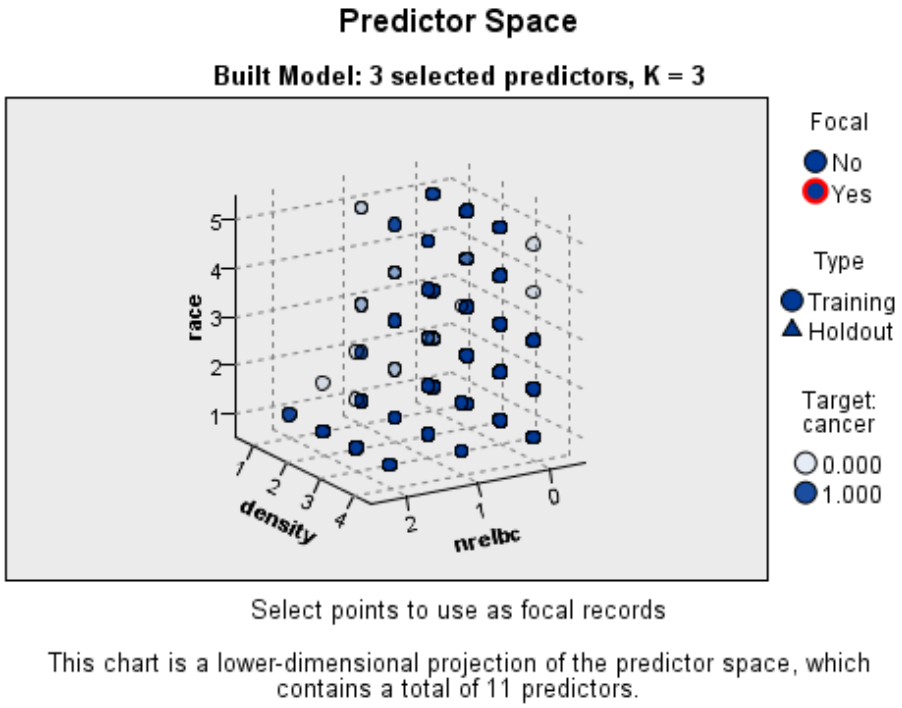
Doğruluk	0,74
Duyarlılık	0,76
Hata oranı	0,26
Yanlış negatif oranı	0,24
Özgüllük	0,73
Yanlış pozitif oranı(1-Özgüllük)	0,27

Modelin genel anlamda doğru sınıflandırma (hastaya hasta ve sağlıklıya sağlıklı deme) oranı 0,74, modelin hata oranı oranı (hasta ya sağlılı ve sağlıklıya hasta deme) 0,26, duyarlılık yani hastaları doğru tahmin etme oranı 0,76, özgüllük yani sağlıklıları tahmin etme oranı 0,73 bulunmuştur.

5.2.1.6 k En Yakın Komşu Algoritması ile İlgili Bulgular

Veri seti test ve eğitim seti olarak ayrılmıştır. Eğitim veri seti katılımcıların %80'inini (1697'den 1347'si), test veri seti %20 (351) sini oluşturmaktadır.

k-NN yönteminde en iyi sonucu k nın 3 olduğu durum vermiştir.



Şekil 5. 7 k-NN boyutsal gösterimi

Şekil 5.7'de en etkili değişkenlerin ırk, meme yoğunluğu ve ailede kanser öyküsü olduğu görülmektedir.

Çizelge 5. 24 k-NN eğitim ve test seti hata matrisi

	Eğitim		Test	
Doğru	1117	83%	217	61,82%
Yanlış	229	17%	134	38,18%
Toplam	1346		351	

Tabloda verinin test ve eğitim seti olarak ayrılmış durumu görülmektedir. Eğitim setinde modelin doğruluk yüzdesi %83, modelin hata yapma yüzdesi %17'dir.

Test modelin doğruluk yüzdesi %61,82, modelin hata yapma yüzdesi %38,18'dir. Eğitim ve test setindeki doğruluk yüzdesi arasındaki fark çok büyük olduğu için bu model aşırı öğrenmiştir. Diğer modeller ile karşılaştırma yapmak gerçeği yansıtmayacaktır.

Çizelge 5. 25 k-NN kontenjans tablosu

Kanser		Sağlıklı	Hasta	Toplam
Sağlıklı	N	605	242	847
	%	71,429	28,571	100
Hasta	N	121	729	850
	%	14,781	85,765	100
Toplam	N	726	971	1697
	%	42,781	57,219	100

Model test ve eğitim seti olmaksızın hasta olmayan kişilerin %71,42 sini, hasta olan kişilerin de %85,76 sını doğru tahmin etmiştir

Çizelge 5. 26 k-NN model performans değerlendirme ölçütleri

Doğruluk	0,79
Duyarlılık	0,75
Hata oranı	0,21
Yanlış negatif oranı	0,25
Özgüllük	0,83
Yanlış pozitif oranı	0,17

Modelin genel anlamda doğru sınıflandırma oranı 0,79, modelin hata oranı 0,21, duyarlılık yani hastaları doğru tahmin etme oranı 0,75, özgüllük yani hasta olmayanları tahmin etme oranı 0,83 bulunmuştur.

Çizelge 5. 27 k-NN ROC eğrisi altındaki alan (AUC)

Partition	Eğitim	Test
Model	AUC	AUC
Kanser	0,916	0,686

Model perfromans değerlendirmesi için AUC değerine bakıldığında eğitim setinde 0,916 test setinde 0,686 olduğu görülmüştür. Model aşırı öğrendiği için test setindeki değerlendirmesi iyi bulunmamıştır.

5.2.2 Çapraz Doğrulama Kullanılarak Elde Edilen Sonuçlar

5.2.2.1 Lojistik Regresyon Yöntemi ile Elde Edilen Sonuçlar

Lojistik regresyonun tanısal performansını göstermek için; sınıflandırma doğruluğu, ROC eğrisi altındaki alan, 3,5 ve 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 5. 28 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	%73,07 (1240)	%73,30 (1244)	%73,07 (1240)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1244 örnek ile k nın 5 olduğu durumda ortaya çıktığı görülmektedir. Model 5 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %73,30 başarı göstermiştir.

Çizelge 5. 29 Farklı k değerleri için model performans değerlendirme ölçütleri

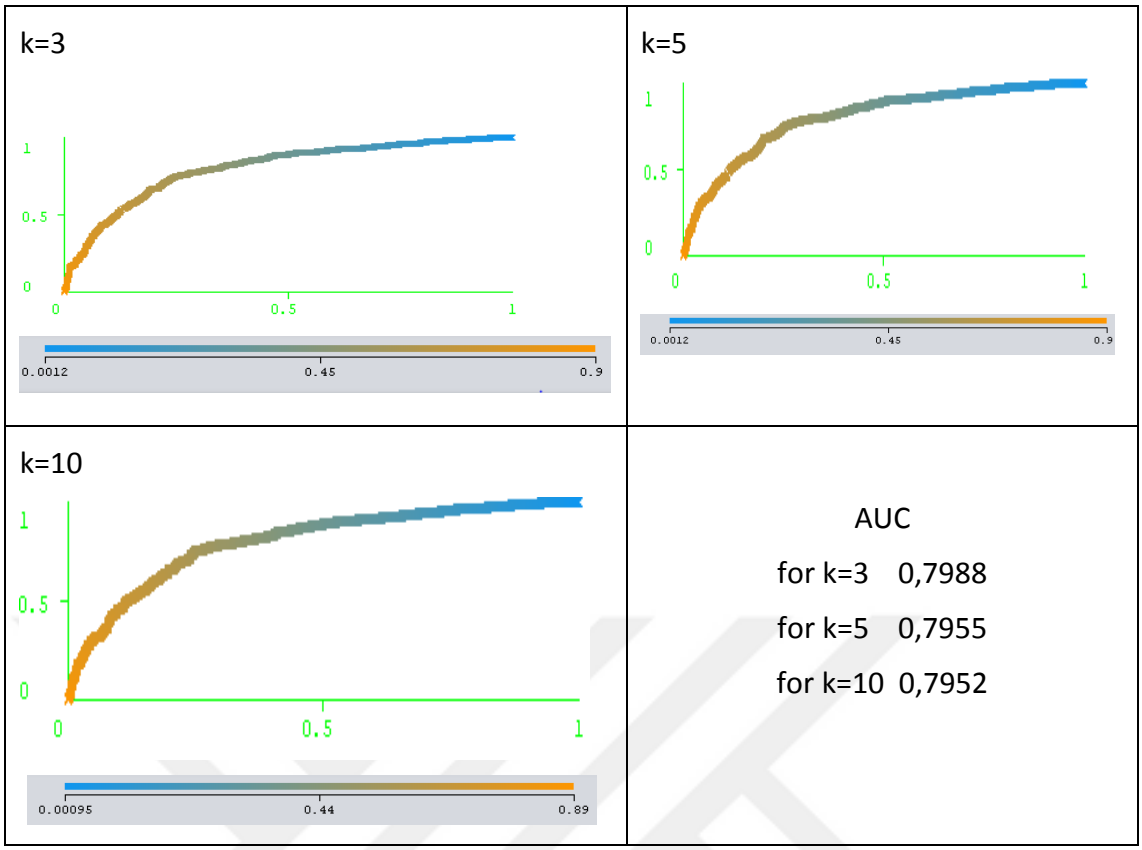
k Kat Çapraz Doğrulama	DP Oranı	YP Oranı	Kesinlik	F-Ölçütü	Sınıf
k=3	0,671	0,209	0,761	0,713	Sağlıklı
	0,791	0,329	0,707	0,746	Hasta
Ağırlıklı Ort	0,731	0,270	0,734	0,730	
k=5	0,677	0,211	0,762	0,717	Sağlıklı
	0,789	0,323	0,710	0,748	Hasta
Ağırlıklı Ort	0,733	0,267	0,736	0,732	
k=10	0,675	0,214	0,759	0,715	Sağlıklı
	0,786	0,325	0,708	0,745	Hasta
Ağırlıklı Ort	0,731	0,269	0,733	0,730	

k'nın 5 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. Doğru pozitif oranı yani modelin hastaları doğru olarak tahmin ettiği değer 0,733 olarak görülmektedir. Yanlış pozitif oranı ise sağlıklı bireyleri hasta olarak tahmin ettiği değerdir. Bu değer en düşük olarak yine k'nın 5 olduğu modelde 0,267 olarak görülmektedir.

Çizelge 5. 30 Farklı k değerleri için kontenjans tablosu

N=1697	**k=3**		**k=5**		**k=10**		**Toplam**
	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	
Sağlıklı	568	279	573	274	572	275	847
Hasta	178	672	179	671	182	668	850

Çizelge 5.30'da farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 5 kat olduğu durumda görülmektedir. k'nın 5 olduğu durumda model 847 hastanın 573 ünü doğru tahmin ederken, 850 sağlıklı bireyin 671 ini doğru tahmin etmiştir.



Şekil 5. 8 Farklı k değerleri için Roc eğrileri

Şekil 5.8'de farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 3 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %79 başarı gösterdiği görülmektedir.

5.2.2.2 C5.0 Karar Ağacından Elde Edilen Sonuçlar

C5.0 algoritmasının tanısal performansını göstermek için; sınıflandırma doğruluğu, ROC eğrisi altındaki alan, 3,5 ve 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 5. 31 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	%76,429 (1297)	%75,957 (1289)	%75,957 (1289)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1297 örnek ile k nın 3 olduğu durumda ortaya çıktığı görülmektedir. Model 3 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %76,42 başarı göstermiştir.

Çizelge 5. 32 Farklı k değerleri için model performans değerlendirme ölçütleri

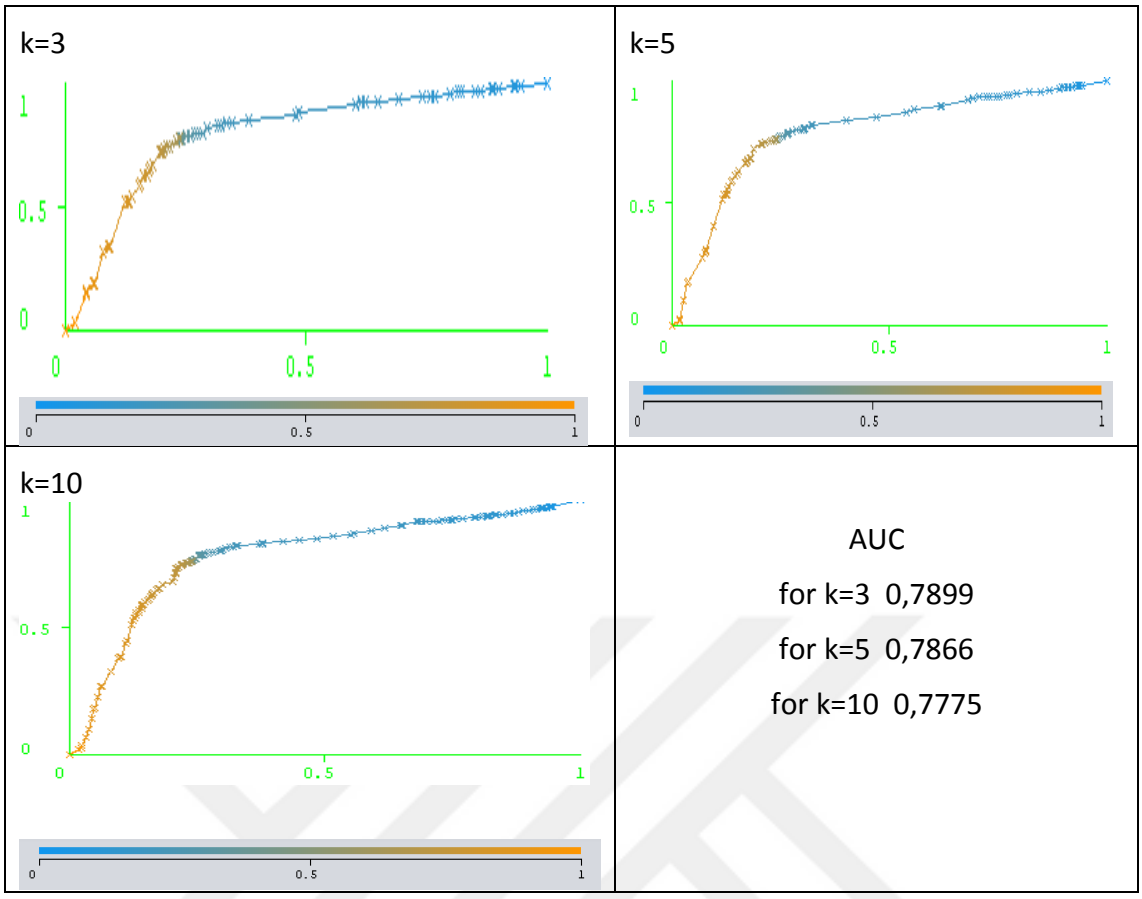
k Kat Çapraz Doğrulama	DP Oranı	YP Oranı	Kesinlik	F-Ölçütü	Sınıf
k=3	0,760	0,232	0,766	0,763	Sağlıklı
	0,768	0,240	0,763	0,766	Hasta
Ağırlıklı Ort	0,764	0,236	0,764	0,764	
k=5	0,759	0,240	0,759	0,759	Sağlıklı
	0,760	0,241	0,760	0,760	Hasta
Ağırlıklı Ort	0,760	0,240	0,760	0,760	
k=10	0,760	0,241	0,759	0,759	Sağlıklı
	0,759	0,240	0,761	0,760	Hasta
Ağırlıklı Ort	0,760	0,240	0,760	0,760	

k'nın 3 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. Doğru pozitif oranı 0,764, yanlış pozitif oranı 0,236 olarak görülmektedir.

Çizelge 5. 33 Farklı k değerleri için kontenjans tablosu

N=1697	k=3		k=5		k=10		Toplam
	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	
Sağlıklı	644	203	643	204	644	203	847
Hasta	197	653	204	646	205	645	850

Çizelge 5.35'de farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 3 kat olduğu durumda görülmektedir. k'nın 3 olduğu durumda 847 sağlıklı bireyin 644 ünü doğru tahmin ederken, 850 hasta bireyin 653 ünü doğru tahmin etmiştir.



Şekil 5. 9 Farklı k değerleri için Roc eğrileri

Şekil 5.9'da farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 3 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %78 başarı gösterdiği görülmektedir.

5.2.2.3 Bayes Ağları Yöntemi Sonuçları

Bayes ağları tanısal performansını göstermek için; sınıflandırma doğruluğu, ROC eğrisi altındaki alan, 3,5 ve 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 5. 34 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	%71,3612 (1211)	%71,2434 (1209)	%71,3023 (1210)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1211 örnek ile k nın 3 olduğu

durumda ortaya çıktığı görülmektedir. Model 3 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %71,36 başarı göstermiştir.

Çizelge 5. 35 Farklı k değerleri için model performans değerlendirme ölçütleri

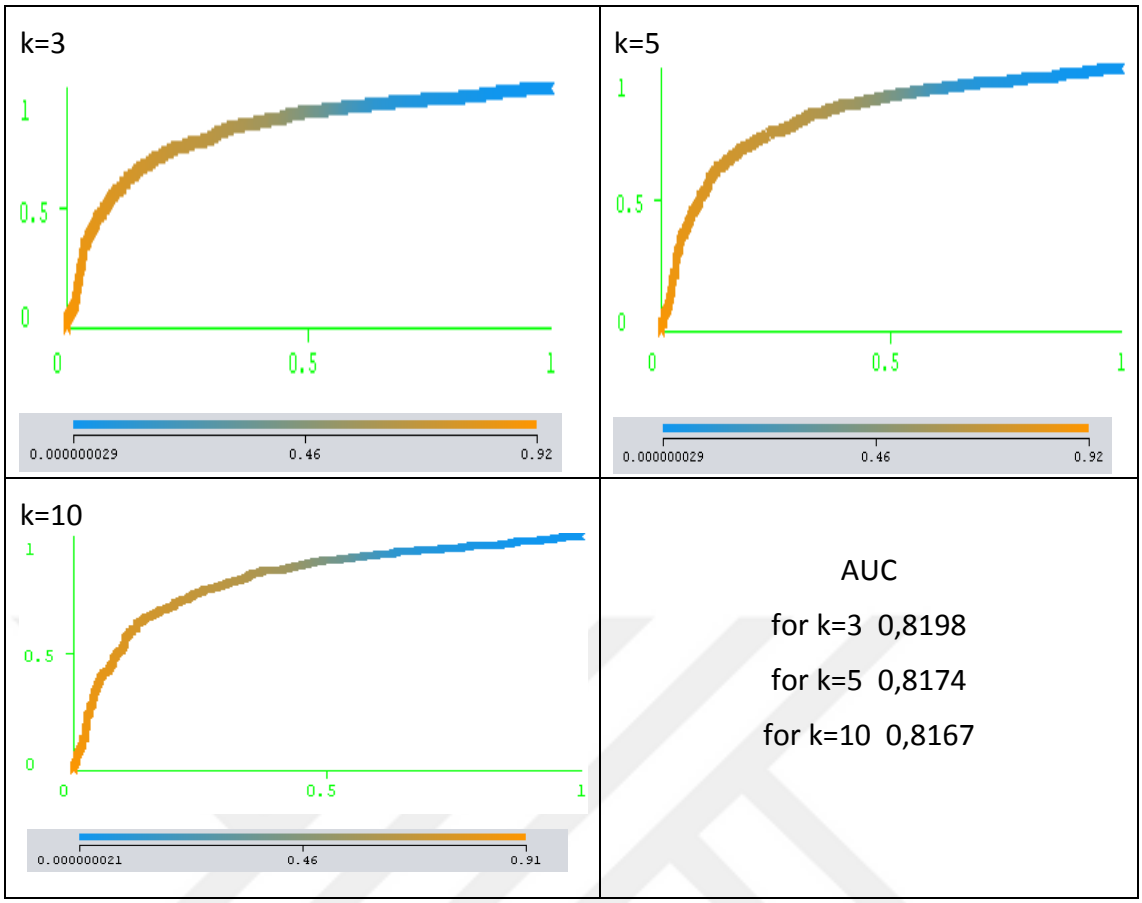
k Kat Çapraz Doğrulama	**DP Oranı**	**YP Oranı**	**Kesinlik**	**F-Ölçütü**	**Sınıf**
k=3	0,551	0,125	0,815	0,658	Sağlıklı
	0,875	0,449	0,662	0,875	Hasta
Ağırlıklı Ort	0,714	0,287	0,738	0,706	
k=5	0,551	0,127	0,812	0,657	Sağlıklı
	0,873	0,449	0,661	0,753	Hasta
Ağırlıklı Ort	0,712	0,288	0,737	0,705	
k=10	0,554	0,128	0,811	0,658	Sağlıklı
	0,872	0,446	0,662	0,753	Hasta
Ağırlıklı Ort	0,713	0,288	0,737	0,706	

k'nın 3 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. k'nın 3 kat olduğu modelde doğru pozitif oranı 0,714 yanlış pozitif oranı 0,287 olarak görülmektedir.

Çizelge 5. 36 Farklı k değerleri için kontenjans tablosu

	k=3		**k=5**		**k=10**		**Toplam**
N=1697	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	
Sağlıklı	467	380	467	380	469	378	847
Hasta	106	744	108	742	109	741	850

Çizelge 5.36'da farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 3 kat olduğu durumda görülmektedir. k'nın 3 olduğu durumda model 847 sağlıklı bireyin 467 sini doğru tahmin ederken, 850 hasta bireyin 744 ünü doğru tahmin etmiştir.



Şekil 5. 10 Farklı k değerleri için Roc eğrileri

Şekil 5.10'da farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 3 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %81 başarı gösterdiği görülmektedir.

5.2.2.4 Destek Vektör Makineleri Yöntemi Sonucu

Destek Vektör Makineleri tanısal performansını göstermek için; sınıflandırma doğruluğu, ROC eğrisi altındaki alan, 3,5 ve 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 5. 37 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	73.1291 % (1241)	72.6576 % (1233)	72.4219 % (1229)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1241 örnek ile k nın 3 olduğu

durumda ortaya çıktığı görülmektedir. Model 3 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %73,12 başarı göstermiştir

k Kat Çapraz Doğrulama	DP Oranı	YP Oranı	Kesinlik	F-Ölçütü	Sınıf
k=3	0,648	0,186	0,777	0,707	Sağlıklı
	0,814	0,352	0,699	0,752	Hasta
Ağırlıklı Ort	0,731	0,269	0,738	0,729	
k=5	0,645	0,192	0,770	0,702	Sağlıklı
	0,808	0,355	0,695	0,748	Hasta
Ağırlıklı Ort	0,727	0,274	0,733	0,725	
k=10	0,653	0,205	0,761	0,703	Sağlıklı
	0,795	0,347	0,697	0,743	Hasta
Ağırlıklı Ort	0,724	0,276	0,729	0,723	

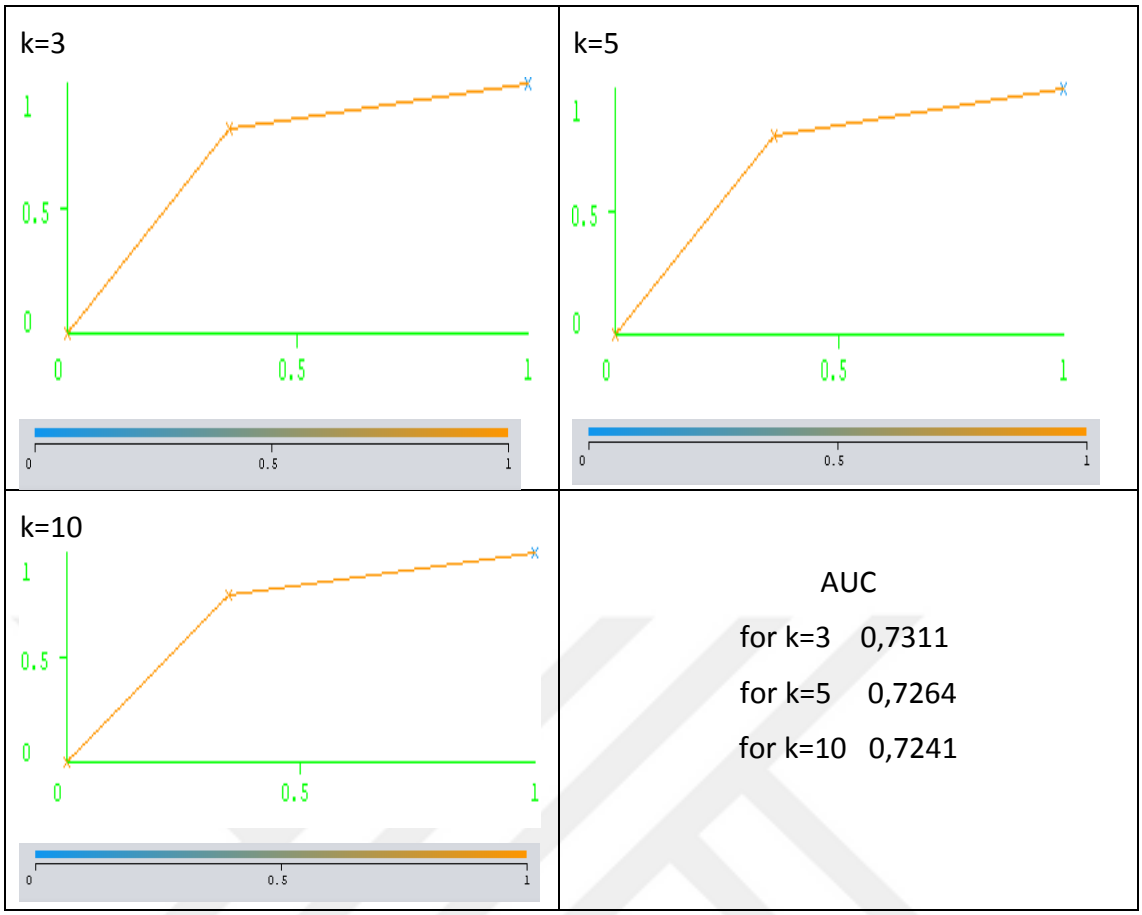
Çizelge 5. 38 Farklı k değerleri için model performans değerlendirme ölçütleri

k'nın 3 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. k'nın 3 kat olduğu modelde doğru pozitif oranı 0,731 ve yanlış pozitif oranı 0,269 olarak görülmektedir.

Çizelge 5. 39 Farklı k değerleri için kontenjans tablosu

	k=3		**k=5**		**k=10**		**Toplam**
N=1697	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	
Sağlıklı	549	298	546	301	553	294	847
Hasta	158	692	163	687	174	676	850

Çizelge 5.39'da farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 3 kat olduğu durumda görülmektedir. k'nın 3 olduğu durumda model 847 sağlıklı bireyin 549 unu doğru tahmin ederken, 850 hasta bireyin 692 sini doğru tahmin etmiştir.



Şekil 5. 11 Farklı k değerleri için Roc eğrileri

Şekil 5.11'de farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 3 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %73 başarı gösterdiği görülmektedir.

5.2.2.5 Yapay Sinir Ağları ile İlgili Bulgular

Yapay Sinir Ağları yönteminin tanısal performansını göstermek için; sınıflandırma doğruluğu, ROC eğrisi altındaki alan, 3,5 ve 10 kat çapraz doğrulama yöntemi kullanılmıştır.

Çizelge 5. 40 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	73,0112 % (1239)	74.8379 % (1270)	74.5433 % (1265)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1270 örnek ile k nın 5 olduğu

durumda ortaya çıktığı görülmektedir. Model 5 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %74,83 başarı göstermiştir.

Çizelge 5. 41 Farklı k değerleri için model performans değerlendirme ölçütleri

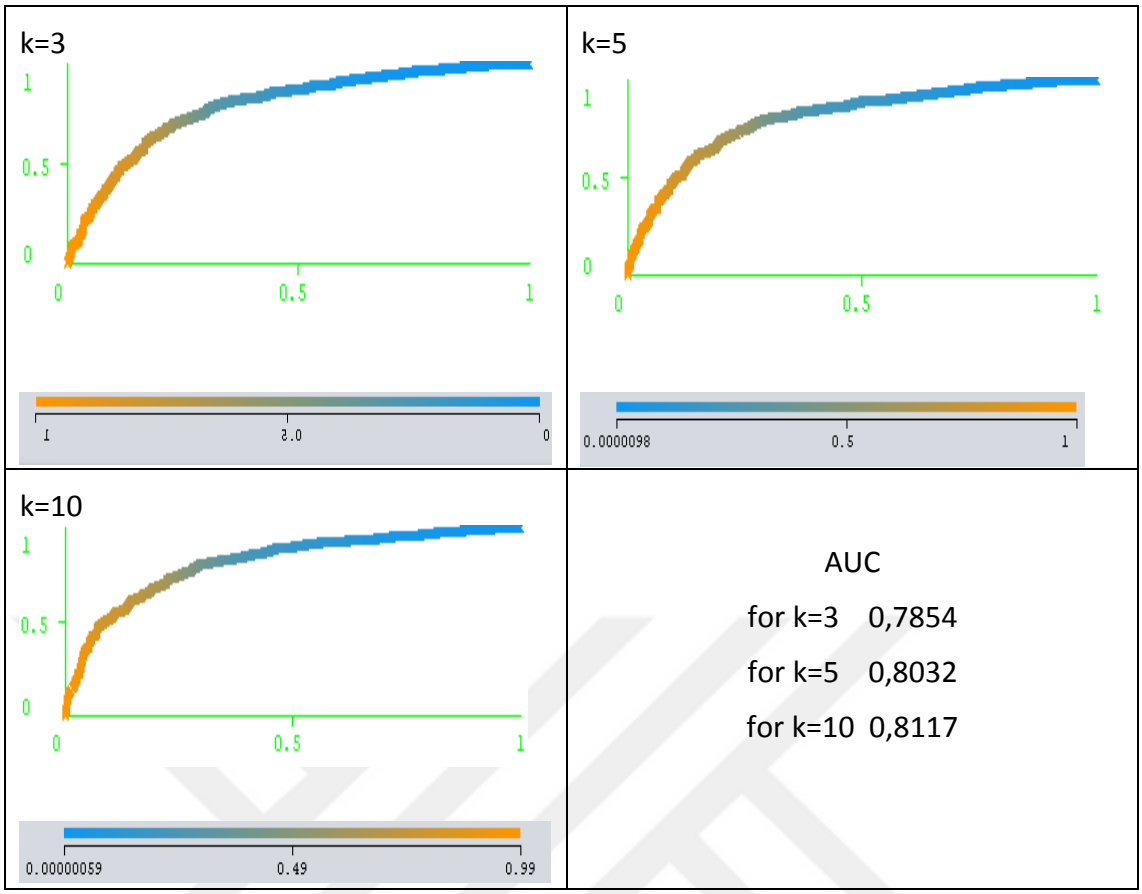
k Kat Çapraz Doğrulama	**DP Oranı**	**YP Oranı**	**Kesinlik**	**F-Ölçütü**	**Sınıf**
k=3	0,746	0,286	0,722	0,734	Sağlıklı
	0,714	0,254	0,714	0,726	Hasta
Ağırlıklı Ort	0,730	0,270	0,730	0,730	
k=5	0,727	0,231	0,727	0,743	Sağlıklı
	0,769	0,273	0,739	0,748	Hasta
Ağırlıklı Ort	0,748	0,252	0,749	0,748	
k=10	0,743	0,252	0,746	0,744	Sağlıklı
	0,748	0,257	0,745	0,746	Hasta
Ağırlıklı Ort	0,745	0,255	0,745	0,745	

k'nın 5 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. k'nın 5 kat olduğu modelde doğru pozitif oranı 0,748 iken yanlış pozitif oranı 0,273 olarak görülmektedir.

Çizelge 5. 42 Farklı k değerleri için kontenjans tablosu

	k=3		**k=5**		**k=10**		
N=1697	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	**Toplam**
Sağlıklı	632	215	616	231	629	218	847
Hasta	243	607	196	654	214	636	850

Çizelge 5.42'de farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 3 kat olduğu durumda görülmektedir. k'nın 3 olduğu durumda model 847 sağlıklı bireyin 632 sini doğru tahmin ederken, 850 hasta bireyin 607 sini doğru tahmin etmiştir.



Şekil 5. 12 Farklı k değerleri için Roc eğrileri

Şekil 5.12'de farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 10 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %81 başarı gösterdiği görülmektedir.

5.2.2.6 k En Yakın Komşu Algoritması ile İlgili Bulgular

k-NN yönteminde yine tanısal performansı göstermek için sınıflandırma doğruluğu ve ROC eğrisi altındaki alan yöntemleri kullanılmıştır. k en yakın komşulukta k nın 5 olduğu değer en iyi sonucu vermiştir.

Çizelge 5. 43 Farklı k değerleri için doğru sınıflandırma yüzdeleri

N=1697	**k=3**	**k=5**	**k=10**
Doğru Sınıflandırma Yüzdesi	%66,9417 (1136)	69,0631 % (1172)	68,9452 % (1170)

Tabloda modelin doğru sınıflandırma yüzdesinin çeşitli k değerlerindeki sonuçları görülmektedir. En iyi doğru sınıflandırma yüzdesinin 1172 örnek ile k nın 5 olduğu

durumda ortaya çıktığı görülmektedir. Model 5 kat çapraz doğrulama ile hastaları hasta ve sağlıklıları sağlıklı olarak sınıflandırma %69,06 başarı göstermiştir.

Çizelge 5. 44 Farklı k değerleri için model performans değerlendirme ölçütleri

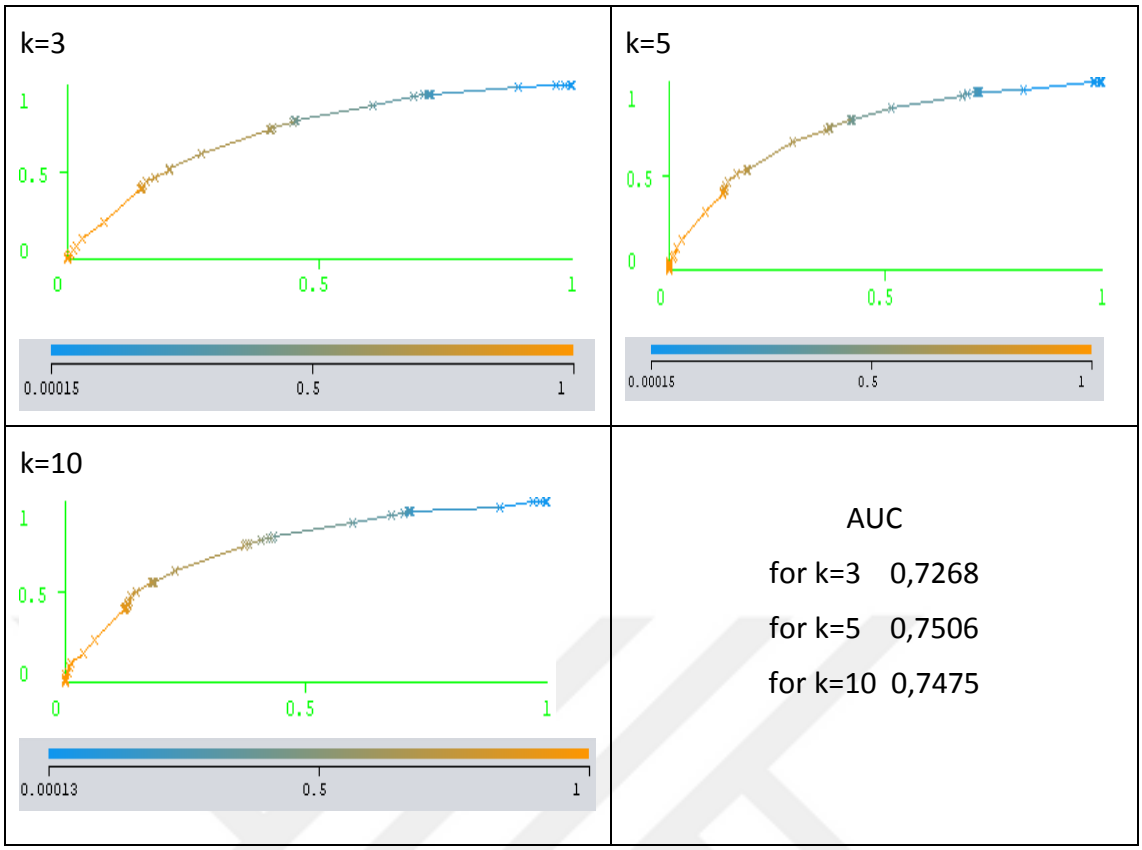
k Kat Çapraz Doğrulama	**DP Oranı**	**YP Oranı**	**Kesinlik**	**F-Ölçütü**	**Sınıf**
3	0,591	0,253	0,700	0,641	Sağlıklı
	0,747	0,409	0,647	0,694	Hasta
Ağırlıklı Ort	0,669	0,331	0,673	0,667	
5	0,599	0,205	0,744	0,664	Sağlıklı
	0,795	0,401	0,665	0,725	Hasta
Ağırlıklı Ort	0,691	0,310	0,694	0,689	
10	0,617	0,239	0,720	0,665	Sağlıklı
	0,761	0,383	0,666	0,711	Hasta
Ağırlıklı Ort	0,689	0,311	0,693	0,688	

k'nın 5 kat olduğu modelde doğru pozitif oranı en yüksek, yanlış pozitif oranı en düşük, kesinlik ve f ölçütü diğer k değerlerini alan modellere göre en iyi sonucu vermektedir. k'nın 5 kat olduğu modelde doğru pozitif oranı yani modelin hastaları doğru olarak tahmin ettiği değer 0,691 yanlış pozitif 0,310 olarak görülmektedir.

Çizelge 5. 45 Farklı k değerleri için kontenjans tablosu

	k=3		**k=5**		**k=10**		**Toplam**
N=1167	Sağlıklı	Hasta	Sağlıklı	Hasta	Sağlıklı	Hasta	
Sağlıklı	501	346	532	315	523	324	847
Hasta	215	635	210	640	203	647	850

Çizelge 5.45'de farklı k değerleri için kontenjans matrisi gösterilmiştir. Hastaları hasta olarak tahmin etmede en yüksek tahmin k'nın 5 kat olduğu durumda görülmektedir. k'nın 5 olduğu durumda model 847 sağlıklı bireyin 532 sini doğru tahmin ederken, 850 hasta bireyin 640 ını doğru tahmin etmiştir.



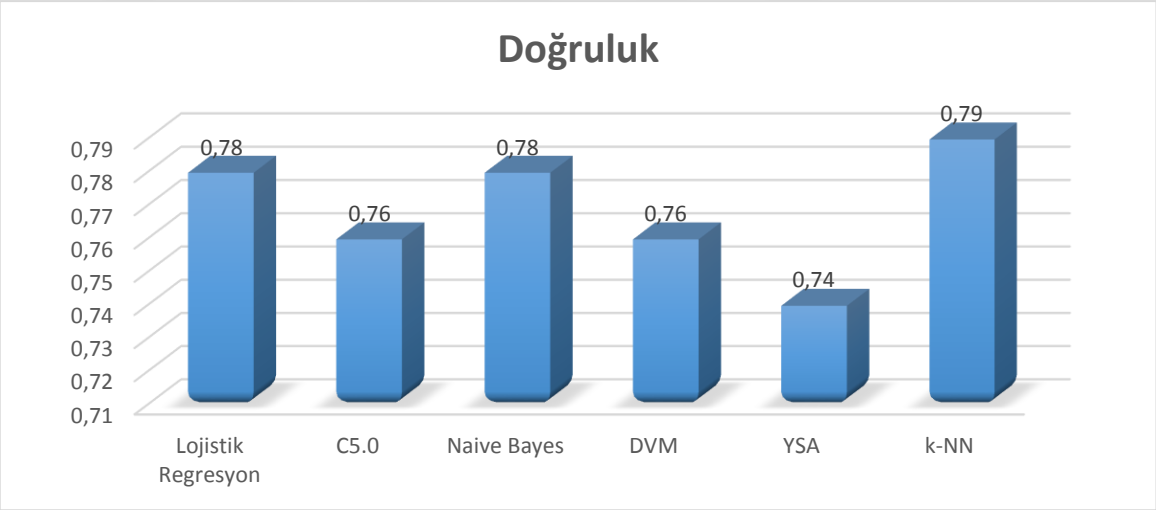
Şekil 5. 13 Farklı k değerleri için Roc eğrileri

Şekil 5.13'de farklı k değerleri için ROC eğrileri gösterilmekte ve AUC değerleri verilmektedir. k'nın 5 kat olduğu model en iyi AUC değerini göstermektedir. Lojistik regresyonun hastalıkları sağlıklardan ayırt etmede %75 başarı gösterdiği görülmektedir.

BÖLÜM 6

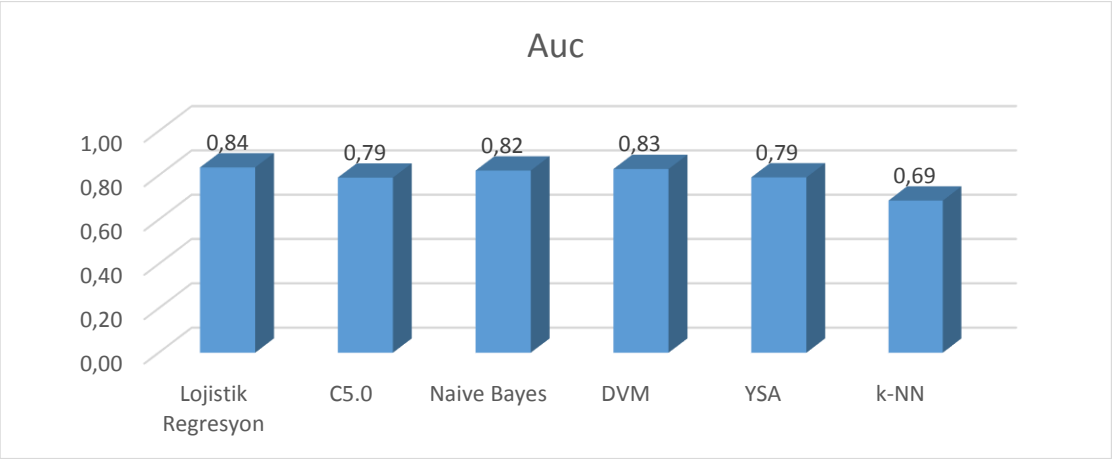
SONUÇ VE ÖNERİLER

6.1 Eğitim ve Test Seti Yöntemi ile Elde Edilen Sonuçların Karşılaştırılması



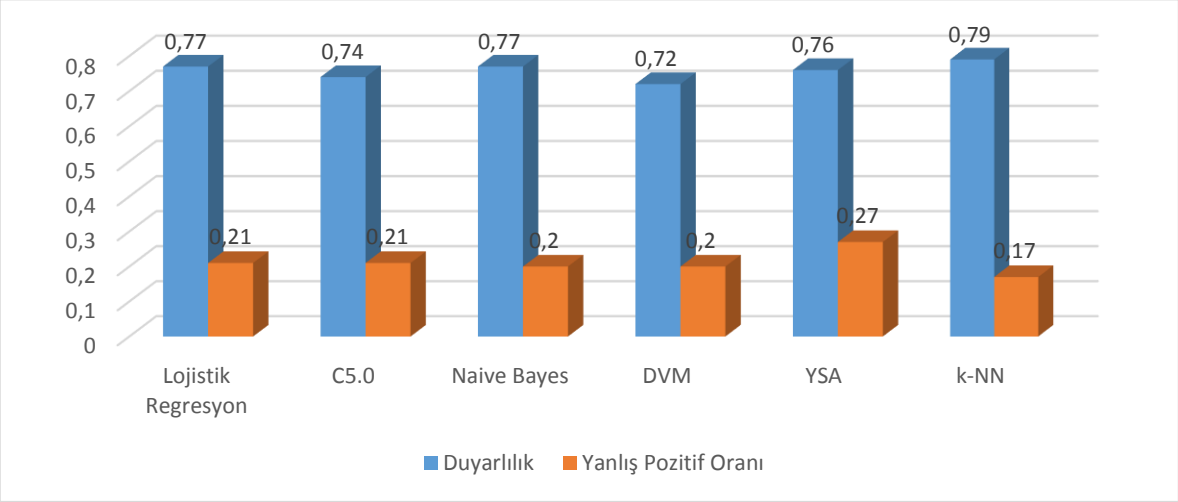
Şekil 6. 1 Tüm modellerin doğruluk karşılaştırması

En iyi model doğruluğu aşırı öğrendiği için k-NN'de görülmektedir. O yüzden k-NN modeli dışında değerlendirme yapılacaktır. Naive bayes ve lojistik regresyonun hastaları hasta ve sağlıklıları sağlıklı olarak tahmin etmesinde %78 ile aynı doğruluğu göstermiştir.



Şekil 6. 2 Tüm modellerin AUC karşılaştırması

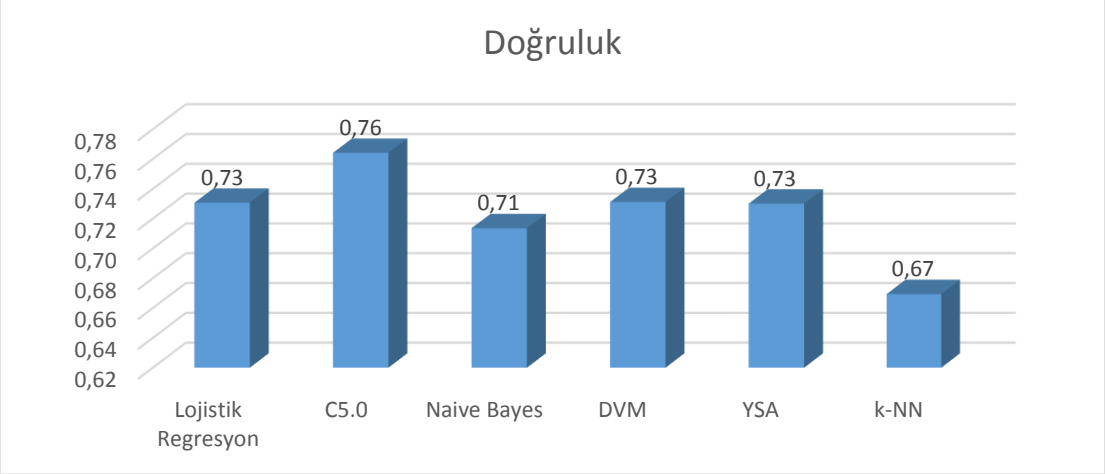
Hastalıklı bireyleri hastalıklı olarak tahmin etmede en iyi sonucu veren sınıflandırma yönteminin destek vektör makinelerinden çok küçük bir farkla lojistik regresyon olduğu görülmektedir. En düşük performansı model aşırı öğrendiği k-NN göstermiştir. Modelin aşırı öğrenmesi gerçekten uzak sonuçlar tahmin etmeye neden olur.



Şekil 6. 3 Tüm modellerin duyarlılık ve yanlış pozitif oranı karşılaştırması

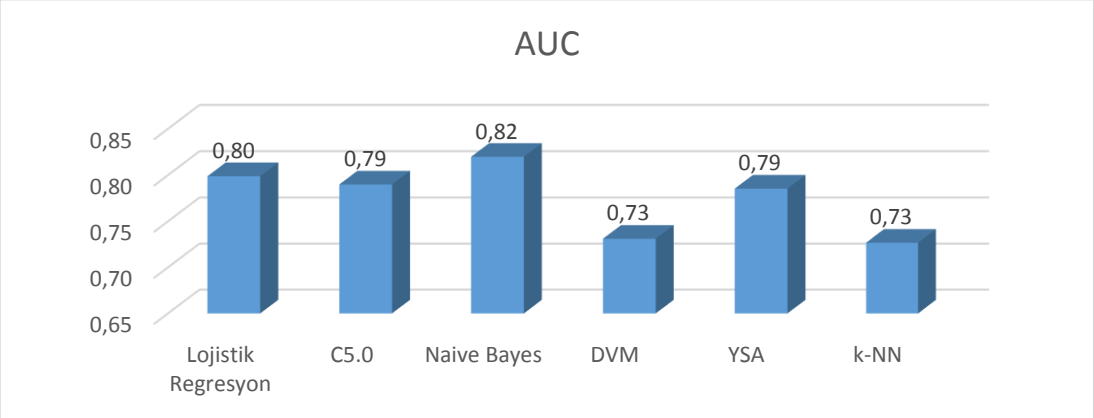
En iyi performans sonucu doğru olarak sınıflandırılan hasta yüzdesinin en yüksek ve hasta olanların sağlıklı olarak sınıflandırılmış yüzdesini veren yanlış pozitif oranının en düşük olduğu durumdur. Bu durumu lojistik regresyona göre çok küçük bir farkla ile Naive Bayes sağlamaktadır.

6.2 Çapraz Doğrulama Yöntemi ile Elde Edilen Sonuçların Karşılaştırılması



Şekil 6. 4 Tüm modellerin doğruluk karşılaştırması

En iyi model doğruluğu Karar ağacı algoriması olan C5.0 da 0,76 olduğu görülmektedir. Lojistik regresyon, DVM ve YSA'nın hastaları hasta ve sağlıklıları sağlıklı olarak tahmin etmesinde %73 ile aynı doğruluğu göstermiştir.



Şekil 6. 5 Tüm modellerin AUC karşılaştırması

Hastalıklı bireyleri hastalıklı olarak tahmin etmede en iyi sonucu veren sınıflandırma yönteminin Naive Bayes (0,82) olduğu görülmektedir. En kötü model performansını ise DVM ve k-NN olduğu görülmektedir.

Yapılan çalışma iki farklı yöntemle değerlendirilmiştir. İlkinde veri seti %80 eğitim ve %20 test seti olarak ayrılmış ve uygulamalar SPSS Moduler 18.0 ile yapılmıştır. Modellerin doğrulukları sırası ile Lojistik regresyon %78, Naive Bayes %78, C5.0 algoritması %76, Destek Vektör Makineleri %76, CRT algoritması %74, Yapay Sinir Ağları %74 bulunmuştur. K-NN de aşırı öğrenme görüldüğü için modeller arası değerlendirmeye katılmamıştır.

Modellerin ROC eğrisi altındaki alan değerlendirmesi(AUC) Lojistik regresyon %83, Destek Vektör Makineleri %82,9, Naive Bayes %82,2, Yapay Sinir Ağları %79,1, C5.0 algoritması %79, CRT algoritması %77,9 Yapay Sinir Ağları %87 ve k-NN %68,6 bulunmuştur. k-NN'in hasta bireyleri hasta olmayan bireylerden ayırt etmeden bu kadar düşük performans göstermesi modelin eğitim setinde aşırı öğrenmesidir.

Modellerin hata oranları; Lojistik regresyon %22, CRT algoritması %26, C5.0 algoritması %24, Naive Bayes %22, Destek Vektör Makineleri %24, Yapay Sinir Ağları %26 ve k-NN %21 bulunmuştur. Modelde en düşük hata oranını aşırı öğrendiği için k-NN göstermişrtir.

Hastalıkları sağlıklıdan ayırmada en az hata yapan ve en doğru model Lojistik Regresyon ve Naive Bayes olduğu görülmektedir.

Duyarlılık oranı en yüksek ve yanlış pozitif oranı en düşük model en iyi ayırt edici sonucu verir. Yine Şekil 6.3'de görüldüğü üzere 0,77 duyarlılık ve 0,2 yanlış pozitif oranı ile naive bayes yönteminin en iyi sonucu verdiğini görülmektedir.

C5.0 algortmasında algoritmasında en etkili değişkenler sırası ile hispanik, ırk ve son mamografi sonucu, Naive Bayes modelinde yaş grupları, Destek Vektör Makinelerinde ırk, beden kütle indeksi ve yaş grupları, ilk doğum yaşı ve hispanik, Yapay Sinir Ağlarında ise ırk, yaş grupları ve hispanik değişkenler en önemli tahminciler olarak bulunmuştur.

İkinci yöntemde ise veri setinde k-kat çapraz doğrulama yöntemi kullanılmış ve uygulamalar WEKA ile yapılmıştır. Çeşitli k değerleri kullanılarak farklı sonuçlara ulaşılmıştır. Model performans değerlendirmeleri ROC eğrisi altında (AUC) doğru pozitif (DP), yanlış negatif (YN), doğru negatif (DN), yanlış pozitif (YP), F-ölçütü ve özgüllük değerleri kullanılarak yorumlanmıştır.

Lojistik regresyon sınıflandırma doğruluğu 5 kat çapraz doğrulamada en iyi sonucu vermiştir (%79). K'nın 5 olduğu durumda doğru pozitif, yanlış pozitif ve kesinlik değerleri en iyi sonucu verirken, AUC k'nın 3 olduğu durumda en iyi sonucu vermiştir.

C5.0 algoritmasında sınıflandırma doğruluğu (%76) ve diğer performans değerlendirme ölçütleri 3 kat çağraz doğrulamada en iyi sonucu vermiştir.

Naive Bayes metodunda sınıflandırma doğruluğu (%71) ve diğer performans değerlendirme ölçütleri 3 kat çağraz doğrulamada en iyi sonucu vermiştir.

Yapay Sinir Ağlarında sınıflandırma doğruluğu (%75) ve diğer performans değerlendirme ölçütleri 5 kat çağraz doğrulamada en iyi sonucu vermiştir.

k-NN sınıflandırma doğruluğu (%69) ve diğer performans değerlendirme ölçütleri 5 kat çağraz doğrulamada en iyi sonucu vermiştir.

Sınıflandırma yöntemlerininin 3 kat çapraz doğrulamaya göre AUC Şekil 8.5'de gösterilmektedir. Hastaları sağlıklılardan ayırmada en iyi model başarısını %82 lik AUC değeri ile Naive Bayes göstermiştir.

12 değişkenli veri setinde çoğu modelde en önemli değişkeninin ırk, hispanik kökenli olma durumu ve yaş gruplarının kanser riskini arttırdığı görülmektedir. Asya/Pasifik Adalı ve hispanik kökenli kadınlarda kanser riskinin diğer gruplara göre oldukça fazla olduğu söylenebilir.

Bu çalışmaya ek olarak sigara, alkol kullanma durumları, sağlıklı beslenme, genetik faktörlerin de incelenmesi ve daha önce literatürde yapılmış tümör bilgilerinden elde edilerek yapılmış tahminlerin birleştirilmesi ile kanser riskinin önceden teşhisi büyük oranda sağlanabilir.

KAYNAKLAR

[1] Kıyan, T. ve Yıldırım, T. (2004). "Breast Cancer Diagnosis Using Statistical Neural Networks", Journal of Electrical & Electronics Engineering, 1149-1153.

[2] Maglogiannis, I., Zafiropoulos, E. ve Anagnostopoulos, I. (2007). "An Intelligent System for Automated Breast Cancer Diagnosis and Prognosis Using SVM Based Classifiers", Springer, 24-36.

[3] Bradley, A., (1997). "The Use of The Area Under The Roc Curve In The Evaluation Of Machine Learning Algorithms", Elsevier, 1145-1159.

[4] Salama, G. I., Abdelhalim, M.B. ve Zeid, M. A., (2012). "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology , 2277 – 0764.

[5] Sawarkar, S. D., Ghatol, A. A. ve Pande, A. P., (2006). "Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine" Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia, 158-163.

[6] Polat, K. ve Güneş, S., (2007). "Breast Cancer Diagnosis Using Least Square Support Vector Machine", Elsevier, 694-701.

[7] Decruyenaere A, P., Peters, P., Vermassen, F. ve Dhaene, T., (2015). "Prediction of Delayed Graft Function After Kidney Transplantation: Comparison Between Logistic Regression and Machine Learning Methods", BMC Medical Informatics and Decision Making.

[8] Hurtado, O.M., Guest, R., Stevenage, S. V. ve Neil, G. J., (2016). "Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship Between Hand Dimensions and Demographic Characteristics", PLOS ONE, 1-25.

[9] Saraoğlu, H.M., Yıldırım,M., Özbeyaz, A. ve Temurtaş F., (2012). "Analysis of Palm Perspiration Effect with SVM for Diabetes in People", International Scholarly and Scientific Research & Innovation, 489-493.

[10] Şatır, E., Azaboy, F., Aydın, A., Arslan, H. ve Hacıefendioğlu, Ş., (2016). "Veri İndirgeme ve Sınıflandırma Teknikleri ile Glokom Hastalığı Teşhisi", El-Cezerî Journal of Science and Engineering, 485-497.

[11] Ülgen, I. G. Osman, O., Özekeş, S., Baslo, B. ve Ertaş, M.,(2012). "Classification of Juvenile Myoclonic Epilepsy Data Acquired Through Scanning Electromyography with Machine Learning Algorithms", Springer Science+Business, 2705-2711.

[12] Çomak,E., Polat, K., Güneş, S. ve Arslan A., (2005). "A New Medical Decision Making System:Least Square Support Vector Machine (LSSVM) with Fuzzy Weighting Pre-Processing", Elsevier , 409-414.

[13] Dreiseitl,S., Kittler, H., Vinterbo,S., Billhardt,H. ve Binder, M., (2001). "A Comparison of Machine Learning Methods for the Diagnosis of Pigmented Skin Lesions", Journal of Biomedical Informatics, 28-36.

[14] Verplancke, T., Looy,S.V., Benoit D., Vansteeland,S., Turck, F.D. ve Decruyenare,J., (2008). "Support Vector Machine Versus Logistic Regression Modeling for Hospital Mortality in Critically İll Patients With Haematological Malignancies", BMC Medical Informatics and Decision Making, 1-8.

[15] Yu, K., Zhang C., Re, C., Rubin, D.L. ve Snyder, M., (2016). "Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features", Nature Communicatıons, 1-10.

[16] Rampun, A., Tiddeman, B., Zwiggelaar, T. ve Malcolm, P., (2016). "Computer Aided Diagnosis of Prostate Cancer: A texton Based Approach", Medical Physics, 5412-5425.

[17] Salah, B., Alshraideh, M., Beidas, R. ve Hayajneh, F., (2011). "Skin Cancer Recognition by Using a Neuro-Fuzzy System", Cancer Informatics, 1-11.

[18] Soni, J., Ansari, U., Sharma, D. ve Soni, S. (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, 43-48.

[19] Wang, L., Zheng, W.L., Ma, H.W. ve Lu,B.L., (2016). "Measuring Sleep Quality from EEG with Machine Learning Approaches", IEEE, 905-912.

[20] Aljaaf, A. J., (2016). "Evaluation of Machine Learning Methods to Predict Knee Loading from the Movement of Body Segments", IEEE, 5168-5173.

[21] Parthiban, G. ve S.K.Srivatsa., (2012). "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients", International Journal of Applied Information Systems, 25-36.

[22] Krawczyk, B., Simic, S., Simic, D. ve Wozniak, M., (2012). "Automatic Diagnosis of Primary Headaches By Machine Learning Methods", Central European Journal of Medicine, 157-165.

[23] Çınar, M., Engin, M., Engin, E. Z. ve Ateşçi, Y. Z.,(2009). "Early Prostate Cancer Diagnosis by Using Artificial Neural Networks and Support Vector Machines", Elsevier, 6357–6361.

[24] Demšar, J., Kattan, M. W., Beck, J. R. ve Bratko, I., (1999). "Machine Learning for Survival Analysis: A Case Study on Recurrence of Prostate Cancer", Springer, 346-350.

[25] Moor, J. H., (2012). The Turing Test:The Elusive Standard of Artificial Intelligence, Springer Science & Business Media.

[26] Alpaydın, E., (2010). Introduction to Machine Learning, Massachusetts Institute of Technology, United States of America.

[27] Mitchell, T., (1997). Machine Learning, NY: McGraw-Hill Science/Engineering/Math.

[28] Shalev-Shwartz, S. ve Ben-David, S., (2014). "Understanding Machine Learning:From Theory to Algorithms", Cambridge University Press.

[29] Nilsson, N. J., (1998). "Introduction to Machine Learning , An Early Draft of A Proposed Textbook", Stanford: Stanford University.

[30] Murphy, K. P., (2012). Machine Learning A Probabilistic Perspective, Cambridge, Massachusetts,London, England: The MIT Press.

[31] Langley, P., (1998). Element of Machine Learning, London: Morgan Kaufmann.

[32] Mohri,M., Rostamizadeh, A. ve Talwalkar, A., (2012). Foundations of Machine Learning,London: The MIT Press.

[33] Schapire, R. (2008). "Theoretical Machine Learning", https://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf, 9 Ocak 2017.

[34] O'Halloran,S.,Logistical Regression II-Multinomial Data. http://www.columbia.edu/~so33/SusDev/Lecture_10.pdf, 25 Mart 2017.

[35] Vapnik, V. N., (2000). The Nature of Statistical Learning Theory, New York: Springer.

[36] Bayer, H., Çoban T., (2015). "Web İstatistiklerinde Makine Öğrenmesi Algoritmaları İle Kritik Parametre Tespiti", Electronic Journal of Vocational Colleges- Special Issue: The Latest Trends in Engineering.

[37] Stephan, C., Cammann, H., Semjonow, A., Diamandis, E. P., Wymenga, L. F., Lein, M. ve Jung, K., (2002). "Multicenter Evaluation of Evaluation of an Artifical Neural Network to Increase the Prostate Cancer Detection Rate and Reduce Unnecessary Biopsies", Cancer Diagnostics:Discovery and Clinical Applications, 1279–1287.

[38] Parthiban, G. ve S.K.Srivatsa., (2012). "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients", International Journal of Applied Information Systems, 25-36.

[39] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet,C. ve Haussler, D., (1999). "Support Vector Machine Classification of Microarray Gene Expression Data", UCSC-CRL-99-09.

[40] Soni,J., Ujma, A., Sharma, D. ve Soni,S., (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, 43-48.

[41] Pandya, R. ve Pandya, J., (2015). "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning", International Journal of Computer Applications , 0975 – 8887.

[42] Lewis, D. D., (1998). "The Independence Assumption in Information Retrieval", Springer Berlin Heidelberg, 4-15.

[43] Maglogiannis, I., Zafiropoulos, E. ve Anagnostopoulos, I., (2007). "An Intelligent System for Automated Breast Cancer Diagnosis and Prognosis Using SVM Based Classifiers", Springer, 24-36.

[44] Wang, S. ve Summers, R. M., (2012). "Machine Learning and Radiology",Elsevier, 933-951.

[45] Regnier-Couderta, O., McCalla, J., Lothiana, R., Lam, T. ve McClinton, S., (2011), "Machine Learning for Improved Pathological Staging of Prostate Cancer:A Performance Comparison on A Range of Classifiers", Elsevier, 25-35.

[46] Schneider,J.http://www.cs.cmu.edu/~schneide/tut5/node42.html. http://www.cs.cmu.edu/~schneide/tut5/node42.html, 28 Aralık 2016

[47] Kartal, E., (2015). Sınıflandırmaya Dayalı Makine Öğrenmesi Teknikleri ve Kardiyolojik Risk Değerlendirmesine İlişkin Bir Uygulama, Doktora, İÜ, Enformatik Anabilim Dalı, İstanbul.

[48] World Kanser Research Fund İnternational.:http://www.wcrf.org/int/Kanser-factsfigures/data-specific-Kansers/breast-Kanser-statistics, 2 Mart,2017.

[49] Yeni Dünya Kanser İstatistikleri, (2017, Nisan). Kanser Daire Başkanlığı: http://kanser.gov.tr/daire-faaliyetleri/kanser-istatistikleri/860-yeni-d%C3%BCnya-kanser-istatistikleri-yay%C4%B1nland%C4%B1.html, 3 Nisan 2017.

[50] Desantis, C.E., Fedewa, S.A., M., Sauer, A.G., M., Kramer, J.L., Smith, R.A. ve Jemal, A., (2015). "Breast Cancer Statistics, 2015: Convergence of Incidence Rates Between Black and White Women", CA: A Cancer Journal for Clinicians, 1542-4863.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Meliha Nur DURAK

Doğum Tarihi ve Yeri : 11/02/1992- Malkara/TEKİREDAĞ

Yabancı Dili : İngilizce

E-posta : mnurd.59@gmail.com

ÖĞRENİM DURUMU

Derece	Alan	Okul/Üniversite	Mezuniyet Yılı
Y. Lisans	İstatistik	Yıldız Teknik Üniversitesi	2017
Y. Lisans	Biyostatistik	İstanbul Üniversitesi	2017
Lisans	İstatistik	Yıldız Teknik Üniversitesi	2015
Lise	Fen	Malkara Anadolu Lisesi	

İŞ TECRÜBESİ

Yıl	Firma/Kurum	Görevi
2016-2017	Monitor Medikal Araştırma ve Danışmanlık	İstatistikçi ve Veri Giriş Uzmanı
2017	Greyder	Veri Analizi ve Raporlama Uzmanı