**logo, metin, simge, sembol, kırpıntı çizim içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**T.C.**

**ONDOKUZ MAYIS ÜNİVERSİTESİ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ BİLGİSAYAR MÜHENDİSLİĞİ**

**ÖRÜNTÜ TANIMA-BM605**

**Beast Cancer Veri Seti ile Analiz Raporu**

**FIRAT KAAN BİTMEZ - 281855**

**SAMSUN, 2024-2025 Eğitim Öğretim Yılı Güz Yarıyılı**

**1. Giriş**

Bu çalışmamızda, UCI Machine Learning Repository sitesinden alınan Breast Cancer Wisconsin (Diagnostic) veri seti üzerinde uygulanan üç farklı boyut indirgeme yönteminin ve beş farklı sınıflandırma algoritmasının performansını analiz etmektedir. Kullanılan boyut indirgeme yöntemleri şunlardır:

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* t-Distributed Stochastic Neighbor Embedding (t-SNE)

Matematiksel altyapılar, sınıflandırma performansları ve görselleştirmeler detaylı olarak ele alınmıştır. Çalışma, Python programlama dili kullanılarak veri işleme, model eğitimi ve değerlendirme adımlarıyla gerçekleştirilmiştir. Ayrıca, bu raporda boyut indirgeme tekniklerinin matematiksel temellerine, kullanılan sınıflandırma algoritmalarının avantaj ve dezavantajlarına da detaylı bir şekilde değinilmiştir.

Boyut indirgeme yöntemlerinin sınıflandırma performansına etkilerini inceleyerek, hangi yöntemin belirli veri türlerinde daha iyi sonuç verdiğini ortaya çıkarmayı amaçlamaktadır. Veri seti üzerinde yapılan analizler sonucunda, boyut indirgeme işlemlerinin sınıflandırma performansını nasıl etkilediği gözlemlenmiştir. Python programlama dili kullanılarak veri işleme, model eğitimi ve değerlendirme adımları gerçekleştirilmiştir. Bu rapor, ilgili tekniklerin matematiksel temellerine de değinerek çalışmanın teorik altyapısını da sunmaktadır. Ayrıca bu raporda, kullanılan sınıflandırma algoritmalarının avantajları ve sınırlamaları da tartışılmıştır.

Boyut indirgeme teknikleri, yüksek boyutlu veri kümelerinde, özellikle verinin görselleştirilmesi ve sınıflandırma modellerinin daha verimli hale getirilmesi açısından büyük önem taşımaktadır. Bu nedenle, PCA, LDA ve t-SNE gibi tekniklerin hem teorik altyapıları hem de uygulamada sağladıkları avantajlar detaylı olarak incelenmiştir. Her bir teknik, verideki boyutların azaltılmasıyla model performansını artırmak, modelin eğitilme süresini kısaltmak ve verideki gürültüyü azaltarak sınıflar arasındaki ayrımı daha net hale getirmek amacıyla kullanılmıştır.

**2. Kullanılan Veri Seti**

Kaynak: UCI Machine Learning Repository - Breast Cancer Wisconsin (Diagnostic)

- Veri Seti Bağlantısı: [[Breast Cancer Wisconsin (Diagnostic)]](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)

- Amaç: Meme kanseri teşhisinde iyi huylu (Benign - B) ve kötü huylu (Malignant - M) tümörleri sınıflandırmak.

- Veri Seti: Toplamda 569 örnek ve 32 sütundan oluşmaktadır. Bu sütunlardan ID ve Diagnosis (etiket) sütunları dışında, geri kalan 30 özellik tümör özellikleriyle ilgilidir. Diagnosis sütunu iyi huylu tümörleri 0, kötü huylu tümörleri 1 olarak etiketlenmiştir.

Veri setindeki özelliklerden bazıları şunlardır:

- ID: Hastaya atanmış bir kimlik numarasıdır ve analiz için gerekli değildir. Sadece hangi hasta olduğu belirlemek için kullanmaktadır.

- Diagnosis: Teşhis sonucunu göstermek için kullanılır analiz sonucunu yansıtmak için gereklidir.İki tip Diagnosis vardır: iyi huylu (Benign) veya kötü huylu (Malignant).

- Radius: Hücre çekirdeği çapı

- Texture: Piksel yoğunluklarının varyasyonu

- Perimeter: Hücre çevresi

- Area: Hücre alanı

- Smoothness: Çekirdek sınırlarının düzgünlüğü

- Compactness: (Perimetre² / Alan) - 1

- Concavity: Çekirdeğin dışbükey bölgelerinin derinliği

- Concave Points: Çekirdek sınırındaki içbükey noktalar

- Symmetry: Çekirdek simetrisi

- Fractal Dimension: Çekirdek yüzey karmaşıklığı

Veri setindeki özellikler, meme kanseri tümörlerinin çeşitli karakteristiklerini ifade eder ve bu özelliklerin bazısı hücre çekirdeğinin fiziksel büyüklüklerine (alan, çevre vb.) dayanmaktadır. Bu özellikler kullanılarak meme kanserinin iyi huylu mu yoksa kötü huylu mu olduğu tespit edilmeye çalışılmıştır. Veri seti, makine öğrenmesi algoritmalarının eğitimi ve test edilmesi için dengeli bir yapıya sahiptir. İyi huylu tümör sayısı 357 iken kötü huylu tümör sayısı 212'dir, bu da veri setinin dengesizliğini gideren uygun bir oran sağlar.

Veri seti üzerinde gerçekleştirilen analizlerde, her bir tümör örneğinin çeşitli ölçütlerle ifade edilmesi, özellikle sınıflandırıcıların daha yüksek doğruluk oranlarına ulaşabilmesi açısından önem taşımaktadır. Bu özelliklerin birçoğu, hem ortalama değerler, hem standart sapmalar hem de en kötü değerler üzerinden hesaplanmış ve bu da veri setini çok boyutlu hale getirmiştir.

**3. Veri Ön İşleme**

Verilerin farklı ölçeklerde olması, sınıflandırıcıların performansını olumsuz etkileyebilir. Bu yüzden tüm özellikler StandardScaler kullanılarak ölçeklendirilmiştir. Veri ön işleme, makine öğrenmesi modellerinin doğru çalışması ve yanlış öğrenmeyi önlemek için kritik bir aşama olarak kabul edilmektedir.

- Veri Yüklenmesi ve Özelliklerin Düzenlenmesi:

- Veri seti CSV formatında yüklenmiş, ID sütunu analiz için anlam taşımadığından kaldırılmıştır.

- Diagnosis etiketi 0 (Benign) ve 1 (Malignant) olacak şekilde kodlanmıştır. Bu, sınıflandırma işlemlerinde kullanılacak hedef değişkenin belirlenmesini sağlar.

- Özelliklerin Ölçeklendirilmesi:

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(features)

joblib.dump(scaler, os.path.join(models\_dir, "scaler.pkl"))

- Tüm özellikler, ortalaması 0 ve standart sapması 1 olacak şekilde standartlaştırılmıştır. Bu, algoritmaların çalışmasının daha verimli ve doğru olmasına yardımcı olur. Aynı zamanda, veri setinin düzgün bir şekilde işlenmesi, öğrenme sürecinde daha iyi performans elde edilmesine olanak tanır.

Veri ön işleme aşamasında veri setinin temizlenmesi, özelliklerin uygun şekilde seçilmesi ve ölçeklendirilmesi, veri setinin daha anlamlı hale getirilmesi açısından önemlidir. Bu aşamalar, makine öğrenmesi modelinin başarısını doğrudan etkileyen faktörler arasında yer alır.

**4. Boyut İndirgeme Teknikleri**

Bu çalışma kapsamında özellikle sınıflandırma problemleri üzerinde çalışırken kullanılan üç farklı boyut indirgeme tekniği uygulandı: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) ve t-Distributed Stochastic Neighbor Embedding (t-SNE). Bu teknikler, veri setinin çok boyutluluğunu azaltarak hem sınıflandırma performansını iyileştirmek hem de verilerin daha kolay anlaşılmasını sağlamak için kullanılmıştır. Her bir boyut indirgeme tekniğinin matematiksel altyapısı, özellikleri ve uygulama sonrası elde edilen çıktılar aşağıda detaylı olarak ele alınmıştır. Ayrıca bu tekniklerin avantajları, dezavantajları ve belirli veri setlerinde nasıl daha verimli kullanılabileceği konusuna da değinilmiştir.

**4.1 Principal Component Analysis (PCA)**

Principal Component Analysis (PCA), çok boyutlu verilerdeki varyansı maksimize ederek boyut indirgeme işlemi yapar. PCA, orijinal veride bulunan özelliklerin lineer kombinasyonlarından oluşan yeni bir özellik uzayı oluşturur ve bu yeni uzaydaki bileşenler orijinal verideki en fazla varyansı taşıyacak şekilde seçilir. PCA kullanılmasındaki temel amaç, veri setindeki özellik sayısını azaltarak hem hesaplama maliyetini düşürmek hem de modelin aşırı uyum problemini engellemektir. PCA, özellikle yüksek boyutlu verilerde boyut indirgeme yaparak, modelin daha hızlı eğitilmesine ve daha az bellek kullanmasına olanak tanır.

PCA, kovaryans matrisi üzerinden hesaplanan özdeğer ve özvektörler kullanılarak boyut indirgeme işlemi yapar. Bu özvektörler, veri setinin ana bileşenlerini oluşturur ve en büyük özdeğere sahip bileşenler, verideki en fazla varyansı ifade eder. PCA uygulaması sonucu 30 orijinal özellikten iki bileşene indirgenmiştir. Bu bileşenler, verinin gözlemlenebilir şekilde işlenmesine olanak tanımış ve veri kaybı en aza indirilerek çoğu varyansın korunması sağlanmıştır. Bu sayede, veri seti hem daha yönetilebilir hale gelmiş hem de sınıflandırma algoritmaları için daha anlamlı bir girdi oluşturulmuştur.

PCA sonucunda elde edilen bileşenler, "PCA - 2 Bileşen" adıyla scatter plot olarak görselde sunulmuştur. Bu görselde verinin iki boyutta nasıl dağıldığı ve sınıflar arası ayrımın ne kadar belirgin olduğu gözlemlenebilir. PCA, özellikle büyük boyutlu veri setlerinde veriyi daha kompakt bir şekilde temsil ederek sınıflandırma modellerinin eğitim sürelerini kısaltmakta ve modellerin performansını iyileştirmektedir. Bununla birlikte, PCA'nın lineer bir yöntem olması, lineer olmayan yapıları yeterince iyi temsil edemeyeceği anlamına gelir; bu durumlarda diğer tekniklerle birlikte kullanılması gerekebilir.

**4.2 Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis (LDA), sınıflar arasındaki farklılığı maksimize ederken sınıf içi varyansı minimize etmeyi amaçlayan bir boyut indirgeme yöntemidir. LDA, özellikle sınıflandırma problemlerinde sınıflar arasındaki ayrımı netleştirmek ve modelin performansını arttırmak için kullanılır. LDA, verinin sınıflara göre daha iyi ayrılmasını sağlayarak, sınıflandırıcıların doğruluğunu artırır ve sınıflar arasındaki sınırların daha belirgin hale gelmesine katkıda bulunur. LDA uygulaması sonucu iki sınıf bulunduğundan bir bileşene indirgeme yapılmıştır.

Matematiksel olarak, LDA sınıf içi ve sınıflar arası varyansı hesaplayarak sınıfları en iyi ayıran projeksiyonu bulur. Bu projeksiyon, sınıf merkezlerinin birbirine olan uzaklığını arttırırken sınıf içindeki özelliklerin dağılımını azaltmaya çalışır. "LDA - 1 Bileşen" başlıklı scatter plot, LDA ile elde edilen tek bileşen üzerinde sınıflar arası farklılığın nasıl görüldüğünü göstermektedir. LDA, özellikle sınıflar arası belirgin ayrımları olan veri setlerinde etkili olmuş ve model performansında anlamlı bir artış sağlamıştır. LDA'nın en büyük avantajı, sınıflar arası varyansı maksimize ederek veri setindeki ayrımı artırmasıdır. Bununla birlikte, LDA'nın lineer ayrım varsayımına dayanması, sınıfların karmaşık ve lineer olmayan sınırlara sahip olduğu durumlarda performansını sınırlayabilir.

**4.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)**

t-Distributed Stochastic Neighbor Embedding (t-SNE), çok boyutlu verilerin düşük boyutlu bir alanda görsel olarak daha anlamlı hale getirilmesini sağlayan bir boyut indirgeme yöntemidir. t-SNE, özellikle verilerin karmaşık yapısını koruyarak benzerliklerin düşük boyutlu bir alanda da görülebilmesini amaçlar. Bu yöntem, çok boyutlu verilerin iki ya da üç boyutta daha iyi anlaşılması için tercih edilir. t-SNE, verilerin yapısal ilişkilerini görselleştirmede oldukça başarılıdır ve bu sayede veri bilimciler verilerdeki gizli kalıpları daha iyi gözlemleyebilirler.

Matematiksel olarak, t-SNE her bir veri noktası arasındaki benzerlikleri ölçmek için bir olasılık dağılımı kullanır ve bu benzerlikleri düşük boyutlu bir alanda korumayı amaçlar. t-SNE, verilerin düşük boyutlu bir görselleştirme için uygun hale getirilmesi sırasında özellikle sınıf içi benzerliklerin korunmasına dikkat eder. "t-SNE - 2 Bileşen" başlıklı görselleştirme, verilerin iki boyutta nasıl gruplandığını ve sınıflar arası benzerliklerin ne kadar korunduğunu gösterir. t-SNE, özellikle verilerin karmaşık yapılarını anlamak ve görsel olarak farklılıkları ortaya çıkarmak için kullanılmış ve sınıflar arası farklılıkların daha belirgin hale getirilmesini sağlamıştır. t-SNE'nin en önemli avantajlarından biri, yüksek boyutlu ve karmaşık verilerde bile verilerin iç yapısını görselleştirmeye yardımcı olmasıdır. Ancak, t-SNE'nin bir dezavantajı, parametre seçimine duyarlı olması ve farklı parametrelerle farklı sonuçlar üretebilmesidir. Ayrıca, t-SNE'nin hesaplama maliyeti diğer yöntemlere göre daha yüksektir ve bu da büyük veri setlerinde zaman alıcı olabilir.

**4.4 Boyut İndirgeme Tekniklerinin Karşılaştırılması**

Her üç boyut indirgeme tekniği de farklı amaçlarla kullanılmış ve her birinin özellikle veri setinin yapısına uygun avantajları olmuştur. PCA, verideki en büyük varyansı koruyarak boyutları azaltmış ve bu sayede sınıflandırma modellerinin eğitim sürelerini kısaltmıştır. PCA, özellikle verilerin lineer olarak ayrılabildiği durumlarda başarılı bir şekilde çalışır ve veri kaybını minimumda tutarak veriyi daha anlamlı hale getirir. Ancak, lineer olmayan yapıları yeterince iyi temsil edememesi, bu tekniğin kullanımını belirli durumlarda sınırlayabilir.

LDA, sınıflar arası farklılığı maksimize ederek model performansında önemli bir iyileşme sağlamış ve sınıf ayrımını daha net hale getirmiştir. LDA, verilerin belirli sınıflara ayrıldığı durumlarda, özellikle iki sınıf arasındaki ayrımı artırmada çok etkilidir. Bu yöntem, sınıf içi varyansları minimize ederek, sınıflar arasındaki farklılığı en üst düzeye çıkarmaktadır. Ancak, LDA'nın varsayımlarının gerçek veri setlerinde her zaman geçerli olmaması, bu yöntemin esnekliğini sınırlayabilir. Örneğin, sınıfların normal dağılıma sahip olmasını ve kovaryansların eşit olmasını varsayar; bu varsayımlar sağlanmadığında, LDA'nın performansı olumsuz etkilenebilir.

t-SNE ise verilerin görsel olarak daha iyi anlaşılmasını sağlamış ve karmaşık yapılardaki benzerlikleri koruyarak sınıfların gruplandırılmasına olanak tanımıştır. t-SNE, veri bilimcilerin verilerdeki kalıpları, kümeleri ve sınıfları görselleştirerek daha derin bir anlayış kazanmalarını sağlar. Bu, özellikle verilerin görsel analizinde ve veri keşif sürecinde çok değerli olabilir. Ancak t-SNE'nin yüksek hesaplama maliyeti ve parametre seçiminin sonuçlar üzerindeki etkisi, bu yöntemin kullanımı sırasında dikkat edilmesi gereken noktalardır. Parametrelerin uygun şekilde ayarlanmaması, t-SNE ile elde edilen görselleştirmelerin yanlış yorumlanmasına neden olabilir.

Boyut indirgeme tekniklerinin seçimi, veri setinin yapısı ve problem türüne bağlı olarak değişebilir. Bu çalışmada, PCA ve LDA teknikleri, sınıflandırma işlemlerinde daha iyi performans sağlamışken, t-SNE özellikle verilerin görsel analizinde daha etkili olmuştur. PCA, verideki en büyük varyansı korurken, LDA sınıflar arasındaki ayrımı güçlendirmiştir. t-SNE ise karmaşık veri yapılarının daha iyi görselleştirilmesine olanak tanımıştır. Elde edilen bu sonuçlar, boyut indirgeme yöntemlerinin doğru şekilde seçilmesinin model performansı ve analiz süreci için kritik bir öneme sahip olduğunu göstermiştir. Doğru teknik seçimi, sadece model performansını arttırmakla kalmaz, aynı zamanda verilerin daha iyi anlaşılmasına ve görselleştirilmesine de katkıda bulunur. Bu nedenle, boyut indirgeme teknikleri, makine öğrenmesi sürecinin önemli bir parçası olarak dikkatle değerlendirilmelidir.

**5. Sınıflandırma Modelleri**

Çalışmanın bu kısmında, veri setine beş farklı makine öğrenmesi algoritması uygulanmıştır. Bu algoritmaların her biri farklı avantaj ve dezavantajlara sahiptir ve verinin yapısına uygun olarak kullanılmıştır:

**1. Logistic Regression:** Lineer ayrım kabiliyeti sağlar. Basit ve yorumlanabilir bir modeldir, fakat lineer olmayan verilerde performansı düşebilir.

**2. Random Forest:** Karar ağacı topluluğudur, çoklu karar ağacı kullanarak tahmin yapar. Aşırı uyuma karşı dayanıklıdır, fakat bazen daha fazla hesaplama gücü gerektirir.

**3. Support Vector Machines (SVM):** Veriyi sınıflar arasındaki uzaklığı maksimize eden bir hiper düzleme oturtur. Lineer olmayan sınırlarda kernel kullanarak çalışabilir. Ancak, büyük veri setlerinde eğitim süresi uzun olabilir.

**4. K-Nearest Neighbors (KNN):** En yakın komşuluk bilgilerini kullanarak tahmin yapar. Verinin dağılımına bağlı olarak etkili olabilir, fakat büyük veri setlerinde hesaplama maliyeti artar.

**5. Naive Bayes:** Özelliklerin bağımsız olduğu varsayılarak Bayes Teoremi'ni kullanır. Hızlı ve etkili bir modeldir, fakat bu bağımsızlık varsayımı her zaman gerçekleşmez.

Sınıflandırma ve performans değerlendirmeleri Stratified K-Fold Çapraz Doğrulama yöntemi ile yapılmıştır. Bu yöntem, verinin dengeli bir şekilde eğitim ve test setlerine bölünmesini sağlayarak model performansının daha doğru bir şekilde değerlendirilmesine olanak tanır.

**6. Kullanılan Kütüphaneler**

Bu çalışmada Python programlama dili kullanılarak çeşitli veri işleme ve makine öğrenmesi kütüphanelerinden faydalanılmıştır:

**NumPy:** Yüksek performanslı sayısal işlemler ve matris manipülasyonu sağlamak için kullanılmıştır.

**Pandas:** Veri yükleme, işleme ve analiz için kullanılmıştır. Veri setindeki özelliklerin düzenlenmesi ve temel veri manipülasyonları için önemli rol oynamıştır.

**Scikit-Learn:** Veri ön işleme, boyut indirgeme teknikleri ve sınıflandırma modelleri bu kütüphane kullanılarak uygulanmıştır. Ayrıca model değerlendirme ve çapraz doğrulama işlemleri de bu kütüphane ile gerçekleştirilmiştir.

**Matplotlib ve Seaborn:** Veri görselleştirme için kullanılmıştır. Boyut indirgeme tekniklerinin sonuçlarını ve sınıflar arası ayrımı daha iyi anlayabilmek için grafikler oluşturulmuştur.

**Joblib:** Model ve ölçekleyici nesnelerin saklanması için kullanılmıştır. Bu sayede, daha sonra kullanılmak üzere eğitimli modellerin kaydedilmesi sağlanmıştır.

**7. Sonuçlar ve Analiz**

Doğruluk Oranları Tablosu:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Yöntem** | **Logistic Regression** | **Random Forest** | **SVM** | **KNN** | **Naive Bayes** |
| Boyut İndirgeme Yok | %97.37 | %95.43 | %97.54 | %96.66 | %92.97 |
| PCA | %95.08 | %93.33 | %95.08 | %93.15 | %91.39 |
| LDA | %97.54 | %96.31 | %97.54 | %97.89 | %97.37 |
| t-SNE | %95.43 | %96.13 | %95.26 | %95.26 | %95.26 |

- En iyi performans LDA ile elde edilmiştir. LDA, sınıflar arasındaki farkı optimize ederek en yüksek başarıyı sağlamıştır. Bu durum, LDA'nın sınıflar arasındaki ayrımı daha net bir şekilde ortaya koyabilmesi ile açıklanabilir.

- t-SNE ise daha iyi bir görsel sunum sağlarken sınıflandırma performansına doğrudan katkısı daha sınırlı olmuştur. Bu teknik, verilerin görsel anlamda daha anlaşılabilir hale gelmesi için kullanılmış, fakat sınıflandırma başarısında LDA kadar etkili olmamıştır.

**8. Görselleştirme**

Verilerin görselleştirilmesi için çeşitli scatter plot grafiklerinden faydalanılmıştır. PCA, LDA ve t-SNE yöntemleriyle oluşturulan görünümler arasında çoklu sınıf ayrımına vurgu yapılmış ve sınıfların nasıl farklıştığı gösterilmiştir. Bu görsel analizler, verideki yapının daha iyi anlaşılmasına ve hangi boyut indirgeme yönteminin sınıflar arası farklılıkları daha iyi ortaya çıkardığını anlamamıza yardımcı olmuştur.

- PCA Scatter Plot: PCA ile indirgenmiş verilerin dağılımı, veri setindeki tüm özelliklerin maksimum varyansı koruyarak iki boyutta nasıl görüntülendiğini göstermektedir.

- LDA Scatter Plot: LDA'nın sınıflar arası farklılığı gösterdiği görselleştirme, iki sınıf arasındaki ayrımı daha net bir şekilde ortaya koymaktadır.

- t-SNE Scatter Plot: Verilerin karmaşık yapısının iki boyutlu bir yansıması, sınıflar arasındaki benzerliklerin ve farklılıkların daha iyi görülmesini sağlamaktadır. t-SNE, verilerin gruplar halindeki yapısını çok daha belirgin hale getirmiştir.

**9. Sonuç ve Öneriler**

**En Başarılı Yöntem:** LDA, sınıf ayrımı konusunda en başarılı sonucu vermiştir. Özellikle iki sınıfın net bir şekilde ayrıştığı durumlarda etkili olmuştur. LDA'nın sınıflar arası farklılıkları maksimize etmesi, modelin genel performansını arttırmıştır.

**PCA ve t-SNE:** PCA daha fazla varyansı yakalama kabiliyeti sunarken, t-SNE görsel çeşitlilik sağlamaktadır. Ancak, her iki yöntem de LDA kadar ayrım sağlayamamıştır. PCA, verilerin boyutunu azaltarak eğitim süresini hızlandırırken, t-SNE daha karmaşık yapıların görülmesini sağlamıştır.

**Öneriler:**

- Daha fazla özellik ve sınıf dengesizliği ile çalışmak için ileri seviye boyut indirgeme teknikleri test edilebilir. Özellikle, verideki gürültünün azaltılması ve sınıfların daha net ayrılması için farklı algoritmalar kullanılabilir.

- Daha geniş veri setlerinde çeşitli hiperparametre optimizasyon teknikleri kullanılabilir. Bu sayede, model performansı daha da arttırılabilir ve çeşitli veri ön işleme teknikleri kullanılarak özelliklerin etkisi analiz edilebilir.

- Verinin dengesiz sınıflara sahip olduğu durumlarda, SMOTE gibi tekniklerle veri dengesi sağlanarak modelin başarısı arttırılabilir. Bu sayede, az sayıda bulunan sınıfların model tarafından daha iyi öğrenilmesi sağlanabilir.

Bu rapor, boyut indirgeme tekniklerinin makine öğrenmesi algoritmalarına etkisini detaylı bir şekilde sunmakta ve farklı yöntemlerin sınıflandırma performansına olan etkilerini incelemektedir. Elde edilen bulgular, makine öğrenmesi sürecinde boyut indirgeme yöntemlerinin doğru bir şekilde seçilmesinin model başarısı için kritik olduğunu göstermektedir.

**10. Kaynaklar**

* UCI Machine Learning Repository, Breast Cancer Wisconsin (Diagnostic) Veri Seti, <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>
* Scikit-Learn Documentation, <https://scikit-learn.org/stable/documentation.html>
* NumPy Documentation, <https://numpy.org/doc/stable/>
* Pandas Documentation, <https://pandas.pydata.org/pandas-docs/stable/>
* Matplotlib Documentation, <https://matplotlib.org/stable/users/index.html>
* Github Repository Link, <https://github.com/firatkaanbitmez/pattern-recognition>