T.C.
İNÖNÜ ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
YAZILIM MÜHENDİSLİĞİ BÖLÜMÜ

MAKİNE ÖĞRENMESİ PROJE RAPORU
GÖĞÜS KANSERİ SINIFLANDIRILMASI

UMUT SEFKAN SAK
02210224071

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÖZET

Bu çalışma, sınıflandırma modellerinin performansını artırmak amacıyla kullanılan boyut indirgeme tekniklerinin etkisini değerlendirmeyi ve K-Nearest Neighbors (KNN) algoritması ile yüksek doğruluk oranlarına ulaşmayı hedeflemektedir. KNN algoritması, basitliği ve hızlı uygulanabilirliği nedeniyle temel sınıflandırma yöntemi olarak tercih edilmiştir. Çalışmada, farklı boyut indirgeme yöntemleri uygulanarak, veri setinin daha anlamlı bir hale getirilmesi ve KNN algoritmasının doğruluğunun optimize edilmesi sağlanmıştır.

Literatürde yer alan çeşitli sınıflandırma yaklaşımları incelenmiş ve bu çalışmaların doğruluk oranları değerlendirilmiştir. Örneğin, Delen vd.'nin karar ağaçları yöntemi ile %93.6, yapay sinir ağları ile %91.2 [7]; Kolay ve Erdoğmuş'un K-means yöntemiyle elde ettiği %45 ile %79 arasında değişen doğruluk oranları [17]; Karabatak ve İnce'nin hibrit modeliyle %95.6 doğruluk oranı [11] ve Papageorgiou vd.'nin Fuzzy Cognite Map (FCM) yöntemiyle %95 doğruluk oranı [16], farklı tekniklerin performansları hakkında bilgi sunmaktadır. Bu çalışmalar, doğruluk oranlarının sınıflandırma problemlerinde kullanılan yöntemlere göre büyük ölçüde değişebileceğini göstermiştir.

Çalışmada, Principal Component Analysis (PCA) ve Neighborhood Components Analysis (NCA) teknikleri kullanılarak veri setindeki boyutlar optimize edilmiş ve modelin sınıflandırma performansı iyileştirilmiştir. NCA uygulaması sonucunda modelin doğruluk oranı **%99,41** seviyesine ulaşmıştır. Bu sonuçlar, boyut indirgeme yöntemlerinin ve KNN algoritmasının birlikte etkili bir şekilde kullanıldığında sınıflandırma başarı oranlarını nasıl artırabileceğini göstermektedir.

Çalışmada elde edilen sonuçlar yalnızca modelin doğruluğunu artırmakla kalmamış, aynı zamanda hesaplama yükünü azaltarak daha verimli bir işlem süreci sunmuştur. Bu bağlamda, çalışma, sınıflandırma problemleri üzerinde çalışan araştırmacılar ve uygulayıcılar için önemli bir kaynak niteliği taşımaktadır.

1.GİRİŞ

1.1 Problem Tanımı

Göğüs kanseri, tüm dünyada kadınlar arasında en yaygın kanser türü olup, erken teşhis edilmediği takdirde ölümcül olabilmektedir. Bu sebeple, kanserin erken evrede tespit edilmesi, tedavi şansını artırmak ve hayatta kalma oranlarını yükseltmek açısından son derece kritik bir öneme sahiptir. Ancak, mevcut teşhis süreçlerinde insan hataları ve klinik gözlemlerin sınırlı kalması gibi faktörler, erken teşhisin önündeki engellerden biridir. Bu sorunu çözmek amacıyla makine öğrenmesi tekniklerinin kullanılması, sağlık sektöründe devrim yaratabilecek bir çözümdür.

Göğüs kanseri özellikle 40 – 49 yaş arası kadınlarda sıklıkla görülen ve dünya genelinde kadınlar arasında kanser kaynaklı en yüksek oranda ölüme sebep olan hastalıktır [1]. 2018 yılında tüm dünyada tespit edilen 18.1 milyon kanser vakası içerisinde %11.6 oranla akciğer kanserinden sonra ikinci sırada yer almaktadır [2].

1.2 Projenin Amacı:

Bu projenin temel amacı, göğüs kanseri hastalarının verilerini analiz ederek kanserin **iyi huylu (benign)** ya da **kötü huylu (malignant)** olarak sınıflandırılmasıdır. Erken tanı, tedavi süreçlerinin başarısını doğrudan etkileyebilmektedir. Bu bağlamda, doğru ve hızlı bir sınıflandırma sistemi, klinik tanı süreçlerinde önemli bir yardımcı araç olabilir.

Makine öğrenmesi, bu tür biyolojik veri setlerini analiz etmek için güçlü bir araçtır. Bu projede, hastaların çeşitli özelliklerini (örneğin, tümör boyutu, şekli ve yapısı) kullanarak, kanserin türünü doğru bir şekilde tahmin edebilen bir model geliştirmek hedeflenmiştir. Erken teşhis ve doğru sınıflandırma, tedavi seçeneklerinin belirlenmesinde büyük önem taşımaktadır. Bu nedenle, geliştirilen modelin doğruluğu, hastaların tedavi sürecine büyük katkı sağlayabilir.

1.3 Projenin Kapsamı

Proje, göğüs kanseri hastalarının çeşitli biyolojik verilerini kullanarak kanserin türünü sınıflandırmaya yöneliktir. Veriler, Kaggle platformunda yer alan bir göğüs kanseri sınıflandırma veri setinden alınmıştır ve bu veri seti, her bir hastanın medikal verilerine dayalı olarak kanserin iyi huylu ya da kötü huylu olduğu etiketlenmiştir.

2. PROJE AŞAMALARI

Bu projede kanser verisi üzerinde, KNN (K En Yakın Komşu) algoritması kullanarak bir sınıflandırma modeli geliştirilmiştir. Proje boyunca veri ön işleme, modelleme, modelin optimize edilmesi, boyut indirgeme ve sonuçların görselleştirilmesi gibi önemli adımlar takip edilmiştir. Aşağıda bu sürecin her bir aşaması detaylı bir şekilde açıklanacaktır.

2.1 Veri Seti Tanımı

Bu projede, kanserin iyi huylu ya da kötü huylu olup olmadığını belirlemek amacıyla kullanılan veri seti, Kaggle üzerinde bulunan Breast Cancer Wisconsin (Diagnostic) Data Set'tir [3] Bu veri seti, her bir hastanın özelliklerine dayalı olarak, kanserin türünü (iyi huylu ya da kötü huylu) etiketlemektedir. Verilerde yer alan bazı özellikler şunlardır:

Radius: Tümörün çapı.

Texture: Tümörün yüzey dokusunun ölçüsü.

Perimeter: Tümörün çevresi.

Area: Tümörün alanı.

Smoothness: Tümörün yüzey pürüzlülüğü.

Compactness: Tümörün kompaktlık ölçüsü.

Concavity: Tümörün iç eğriliği.

Symmetry: Tümörün simetrisi.

Bu özellikler, hastaların sağlık durumları hakkında önemli bilgiler sunarak, kanserin iyi huylu ya da kötü huylu olup olmadığını sınıflandırmak için kullanılacaktır.

Proje, makine öğrenmesi kullanarak, kanserin malign (kötü huylu) ya da benign (iyi huylu) olduğu tahminini yapmaya odaklanmaktadır. Bu sınıflandırma, hastaların medikal geçmişlerine ve tümörlerin fiziksel özelliklerine dayalı veriler kullanılarak yapılacaktır. KNN, PCA ve NCA yöntemleri ile en iyi sınıflandırma modelini geliştirmek hedeflenmiştir.

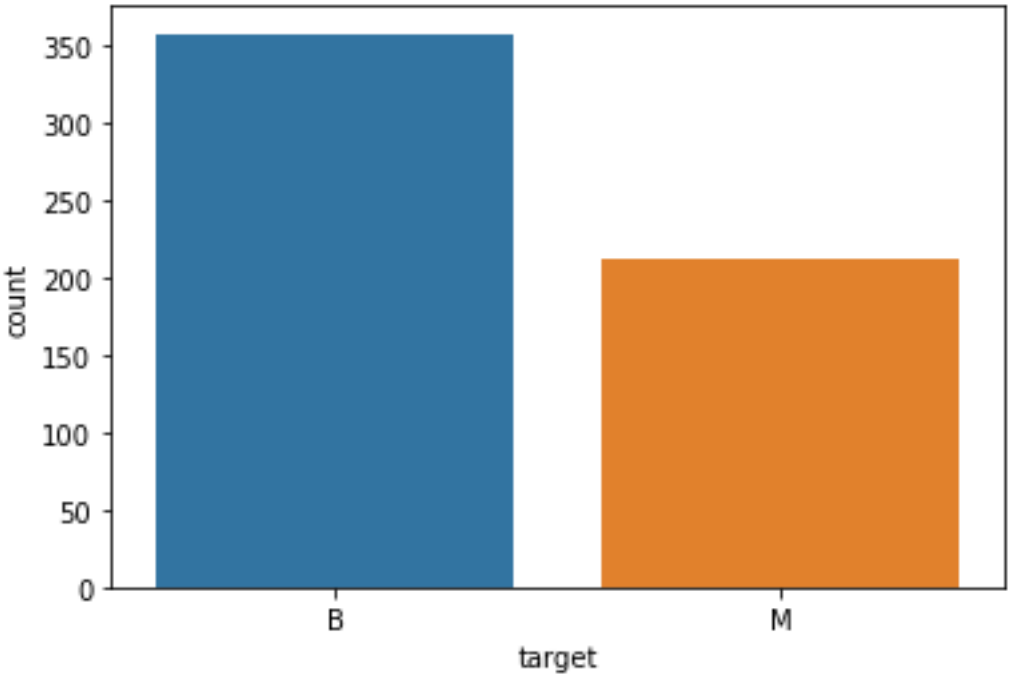
2.2 Veri Seti ve Gerekli Kütüphanelerin Yüklenmesi

Projeye başlamadan önce, **pandas, numpy, seaborn, matplotlib** gibi veri analizi ve görselleştirme için gerekli kütüphaneler ile sklearn kütüphanesinin çeşitli modülleri yüklenmiştir. Verileri okuma ve işleme işlemleri için **pandas** kütüphanesi kullanılmıştır

2.3 Keşifsel Veri Analizi (Exploratory Data Analysis)

Veri setini daha iyi anlayabilmek için keşifsel veri analizi (EDA) yapılmıştır. İlk adımda, veri setindeki kolonlar düzenlenmiş ve gereksiz sütunlar (örneğin, Unnamed: 32, id) silinmiştir. Ayrıca, hedef değişken olan diagnosis kolonunun adı target olarak değiştirilmiş ve target değerleri "M" (malignant) ve "B" (benign) olarak iki sınıfa dönüştürülmüştür.

Bu adımda ayrıca, hedef değişkenin sınıf dağılımı görselleştirilmiş ve veri setindeki sayılar yazdırılmıştır. Verinin iyi huylu ve kötü huylu kanserler arasındaki dağılımı incelenmiş, ardından bazı temel EDA yöntemleri kullanılarak veriler arasındaki ilişkiler, dağılımlar ve korelasyon matrisleri çizilip incelenmiştir.



Şekil 1 : Kanser Sınıflandırması Dağılımı

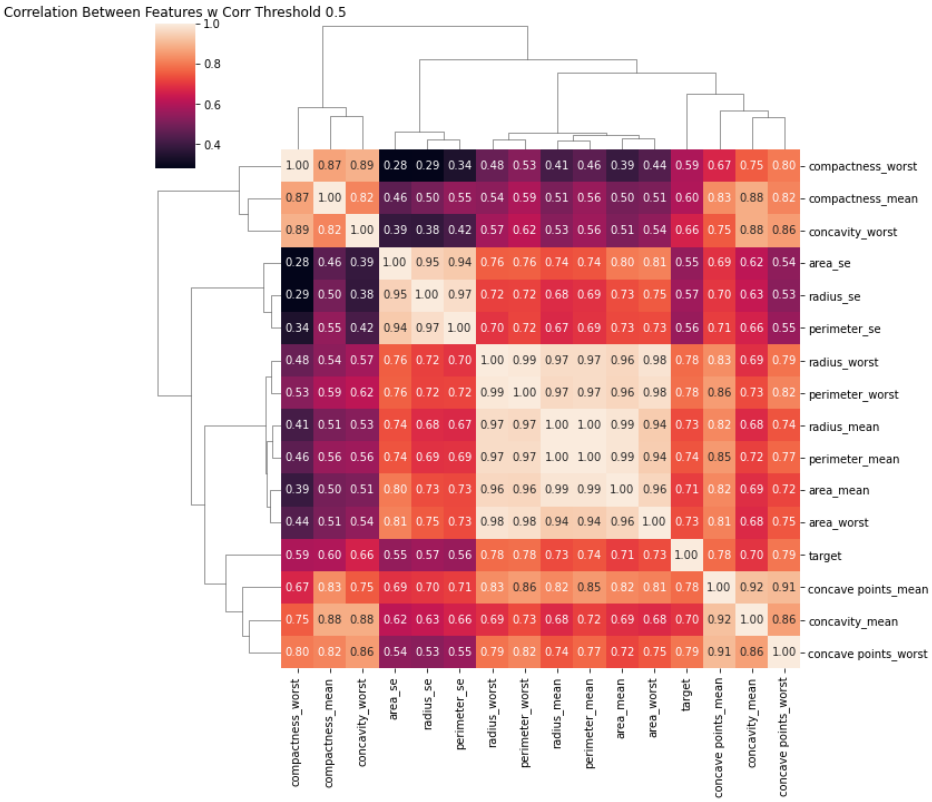
Korelasyon matrisi, iki özellik (feature) arasındaki ilişkinin ölçülmesinde kullanılan bir yöntemdir. Korelasyon değeri 1'e yakınsa, bu iki özellik arasında pozitif doğrusal bir ilişki olduğu anlamına gelir. Eğer korelasyon değeri -1 ise, bu durumda özellikler arasında ters

doğrusal bir ilişki vardır. Korelasyon değeri 0 ise, bu iki özellik arasında herhangi bir ilişki bulunmadığını gösterir.

Bu ölçüm, model geliştirme sürecinde önemli bir rol oynamaktadır. Özellikler arasındaki yüksek korelasyon, bu özelliklerin modelimize benzer şekilde katkı sağladığını gösterir. Bu nedenle, makine öğrenmesi modellerinde çeşitliliği artırmak amacıyla, birbirleriyle yüksek korelasyona sahip olan özelliklerden kaçınılmalıdır. Çeşitlilik, ilişkisi düşük olan özelliklerin seçilmesiyle sağlanabilir. Örneğin, bir korelasyon grafiğinde symmetry_worst ve fractal_dimension_se arasındaki korelasyon değeri 0.11 olarak gözlemlenebilir, bu da bu özelliklerin birbirleriyle zayıf bir ilişkiye sahip olduğunu ve modelde farklı katkılar sağlayabileceğini gösterir.

Bu yaklaşım, modelin performansını artırmaya ve aşırı öğrenme (overfitting) riskini azaltmaya yardımcı olabilir.

Projede bir eşik değeri(0.5) verilerek berlirli değerden yüksek ilişkisi olan özelliklerin korelasyon matrisi çizilmiştir



Şekil 2: Korelasyon Matrisi

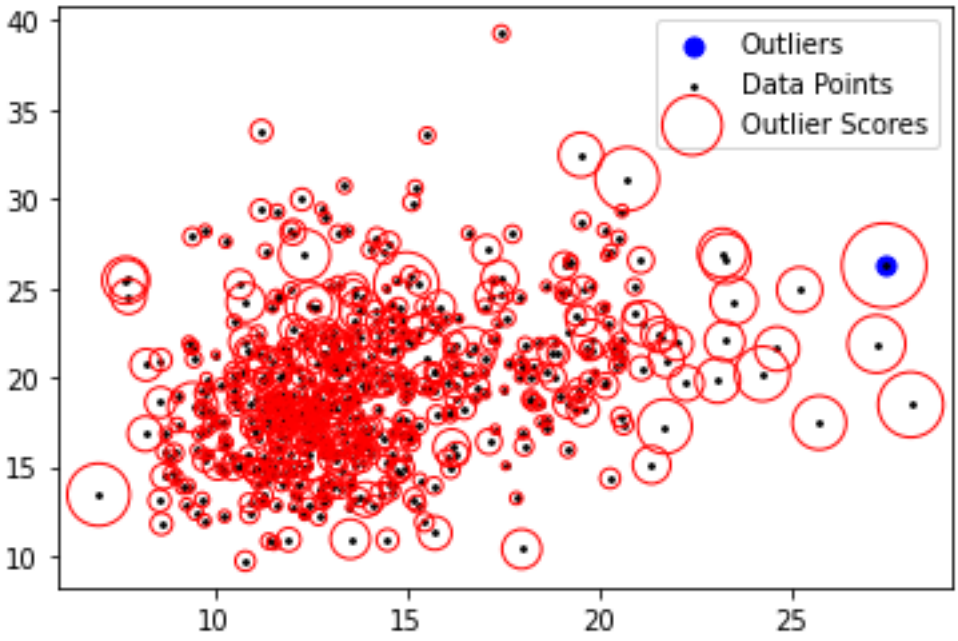
2.4 Aykırı Değer Tespiti (Outlier Detection)

Aykırı değerler (outlier), veri seti içerisinde diğer verilere göre önemli ölçüde farklılık gösteren verilerdir. Bu tür veriler, doğru bir şekilde ayıklanmadıkları takdirde, modelin doğruluğunu olumsuz etkileyebilir ve yanlış yönlendirebilir.

Bu bağlamda, **Density-Based Outlier Detection** (Yoğunluk Tabanlı Aykırı Değer Tespiti) yöntemlerinden biri olarak **Local Outlier Factor** (LOF) kullanılacaktır. LOF, özellikle eğimli (skewed) verilerdeki aykırı değerleri tespit etmek için etkili bir yöntemdir. Çünkü bu tür verilerde, aykırı değerler genellikle veri setinin çoğunluğundan uzak bölgelerde yoğunlaşır ve LOF, bu tür verileri doğru bir şekilde tanımlayabilmektedir. Bu özellik, verimizin özelliklerine uygun olarak, aykırı değerleri başarılı bir şekilde tespit etmeyi sağlar.

Bu algoritma, her bir veri noktasının komşularına olan uzaklığını hesaplar ve bu uzaklıklara göre outlier olup olmadığını değerlendirir. Aykırı değerler, belirli bir eşik değeri (threshold) ile tespit edilmiştir ve gerekirse veri setinden çıkarılmıştır.

Veri setinde yer alan 569 değerin yaklaşık 30 tanesi aykırı değer (outlier) olarak tespit edilmiştir. Bu aykırı değerlerin tamamen silinmesi veri kaybına yol açabilir, bu nedenle bu veriler üzerinde dikkatli bir şekilde işlem yapılmalıdır. Bunun yerine, belirli bir eşik değeri (threshold) tanımlayarak yalnızca bu eşiği aşan aykırı değerleri çıkarma yaklaşımını benimsemek daha uygun olacaktır. Aykırı değerlerin daha iyi anlaşılabilmesi ve görselleştirilmesi amacıyla uygun grafiksel yöntemlerle bu veriler görselleştirilmiştir.



Şekil 3: Aykırı Değerler

2.5 Train Test Split İşlemleri

Modeli geliştirmeden önce, veri seti eğitim (training) ve test (testing) veri setlerine ayrılmıştır. Bu işlem için train_test_split fonksiyonu kullanılmıştır. Eğitim verileri modelin eğitilmesi için, test verileri ise modelin doğruluğunun değerlendirilmesi için kullanılmıştır. Verinin %30'u test, %70'i eğitim için ayırılmıştır.

2.6 Standardizasyon İşlemi

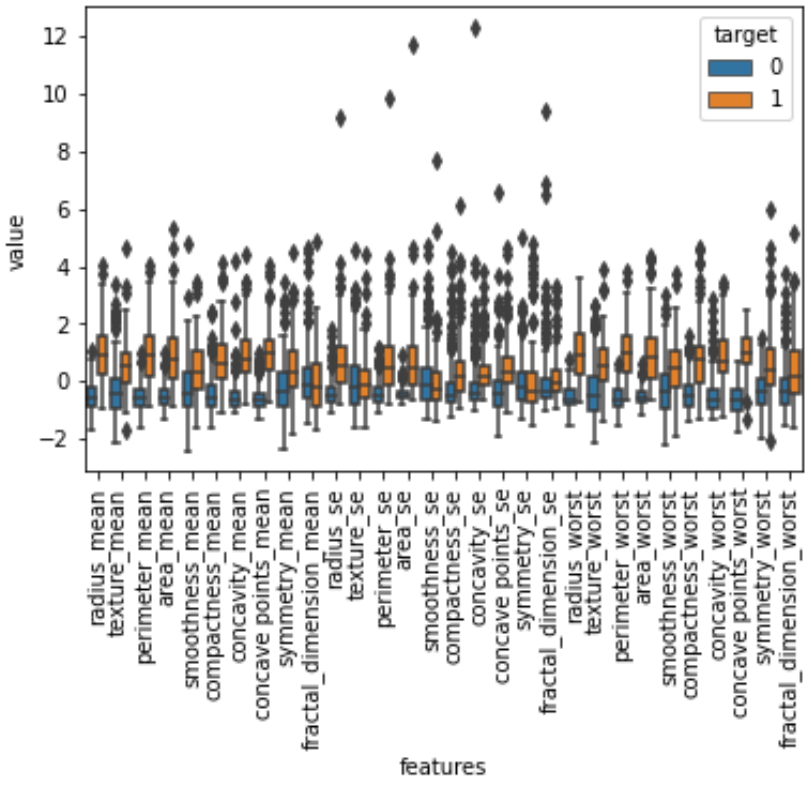
Veri setinde yer alan bazı özelliklerin farklı ölçeklerde olması, makine öğrenmesi modellerinin performansını olumsuz yönde etkileyebilir. Özellikle, bazı algoritmalar, farklı ölçeklerdeki verilerle çalışırken, yüksek değerli özelliklerin modelin karar süreçlerinde daha fazla ağırlık taşımasına neden olabilir. Bu durum, modelin genel doğruluğunu düşürebilir ve öğrenme sürecini zorlaştırabilir.

Bu sorunu ortadan kaldırmak ve tüm özelliklerin eşit bir şekilde model üzerinde etkili olabilmesini sağlamak amacıyla Standardization (Standartlaştırma) işlemi uygulanmıştır. Standardizasyon, her bir özelliğin, ortalama değeri 0 ve standart sapması 1 olacak şekilde dönüştürülmesini sağlar. Bu işlem, StandardScaler gibi bir araç kullanılarak gerçekleştirilmiştir. StandardScaler, her bir özelliğin dağılımını, o özelliğin ortalamasından çıkararak ve ardından bu farkı standart sapmaya bölerek standartlaştırır. Sonuç olarak, tüm özellikler aynı ölçeğe getirilmiş olur, bu da modelin daha verimli bir şekilde öğrenmesini ve daha doğru sonuçlar üretmesini sağlar. Bu işlem, özellikle uzak mesafelerle çalışan algoritmalar (örneğin, K-en Yakın Komşu (KNN), Destek Vektör Makineleri (SVM) gibi) için büyük bir öneme sahiptir.

Yapılan işlemlerden sonra veri setindeki özelliklerin (feature) dağılımlarını görselleştirmek ve anlamak amacıyla boxplot kullanılmıştır. Boxplot, her bir özelliğin, veri setindeki farklı sınıflar (0 ve 1) için dağılımlarını görsel olarak incelememize olanak tanır. Bu görselleştirme, özellikle her sınıfın merkezi eğilimlerini (medyan) ve yayılmalarını (çeyrekler arası aralık) görmeyi sağlar.

Boxplot ayrıca aykırı değerleri (outlier) tespit etmek için etkili bir araçtır. Aykırı değerler, kutu dışındaki noktalar olarak görselleştirilir ve bu sayede veri setindeki olağan dışı gözlemler kolaylıkla fark edilebilir. Aykırı değerlerin varlığı, veri setinin analizine yönelik önemli bir ipucu sunar ve bu değerlerin nasıl işleneceği konusunda kararlar almayı sağlar.

Ayrıca, boxplot yardımıyla hangi özelliklerin sınıflar arasında daha belirgin bir şekilde ayrılabildiğini gözlemleyebiliriz. Özellikler arasındaki dağılımsal farklar, sınıfların birbirinden nasıl ayrılabileceği konusunda bilgi verir. Bu da, modelin öğrenme sürecinde önemli olan özelliklerin seçilmesi için yol gösterir. Özellik çıkarımı (feature extraction) aşamasında, dağılımlarında büyük farklar gözlemlenmeyen ve sınıfları birbirinden iyi ayırt edemeyen özellikler, modelin performansını olumsuz etkilememek için genellikle dışarıda bırakılır. Bu sayede, yalnızca modelin doğruluğunu artırabilecek, anlamlı özellikler üzerinde yoğunlaşılır.



Şekil 4: Boxplot

2.7 KNN Algoritmasının Uygulanması

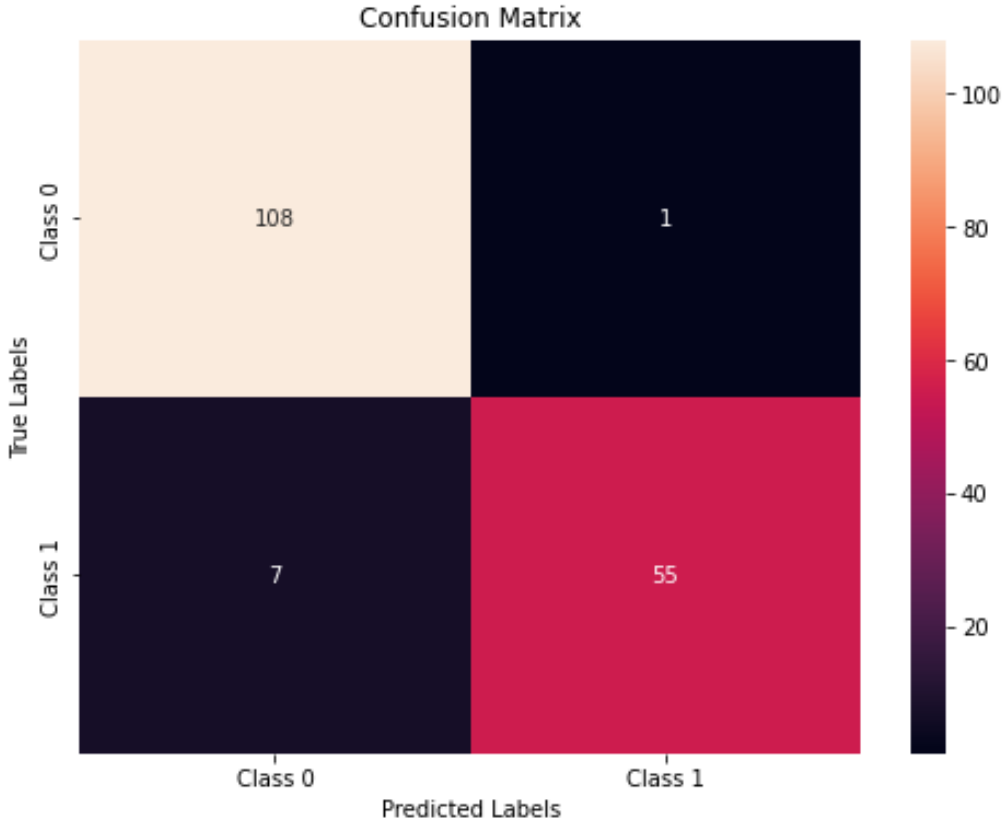
K-en Yakın Komşu (KNN) algoritması, bu çalışmada modelin temel sınıflandırma algoritması olarak kullanılmıştır. KNN, sınıflandırma görevlerinde sıklıkla tercih edilen bir algoritmadır. Bu algoritma, bir veri noktasının sınıfını belirlerken, en yakın k komşusunun sınıfına bakarak karar verir. Komşuluk, genellikle Euclidean mesafesi gibi bir uzaklık ölçütü ile hesaplanır.

Bu çalışmada, KNN algoritması başlangıçta n_neighbors=2 parametresi ile uygulanmıştır. Burada n_neighbors, sınıflandırma kararının verileceği en yakın iki komşu noktasının sayısını ifade etmektedir. Bu parametre, modelin karar verme sürecinde dikkate alınacak komşu sayısını belirler ve bu sayede modelin genel performansı üzerinde önemli bir etkisi vardır.

KNN Algoritmasının Tercih Edilme Sebepleri:

- **Eğitim Gerektirmez:** Hızlı çalışır, çünkü veri seti üzerinde eğitim yapılmaz.
- **Kolay Uygulama:** Basit implementasyon ile hızlıca uygulanabilir.
- **Kolay Parametre Ayarlaması:** k değeri gibi parametreler kolayca ayarlanabilir ve optimize edilebilir.

KNN yöntemi uygulandıktan sonra **%95,32** değerinde bir doğruluk oranı elde edilmiştir. Yöntem sonucu elde edilen karmaşıklık matrisi Şekil 5'te gösterilmiştir.



Şekil 5: Knn Sonucu Karmaşıklık Matrisi

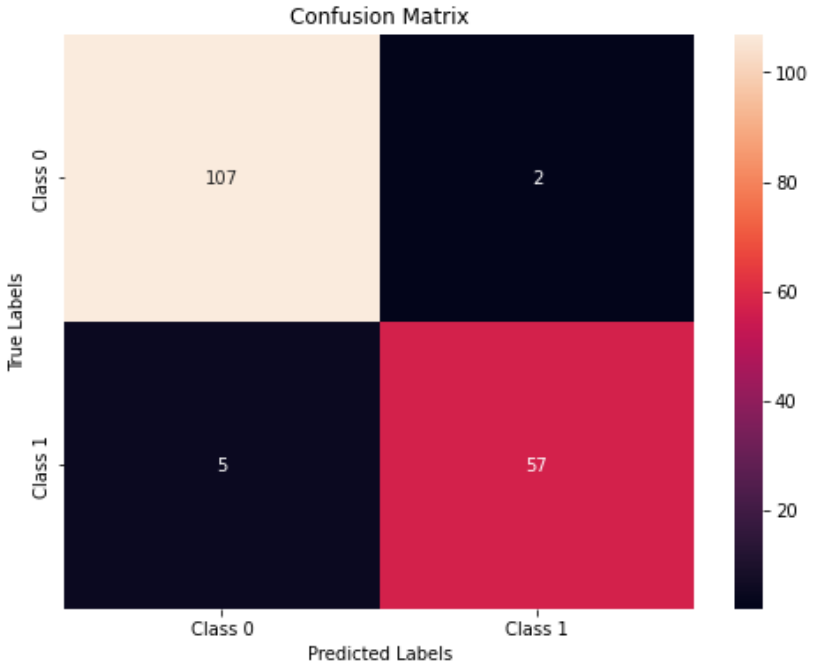
2.8 KNN için En İyi Parametreleri Tespit Etme

KNN algoritmasında, en uygun k değerinin ve ağırlık fonksiyonunun (weights) belirlenmesi önemlidir. Bunun için Çapraz Doğrulama (GridSearchCV) kullanılarak farklı k değerleri ve ağırlık fonksiyonları test edilmiştir. Bu işlem, modelin en iyi parametrelerle eğitilmesini sağlamaktadır.

Çapraz doğrulama ile overfitting ve underfitting riskleri değerlendirilmiştir.

KNN ile **%95,32** değerinde bir doğruluk elde edilirken En iyi parametreleri tespit etme sonucunda **%95,90** değerinde bir doğruluk oranı elde edilmiştir.

En iyi parametlerin tespiti sonucu oluşan karmaşıklık matrisi Şekil 6'da gösterilmiştir.

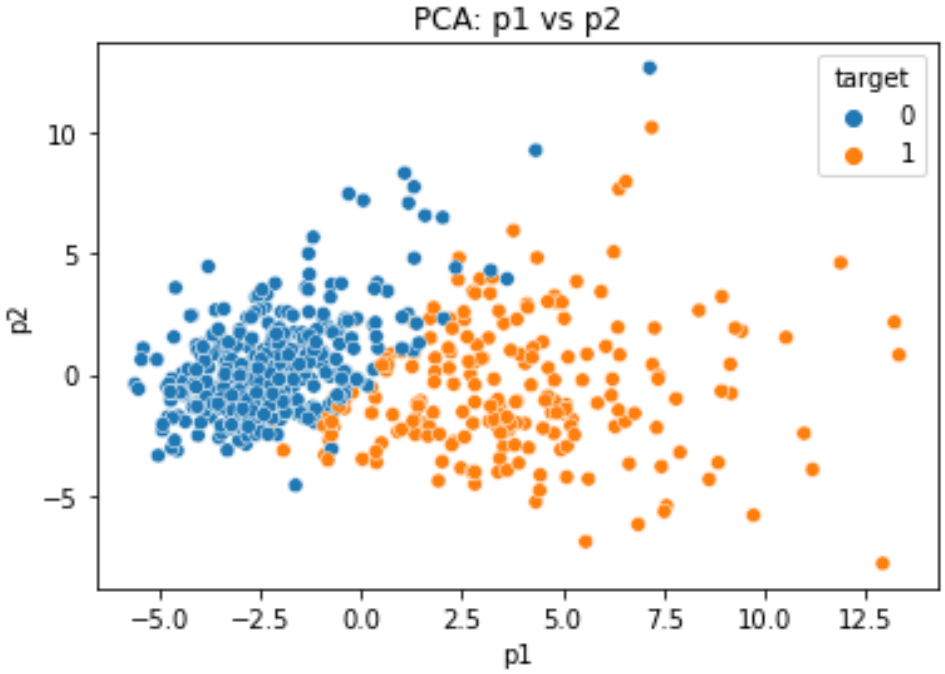


Şekil 6: En İyi Parametreler Sonucu Karmaşıklık Matrisi

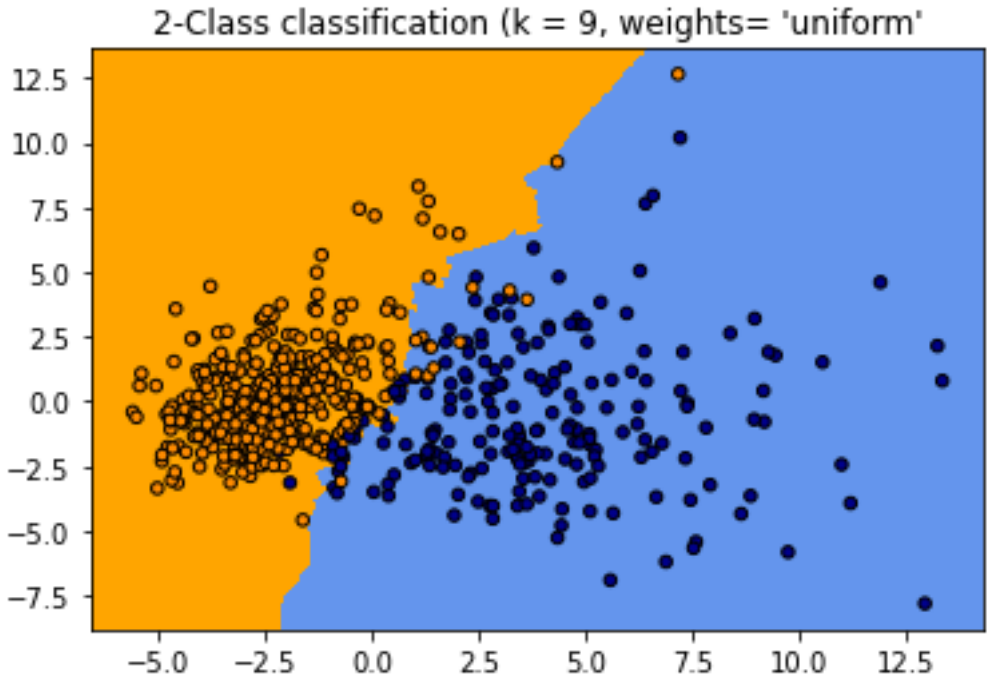
2.9 Principal Component Analysis (PCA) İşlemi

Veri setindeki boyutları (özellikleri) azaltmak amacıyla Principal Component Analysis (PCA) tekniği uygulanmıştır. PCA, yüksek boyutlu veriyi daha düşük boyutlara indirgemek için kullanılan istatistiksel bir yöntemdir. Bu işlem, verinin en fazla varyans gösteren yönlerini (ana bileşenlerini) belirleyerek, özellikler arasındaki korelasyonu azaltır. Bu sayede, modelin eğitilmesi hızlanır ve daha az hesaplama kaynağı gerektirir. Ayrıca, boyut indirgeme, modelin aşırı öğrenme (overfitting) sorununu azaltmaya da yardımcı olabilir. PCA, büyük veri setlerinde performans iyileştirmesi sağlamak için yaygın olarak tercih edilmektedir.

PCA sonucu 30 boyutlu verinin 2 boyuta indirgendikten sonra görselleştirilmesi Şekil 7 ve Şekil 8'de gösterilmiştir.



Şekil 7: PCA Sonucu Verinin Görselleştirilmesi (1)



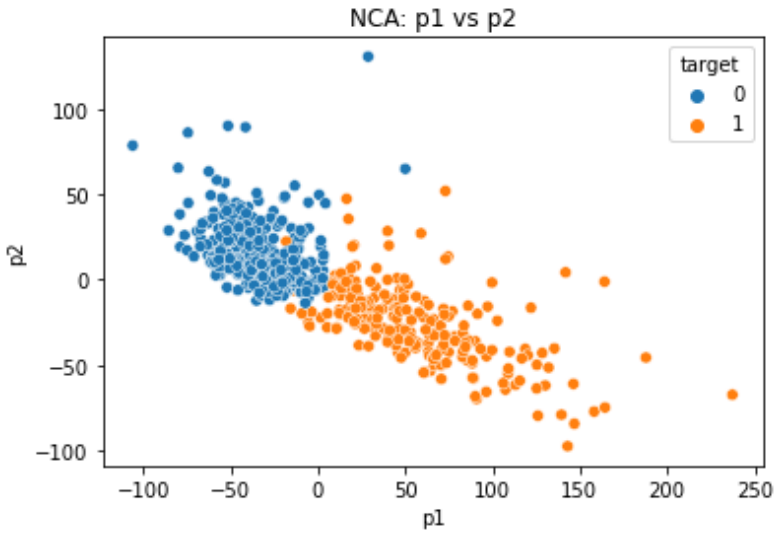
Şekil 8: : PCA Sonucu Verinin Görselleştirilmesi (2)

PCA sonucu modelimizin doğruluk oranı **%92,40** olmuştur.

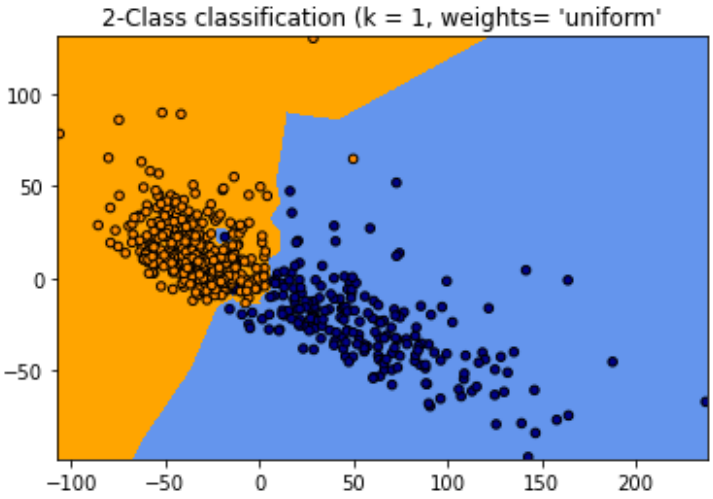
2.10 Neighborhood Components Analysis (NCA) İşlemi

NCA (Neighborhood Components Analysis), özellikle KNN algoritması ile birlikte kullanılan bir boyut indirgeme yöntemidir. Bu teknik, veri setindeki önemli özellikleri öğrenerek, KNN modelinin daha etkili çalışmasını sağlar. NCA, verinin daha anlamlı bir şekilde temsil edilmesine olanak tanır ve sınıflandırma performansını iyileştirir. Bu yöntem, boyutları azaltarak modelin doğruluğunu optimize eder, böylece daha hızlı ve verimli sonuçlar elde edilir. KNN algoritmasının performansını artırmak için NCA, özellikle büyük veri setlerinde faydalı bir araçtır.

NCA sonucu verinin görselleştirilmesi Şekil 9 ve Şekil 10'da gösterilmiştir.

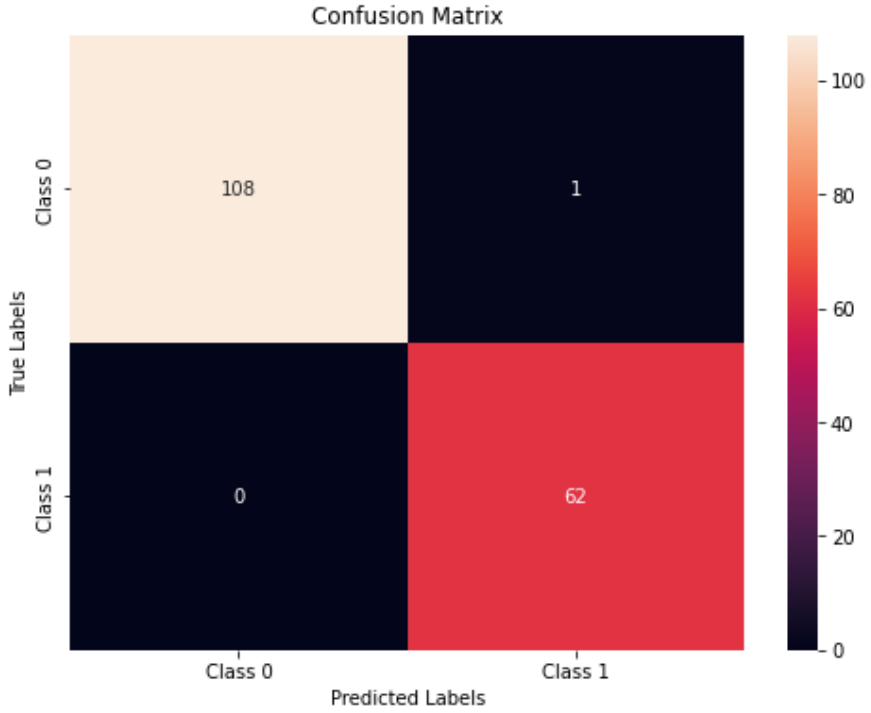


Şekil 9: NCA Sonucu Verinin Görselleştirilmesi (1)



Şekil 10: NCA Sonucu Verinin Görselleştirilmesi (2)

NCA sonucu elde edilen karmaşıklık matrisi Şekil 11'de gösterilmiştir.



Şekil 11: NCA Sonucu Karmaşıklık Matrisi

NCA (Neighborhood Components Analysis) uygulandıktan sonra, modelin doğruluk oranı **%99,41**'e yükselmiş ve bu, yüksek bir başarı seviyesini temsil etmektedir. Bu sonuç, NCA'nın boyut indirgeme sürecinde modelin performansını önemli ölçüde iyileştirdiğini göstermektedir.

2.11 F-1 Skoru:

Karmaşıklık matrisinin elemanlarının anlamı:

True Negatives (TN) (Doğru şekilde negatif tahmin edilen): 108

False Positives (FP) (Yanlış şekilde pozitif tahmin edilen): 1

False Negatives (FN) (Yanlış şekilde negatif tahmin edilen): 0

True Positives (TP) (Doğru şekilde pozitif tahmin edilen): 62

2.11.1 Precision ve Recall Hesaplama:

- **Precision (P):** P = TP/(TP+FP) => 62/62+1 = 62/63 ≈ 0.9841
- **Recall (R):** R = TP/(TP+FN) = 62/(62+0) = 62/62 = 1.0
- **F1 Skoru:** F1 = 2* ((P*R)/(P+R)) = 2*((0.9841*1)/(0.9841+1)) ≈ 0.9919

Modelinizin **F1 skoru ≈ 0.9919** olarak hesaplanmıştır.

2.12 Sonuçlar ve Karşılaştırmalar

Ravdin ve Clark düşük ve yüksek risk taşıyan göğüs kanseri hastalarının tespitine yönelik bir yapay sinir ağı modeli ortaya koymuşlardır [4]. Mangasarian vd. ise kötü huylu tümörler için tekrarlamayan vakaları, tekrarlayan vakalar için ise tekrarlama zamanlarını tahmin etmeye yönelik doğrusal programlama tabanlı bir sistem geliştirmişlerdir [5]. Ravi ve Zimmermann tümör veri seti üzerinde geliştirdikleri üç fazlı bulanık veri işleme modelinde önce özellik uzayında boyut indirgeme yoluna gitmiş, ardından bulanık kuralları otomatik olarak oluşturup daha az kuralla daha yüksek bir sınıflama gücü elde etmişlerdir [6]. Delen vd. geniş bir göğüs kanseri veri seti üzerinde iki popüler veri madenciliği algoritması olan yapay sinir ağları ve karar ağaçlarını kullanarak tahmin modelleri geliştirmişlerdir. Karar ağaçları ile **%93.6**, yapay sinir ağı modeli ile **%91.2** doğruluk elde etmişlerdir [7]. Polat ve Güneş, En küçük kare destek vektör makinesi (LS-SVM) sınıflama algoritması kullanarak göğüs kanseri verileri üzerinde %98 oranında başarı elde etmişlerdir [8]. Khan vd. geliştirdikleri bulanık karar ağaçları ile göğüs kanseri verilerini sınıflamış ve bağımsız sınıflayıcılara göre daha başarılı olduklarını ortaya koymuşlardır [9]

Chauhan vd. yapay sinir ağlarında parametre ayarlamaları için diferansiyel evrim modeli kullanarak gerçekleştirdikleri sistemi, göğüs kanseri veri seti dâhil üç farklı veri seti ile test ederek, geleneksel yapay sinir ağı modelinden daha başarılı olduğunu ortaya koymuşlardır [10]. Karabatak ve İnce, ilişki kuralları ile yapay sinir ağlarını birleştirerek ürettikleri hibrit model ile göğüs kanseri verilerini sınıflamışlar ve modellerinin **%95.6** oranda doğru sınıflama yaptığını ortaya koymuşlardır [11]. Powell vd. Kaliforniya'da yaşayan ve içerisinde yüksek oranda göğüs kanseri olan, hiç doğum yapmamış ve geç doğum yapmış kadınlara ait veriler içeren veri seti ile Breast Cancer Risk Assessment Tool (BCRAT) [12], International Breast Intervention Study (IBIS) [13] ve BRCAPRO [14] göğüs kanseri risk değerlendirme modellerini 5 yıl boyunca yaptıkları uygulamalarla karşılaştırmışlardır ve performanlarını test etmişlerdir.[15].

Papageorgiou vd. çalışmalarında Fuzzy Cognite Map (FCM) kullanarak geliştirdikleri bir sağlık asistanı ile 40 adet hastanın verilerini işleyerek **%95** oranında doğruluk elde etmişlerdir [16]. Kolay ve Erdoğmuş çalışmalarında göğüs kanseri veri setini herhangi bir ön işleme yapmadan, Matlab ve Weka programları üzerinde K-means yöntemi ile sınıflandırmış ve çeşitli parametre değişimleri ile **%45** ile **%79** arasında değişen başarılar elde etmişlerdir [17]. Alharbi ve Tchier, yaptıkları çalışmada bulanık mantık ve evrimsel

genetik algoritma tabanlı göğüs kanserinin erken teşhisine yardımcı olan bir sistem geliştirerek Suudi Arabistan göğüs kanseri teşhis veri tabanı üzerinde uygulamışlardır. Bulanık-genetik hibrit algoritma ile **%97** doğruluk ve **%91** güvenilirlik elde etmişlerdir [18]. Akyol, 2018 yılında yaptığı çalışmada Özyinelemeli Özelik Elemesi yöntemi ile meme kanseri veri seti üzerinde öznitelik tespiti yapıp rastgele orman yöntemini uygulayarak sınıflama yapmış ve **%98** başarı elde etmiştir [19].

Bu çalışmada her biri 30 adet özellik içeren 569 göğüs kanseri veri örneği makine öğrenmesi teknikleri ile sınıflandırılarak, modellerin başarıları karşılaştırılmıştır. KNN ile **%95,32** başarı elde edilmiştir KNN en iyi parametreler seçimi sonucu **%95,90** başarı elde edilmiştir. PCA yöntemi sonucu başarı oranı **%92,40** olmuştur. Son olarak kullanılan NCA yöntemi sonucu başarı oranı **%99,41** olmuştur

3. MATERYAL VE METOT

Bu çalışmada, KNN algoritmasının uygulanması için kullanılan materyaller ve yöntemler aşağıda belirtilmiştir.

3.1 Materyal

Çalışmada kullanılan veriler, belirli bir sınıflandırma problemine ilişkin bir veri seti ile sağlanmıştır. Veri seti, genellikle çeşitli özellikler (features) içeren ve sınıf etiketleri (labels) ile etiketlenmiş örneklerden oluşmaktadır. Bu örnekler, modelin eğitilmesi ve test edilmesi için kullanılmaktadır.

Kullanılan yazılım araçları şunlardır:

- **Python:** Veri işleme, model oluşturma ve değerlendirme işlemleri Python programlama dili ile gerçekleştirilmiştir.
- **Scikit-learn:** KNN algoritmasının uygulanması için Scikit-learn kütüphanesi kullanılmıştır. Bu kütüphane, makine öğrenmesi algoritmalarını kolayca uygulamaya imkan tanır.
- **Pandas ve Numpy:** Veri analizi ve işleme için Pandas ve Numpy kütüphaneleri kullanılmıştır.
- **Matplotlib ve Seaborn:** Verilerin görselleştirilmesi ve modelin değerlendirilmesi için Matplotlib ve Seaborn kütüphanelerinden yararlanılmıştır.

3.2 Metot (Yöntem)

- **Veri hazırlığı:** İlk adımda, veriler analiz edilip ön işleme tabi tutulmuştur. Verilerin eksik değerleri tamamlanmış, kategorik veriler sayısal verilere dönüştürülmüş ve veriler uygun şekilde ölçeklendirilmiştir. Eğitim ve test veri setleri, genellikle %70 eğitim, %30 test oranı ile bölünmüştür.

- **Modelin Eğitilmesi:** KNN algoritması, K değeri seçilerek eğitilmiştir. KNN modelinin başarısı, farklı K değerleriyle test edilip, doğruluk oranları karşılaştırılarak ölçülmüştür. Euclidean mesafesi gibi mesafe ölçütleri kullanılarak her test örneği için en yakın K komşusu bulunmuştur.

- **Modelin Değerlendirilmesi:** Modelin başarımı, doğruluk (accuracy) oranı, kesme noktası (threshold), ve karmaşıklık matrisi (confusion matrix) ile değerlendirilmiştir. Doğruluk, doğru sınıflandırılan örneklerin toplam örneklere oranı olarak hesaplanmıştır. Karmaşıklık matrisi, modelin doğru ve yanlış sınıflandırmalarını görselleştirerek performansı daha ayrıntılı bir şekilde değerlendirmeyi sağlamaktadır.

- **Sonuçların Yorumlanması:** Sonuçlar, elde edilen doğruluk oranları ve karmaşıklık matrisi üzerinden analiz edilmiştir. Ayrıca, farklı K değerlerinin model başarısına etkisi gözlemlenmiş ve en uygun K değeri belirlenmiştir.

3.3 Başarı Ölçütleri

Çalışmanın başarısını ölçmek için şu kriterler kullanılmıştır:

- **Doğruluk (Accuracy):** Modelin doğru sınıflandırma oranı. **(%99,41)**
- **Kesme Noktası (Thresholding):** Modelin sınıflandırma kararlarını belirlemek için kullanılan sınır değeri.
- **Karmaşıklık Matrisi (Confusion Matrix):** Doğru ve yanlış sınıflandırmaların görselleştirilmesi. [Şekil 5, Şekil 6, Şekil 11]
- **F1 Skoru:** Precision ve recall değerlerinin harmonik ortalaması, modelin dengeyi ne kadar iyi sağladığını gösterir. **(%99.54)**

Bu materyaller ve metotlar kullanılarak KNN algoritmasının başarısı değerlendirilmiş ve sınıflandırma problemlerine çözüm üretme kapasitesi ölçülmüştür.

4.Elde Edilen Deneysel Çalışmalar

Bu çalışmada, farklı boyut indirgeme ve sınıflandırma tekniklerinin veri setine uygulanmasıyla elde edilen deneysel sonuçlar detaylı bir şekilde incelenmiştir. İlk olarak, KNN uygulanarak %95,32 doğruluk oranı elde edilmiştir. Sonrasın Principal Component Analysis (PCA) kullanılarak veri setindeki boyutlar indirgenmiş ve bu işlem, modelin eğitim süresini kısaltırken, aynı zamanda doğruluğunu azaltmıştır. PCA'nın, verinin ana bileşenlerini belirleyerek sınıflandırma performansını optimize ettiği gözlemlenmiştir.

Bunun ardından, Neighborhood Components Analysis (NCA) tekniği uygulanmıştır. NCA, özellikle KNN algoritması ile birlikte kullanılarak, veri setindeki önemli bileşenleri öğrenmeye ve modelin doğruluğunu daha da artırmaya odaklanmıştır. Bu teknik, verinin daha anlamlı bir şekilde temsil edilmesine olanak tanıyarak, sınıflandırma doğruluğunu %99,41 gibi yüksek bir seviyeye çıkarmıştır.

Yapılan deneylerde, NCA'nın KNN algoritması üzerinde büyük bir iyileşme sağladığı ve boyut indirgeme işlemi ile birlikte modelin daha verimli ve doğru sonuçlar verdiği ortaya çıkmıştır. Ayrıca, her iki boyut indirgeme tekniği, modelin işlem yükünü azaltarak daha hızlı tahminler yapmasını sağlamıştır. Bu deneysel sonuçlar, kullanılan boyut indirgeme yöntemlerinin, modelin genel başarısını artırmada ne kadar etkili olduğunu ve doğru parametre ayarlarının model performansını nasıl optimize edebileceğini açıkça ortaya koymaktadır.

Sonuç olarak, elde edilen deneysel bulgular, boyut indirgeme ve sınıflandırma tekniklerinin birlikte kullanıldığında, modelin doğruluğunu ve işlem verimliliğini önemli ölçüde iyileştirdiğini göstermektedir.

5. KAYNAKÇA

[1] Jemal, A., Siegel, R., Xu, J., Ward, E., Cancer statistics 2010, CA: a cancer journal for clinicians, 60(5), 277-300, 2010.

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians, 68(6), 394- 424, 2018

[3] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[4] Ravdin, P. M., Clark, G. M., A practical application of neural network analysis for predicting outcome of individual breast cancer patients, Breast cancer research and treatment, 22(3), 285-293, 1992.

[5] Mangasarian, O. L., Street, W. N., Wolberg, W. H., Breast cancer diagnosis and prognosis via linear programming, OperationsResearch, 43(4), 570-577, 1995.

[6] Ravi, V., Zimmermann, H. J., Fuzzy rule based classification with FeatureSelector and modified threshold accepting, European Journal of Operational Research, 123(1), 16-28, 2000.

[7] Delen, D., Walker, G., Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127, 2005.

[8] Polat, K., Güneş, S., Breast cancer diagnosis using least square support vector machine, Digital signal processing, 17(4), 694- 701, 2007.

[9] Khan, M. U., Choi, J. P., Shin, H., Kim, M., Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, In Engineering in Medicine and Biology Society 30th Annual International Conference of the IEEE , 5148-5151, 2008.

[10] Chauhan, N., Ravi, V., Chandra, D. K., Differential evolution trained wavelet neural networks: Application to bankruptcy prediction in banks, Expert Systems with Applications, 36(4), 7659-7665, 2009.

[11] Karabatak, M., Ince, M. C., An expert system for detection of breast cancer based on association rules and neural network, Expert systems with Applications, 36(2), 3465-3469, 2009.

[12] Costantino, J. P., Gail, M. H., Pee, D., Anderson, S., Redmond, C. K., Benichou, J., Wieand, H. S., Validation studies for models projecting the risk of invasive and total breast cancer incidence, Journal of the National Cancer Institute, 91(18), 1541-1548, 1999.

[13] Tyrer, J., Duffy, S. W., Cuzick, J., A breast cancer prediction model incorporating familial and personal risk factors, Statistics in medicine, 23(7), 1111-1130, 2004.

[14] Parmigiani, G., Berry, D. A., Aguilar, O., Determining carrier probabilities for breast cancer–susceptibility genes BRCA1 and BRCA2, The American Journal of Human Genetics, 62(1), 145-158, 1998.
[15] Powell, M., Jamshidian, F., Cheyne, K., Nititham, J., Prebil, L. A., Ereman, R., Assessing breast cancer risk models in Marin County, a population with high rates of delayed childbirth, Clinical breast cancer, 14(3), 212-220, 2014.
[16] Papageorgiou, E. I., Jayashree Subramanian, Karmegam, A., & Papandrianos, N., A risk management model for familial breast cancer: A new application using Fuzzy Cognitive Map method, Computer Methods and Programs in Biomedicine, 122(2), 123– 135, 2015
[17] Kolay, N., Erdoğmuş, P., The classification of breast cancer with Machine Learning Techniques. In Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 1-4, 2016.
[18] Alharbi, A., Tchier, F., Using a genetic-fuzzy algorithm as a computer aided diagnosis tool on Saudi Arabian breast cancer database, Mathematical biosciences, 286, 39-48, 2017.
[19] AKYOL, K., Meme Kanseri Tanısı İçin Özniteliklerin Öneminin Değerlendirilmesi Üzerine Bir Çalışma, Academic Platform Journal of Engineering and Science, 6(2), 109–115, 2018.