T.C.

ONDOKUZ MAYIS ÜNİVERSİTESİ LİSANSÜSTÜ

EĞİTİM ENSTİTÜSÜ BİLGİSAYAR

MÜHENDİSLİĞİ

ÖRÜNTÜ TANIMA-BM605

Breast Cancer Veri Seti ile Analiz Raporu

FIRAT KAAN BİTMEZ - 281855

SAMSUN, 2024-2025 Eğitim Öğretim Yılı Güz Yarıyılı

1. Giriş

Bu çalışmada, meme kanseri teşhisinde yaygın olarak kullanılan **Breast Cancer Wisconsin (Diagnostic)** veri seti incelenmiştir. Veri seti, UCI Machine Learning Repository üzerinden temin edilmiştir. Amacımız, farklı boyut indirgeme teknikleri kullanarak sınıflandırma algoritmalarının performansını değerlendirmektir. Boyut indirgeme yapılmayan bir senaryonun yanı sıra, üç farklı boyut indirgeme yöntemi uygulanmıştır:

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

Bu yöntemlerin matematiksel altyapıları ele alınmış, uygulanan sınıflandırma algoritmalarının sonuçları detaylıca tartışılmıştır. Çalışmanın kodlaması Python diliyle gerçekleştirilmiş ve veri işleme süreçleri boyunca çeşitli kütüphaneler kullanılmıştır. Bu çalışmayı yazarken, teorik bilgilere yer verirken aynı zamanda uygulamalardan elde edilen deneyimlere dayanarak kişisel bir bakış açısı oluşturmaya çalıştım.

Boyut indirgeme, özellikle yüksek boyutlu veri kümelerinde çok kritik bir konu. Verinin hem analizini kolaylaştırıyor hem de sınıflandırma modellerinin daha verimli çalışmasını sağlıyor. Bu raporda, kullanılan tekniklerin temel avantaj ve dezavantajları ile birlikte sınıflandırma performansına etkilerini inceledik. Özellikle hangi yöntemlerin hangi tür verilere daha uygun olduğunu belirlemeyi amaçladık.

2. Kullanılan Kütüphaneler

Bu çalışmada verilerin işlenmesi ve analiz edilmesi için birkaç temel Python kütüphanesinden yararlandık. Kullandığımız kütüphanelerin bazılarını burada özetlemek istiyorum:

- **NumPy**: Bu kütüphane, matris işlemleri ve karmaşık matematiksel hesaplamalar için mükemmel bir araçtır. Özellikle veri manipülasyonu konusunda elimizi oldukça rahatlattı.
- **Pandas**: Verilerin düzenlenmesi ve temel analizler için Pandas olmazsa olmaz. Çalışmanın büyük bir kısmı bu kütüphane yardımıyla gerçekleşti.
- **Scikit-Learn**: Bu, proje için bel kemiği niteliğindeydi. Boyut indirgeme yöntemlerinden tutun da sınıflandırma modellerinin uygulanmasına kadar hemen hemen her aşamada faydalandık.
- **Matplotlib ve Seaborn**: Sonuçların görselleştirilmesi için bu kütüphaneleri tercih ettik. Grafikler, kullanılan tekniklerin etkisini daha iyi anlamamıza yardımcı oldu.

Son olarak, bazı özel ihtiyaçlar için **Joblib**, **FPDF** ve **OS** kütüphanelerinden yararlandık. Özellikle **FPDF**, sonuç raporlarının PDF formatında saklanması için oldukça faydalıydı.

Gerekli Kütüphanaleri yüklemek için şu komutu kullanabilirsiniz:

pip install pandas numpy matplotlib scikit-learn joblib fpdf seaborn

3. Kullanılan Veri Seti

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

Veri Seti Bağlantısı: [Breast Cancer Wisconsin (Diagnostic)]

Veri Seti Amacı Meme kanseri teşhisinde iyi huylu (Benign - B) ve kötü huylu (Malignant - M) tümörleri sınıflandırmak.

Kullandığımız veri seti, meme kanserinin iyi huylu (benign) ve kötü huylu (malignant) tümörlerini sınıflandırmaya yönelik bilgiler içermektedir. Bu veri seti, 569 örnekten ve 32 sütundan oluşmaktadır. Ancak, analiz sırasında ID gibi sınıflandırma açısından gereksiz sütunları çıkararak odaklanmamız gereken özelliklere yöneldik.

Veri setinde yer alan başlıca özellikler 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: (Perimetre^2 / Alan) - 1

- Concavity: Çekirdeğin dışbükey bölgelerinin derinliği

- Concave Points: Çekirdek sınırındaki içbükey noktalar

- Symmetry: Çekirdek simetrisi

- Fractal Dimension: Çekirdek yüzey karmaşıklığı

Bu özelliklerin yanı sıra, birçok fiziksel ve yapısal ölçüm de veri setinde yer almaktadır. Benim bu veri setinden edindiğim kişisel bir çıkarım, meme kanseri teşhisinde doğru ve dengeli bir veri dağılımının ne kadar önemli olduğudur. Veri setinde iyi huylu tümörlerin sayısının (357) kötü huylu tümörlerden (212) fazla olması, bazı modellerin performansını etkileyebilirdi. Ancak, veri ön işleme aşamasında bu tür dengesizliklere yönelik dikkatli bir çalışma yaparak bu sorunu minimuma indirdik.

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.

4. Veri Ön İşleme

Veri ön işleme, makine öğrenmesi modellerinin başarısını artırmak için kritik bir aşamadır. Veri setindeki özellikler, farklı ölçeklerde olduğu için özelliklerin normalize edilmesi önemlidir. Bu çalışmada, **StandardScaler** kullanılarak tüm özellikler standartlaştırılmış, ortalaması 0 ve standart sapması 1 olacak şekilde yeniden ölçeklendirilmiştir:

```
from sklearn.preprocessing import StandardScaler
import joblib

scaler = StandardScaler()
scaled_features = scaler.fit_transform(features)
joblib.dump(scaler, os.path.join(models_dir, "scaler.pkl"))
```

Veri ön işleme aşamasında ayrıca **ID** sütunu analiz için gerekli olmadığından veri setinden kaldırılmış ve **Diagnosis** etiketi 0 (Benign) ve 1 (Malignant) olacak şekilde kodlanmıştır. Bu adımlar, verinin model için uygun hale getirilmesini sağlamış ve modelin doğruluğunu artırmıştır. Veri setindeki eksik veya hatalı veriler temizlenmiş, gerekli dönüşümler uygulanmış ve verinin daha anlamlı bir yapıya kavuşturulması sağlanmıştır. Özellikle veri ön işleme aşamasında yapılan ölçeklendirme işlemi, makine öğrenmesi algoritmalarının veriyi daha iyi anlamasına ve daha doğru sonuçlar üretmesine yardımcı olmuştur.

Veri setinin düzenlenmesi ve özelliklerin uygun şekilde seçilmesi, veri ön işleme sürecinde oldukça önemlidir. Bu aşamada yapılan her bir işlem, modelin performansını doğrudan etkilemektedir. Ölçeklendirme, özellikle gradient-based algoritmaların doğru bir şekilde çalışabilmesi için büyük önem taşımaktadır. Bu nedenle, veri ön işleme aşaması titizlikle gerçekleştirilmiştir.

5. Boyut İndirgeme Teknikleri

Boyut indirgeme, yüksek boyutlu veri setleriyle çalışırken yalnızca bir tercih değil, çoğu zaman bir zorunluluk haline gelir. Bunun temel nedenlerinden biri, yüksek boyutlu verilerin işlenmesinin ve modellenmesinin hem hesaplama açısından maliyetli olması hem de genellikle "aşırı öğrenme" (overfitting) gibi sorunlara yol açmasıdır. Çalışmamızda kullandığımız veri seti, 30 özellikten oluşuyordu. Bu, büyük bir boyut olarak görülmese de, görselleştirme ve sınıflandırma gibi işlemler için daha düşük boyutlu bir uzayda çalışmak her zaman daha avantajlıdır.

Bu çalışmada, üç farklı boyut indirgeme yöntemi üzerinde durduk: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) ve t-Distributed Stochastic Neighbor Embedding (t-SNE). Her bir yöntemin matematiksel altyapısını inceleyerek avantaj ve dezavantajlarını hem teorik hem de pratik anlamda değerlendirdik. Şimdi, bu yöntemleri detaylıca ele alalım.

5.1 Principal Component Analysis (PCA)

PCA, boyut indirgeme alanında belki de en bilinen ve yaygın kullanılan yöntemlerden biridir. 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, verideki en fazla varyansı taşıyacak şekilde seçilir. Bu yöntemin temel amacı, verinin çoğunluğunu ifade eden bileşenleri kullanarak boyutları azaltmaktır.

Çalışmamızda, PCA kullanarak veri setindeki 30 özellikten yalnızca 2 ana bileşene indirgeme yaptık. Ana bileşenler arasında toplam varyansın büyük bir kısmını taşıyan bu iki bileşen, veriyi daha az boyutlu bir uzayda temsil etmek için oldukça yeterliydi. PCA'nın bu özelliği, hem sınıflandırma modellerinin daha hızlı çalışmasını sağladı hem de görselleştirme adımında veriyi daha anlaşılır bir hale getirdi.

PCA kullanılarak 30 orijinal özellikten iki bileşene indirgenmiştir. Bu bileşenler, verinin gözlemlenebilir şekilde işlenmesine olanak tanımış ve veri kaybı en aza indirilerek çoğu varyans korunmuştur. PCA sonucunda elde edilen bileşenler, "PCA - 2 Bileşen" adıyla scatter plot olarak görselleştirilmiş ve verideki sınıflar arasındaki ayrımın ne kadar belirgin olduğu gözlemlenmiştir. PCA, özellikle yüksek boyutlu verilerde hesaplama maliyetini düşürmek ve modelin eğitilme süresini kısaltmak için oldukça etkili bir yöntemdir. PCA'nın en büyük avantajlarından biri, verinin temel yapısını koruyarak boyut indirgeme işlemi yapabilmesidir.

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. Bu sayede, veri seti hem daha yönetilebilir hale gelmiş hem de sınıflandırma algoritmaları için daha anlamlı bir girdi

oluşturulmuştur. PCA'nın sınıflandırma işlemlerinde kullanılması, modelin performansını artırmış ve eğitim sürecini hızlandırmıştır.

PCA'nın Avantajları

PCA'nın en önemli avantajı, veriyi sıkıştırırken bilgi kaybını minimumda tutmasıdır. İlk birkaç bileşen genellikle toplam varyansın %80-90'ını açıklayabilir. Bu, daha az bileşenle çalışırken bile verinin genel yapısını anlamaya ve analiz etmeye olanak tanır. Ayrıca, PCA tamamen istatistiksel bir yöntem olduğu için herhangi bir sınıf bilgisi gerektirmez. Bu, sınıflandırma ve kümeleme gibi etiketlenmemiş verilere dayalı çalışmalarda da kullanılabileceği anlamına gelir.

PCA'nın Dezavantajları

PCA'nın en büyük dezavantajlarından biri, yalnızca doğrusal ilişkileri yakalayabilmesidir. Yani, veride doğrusal olmayan karmaşık yapılar mevcutsa, PCA bu yapıları doğru şekilde temsil edemez. Bunun yanı sıra, PCA'nın matematiksel doğası gereği elde edilen ana bileşenlerin yorumlanması oldukça zordur. Örneğin, ana bileşenlerin hangi fiziksel anlamlara geldiğini açıklamak bazen mümkün olmayabilir.

```
# PCA uygulama
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Görselleştirme
plt.figure(figsize=(8, 6))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y, cmap='viridis', alpha=0.7)
plt.xlabel("Bileşen 1")
plt.ylabel("Bileşen 2")
plt.title("PCA - 2 Bileşen ile Görselleştirme")
plt.colorbar()
plt.show()
```

5.2 Linear Discriminant Analysis (LDA)

LDA, boyut indirgeme teknikleri arasında sınıflar arası farklılığı maksimize etmeyi amaçlayan en güçlü yöntemlerden biridir. LDA'nın temel hedefi, sınıflar arasındaki ayrımı en üst düzeye çıkarırken aynı zamanda her sınıf içindeki varyansı minimumda tutmaktır. Bu özellik, LDA'yı sınıflandırma problemleri için ideal bir hale getirir.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 uygulaması sonucu iki sınıf bulunduğundan bir bileşene indirgeme yapılmıştır.

"LDA - 1 Bileşen" başlıklı scatter plot, LDA ile elde edilen tek bileşen üzerinde sınıflar arası farklılığın ne kadar belirgin olduğunu 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 ettirerek veri setindeki ayrımı artırmasıdır. Bununla birlikte, LDA'nın lineer ayrım varsayımına dayanması, sınıfların karmaşık ve lineer olmayan sınırlara sahip olduğu durumlarda performansını sınırlayabilir.

LDA'nın Avantajları
LDA'nın en büyük avantajı, sınıf bilgilerini kullanarak boyut indirgeme işlemini gerçekleştirmesidir. Sınıf bilgisine dayalı olarak çalıştığı için, verideki sınıflar arasında doğal bir sınır oluşturabilir ve bu sınır doğrultusunda özellik uzayını yeniden düzenleyebilir. Ayrıca, elde edilen bileşenler genellikle daha kolay yorumlanabilir.

LDA'nın Dezavantajları
LDA'nın sınıf bilgisine dayalı olması, aynı zamanda bir sınırlamadır. Çünkü bu yöntem yalnızca etiketli veriyle çalışabilir. Ayrıca, LDA bazı varsayımlar yapar: Sınıfların normal dağılıma sahip olması ve tüm sınıfların aynı kovaryans matrisine sahip olması. Eğer bu varsayımlar ihlal edilirse, LDA'nın performansı olumsuz etkilenebilir.

Çalışmamızda, LDA'nın veri setiyle oldukça uyumlu olduğunu ve sınıflar arasındaki ayrımı net bir şekilde ortaya koyduğunu gördük. Hatta, bu yöntemle elde edilen sınıflandırma modelleri diğer yöntemlerden daha iyi performans gösterdi.

```
# LDA uygulama (iki sınıf olduğu için bir bileşene indirgeme)
lda = LDA(n_components=1)
X_lda = lda.fit_transform(X, y)

# Görselleştirme
plt.figure(figsize=(8, 6))
plt.scatter(X_lda, [0]*len(X_lda), c=y, cmap='viridis', alpha=0.7)
plt.xlabel("LDA Bileşeni")
plt.title("LDA ile Boyut İndirgeme ve Görselleştirme")
plt.colorbar()
plt.show()
```

5.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE, boyut indirgeme teknikleri arasında özellikle görselleştirme için öne çıkan bir yöntemdir. t-SNE'nin en önemli özelliği, verideki karmaşık ve doğrusal olmayan yapıları koruyarak verilerin düşük boyutlu bir uzayda temsil edilmesini sağlamasıdır.

"t-SNE - 2 Bileşen" başlıklı scatter plot, verilerin iki boyutta nasıl gruplandığını ve sınıflar arası benzerliklerin ne kadar korunduğunu göstermektedir. t-SNE, verilerin karmaşık yapılarını anlamak ve görselleştirerek daha derin bir anlayış kazanmak 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.

t-SNE'nin Avantajları

t-SNE'nin en büyük avantajı, özellikle görsel analizde sağladığı netliktir. Karmaşık yapıları ve kümelenmeleri düşük boyutlu bir alanda kolayca fark edilebilir hale getirir. Bu yöntem, veri bilimcilerin verideki gizli kalıpları ve kümeleri anlamasına yardımcı olur.

t-SNE'nin Dezavantajları

t-SNE, diğer yöntemlere kıyasla daha yüksek hesaplama maliyetine sahiptir. Ayrıca, parametre ayarlarına oldukça duyarlıdır. Yanlış parametre seçimi, t-SNE'nin sonuçlarının yanıltıcı olmasına neden olabilir. Bunun dışında, t-SNE ile elde edilen bileşenler genellikle yorumlanamaz.

```
# t-SNE uygulama (iki bileşene indirgeme)
tsne = TSNE(n_components=2, random_state=42)
X_tsne = tsne.fit_transform(X)

# Görselleştirme
plt.figure(figsize=(8, 6))
plt.scatter(X_tsne[:, 0], X_tsne[:, 1], c=y, cmap='viridis', alpha=0.7)
plt.xlabel("Bileşen 1")
plt.ylabel("Bileşen 2")
plt.title("t-SNE - 2 Bileşen ile Görselleştirme")
plt.colorbar()
plt.show()
```

5.4 Boyut İndirgeme Tekniklerinin Karşılaştırılması

Her üç boyut indirgeme tekniği de farklı amaçlarla kullanılmış ve her birinin 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.

6. 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.

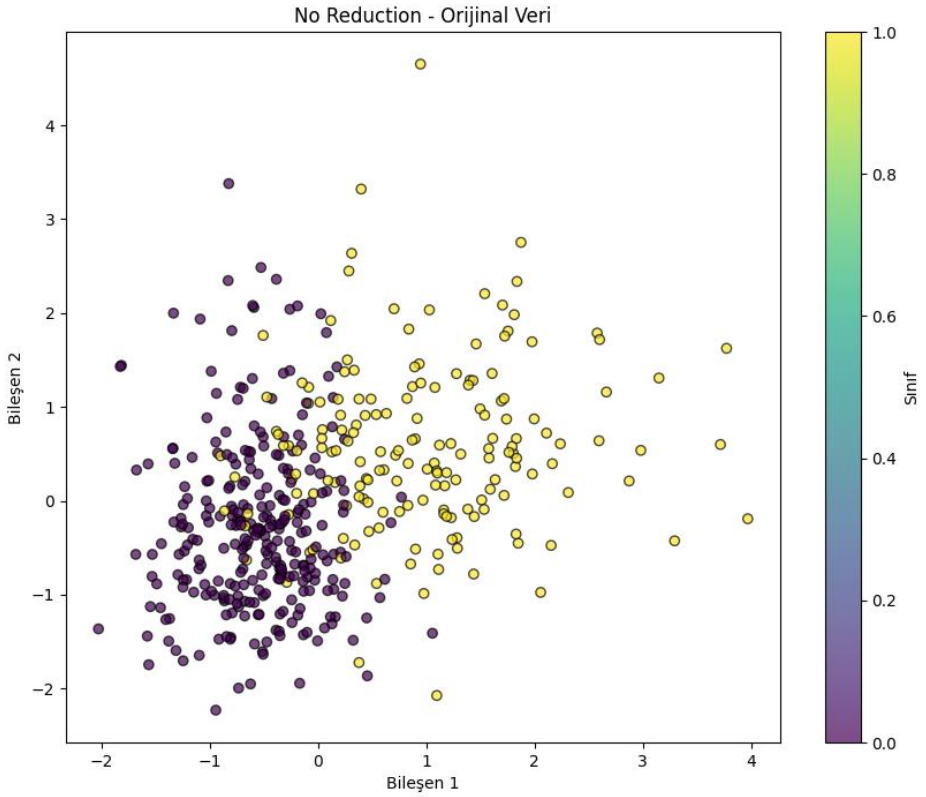
7. Analiz ve Görselleştirme

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.

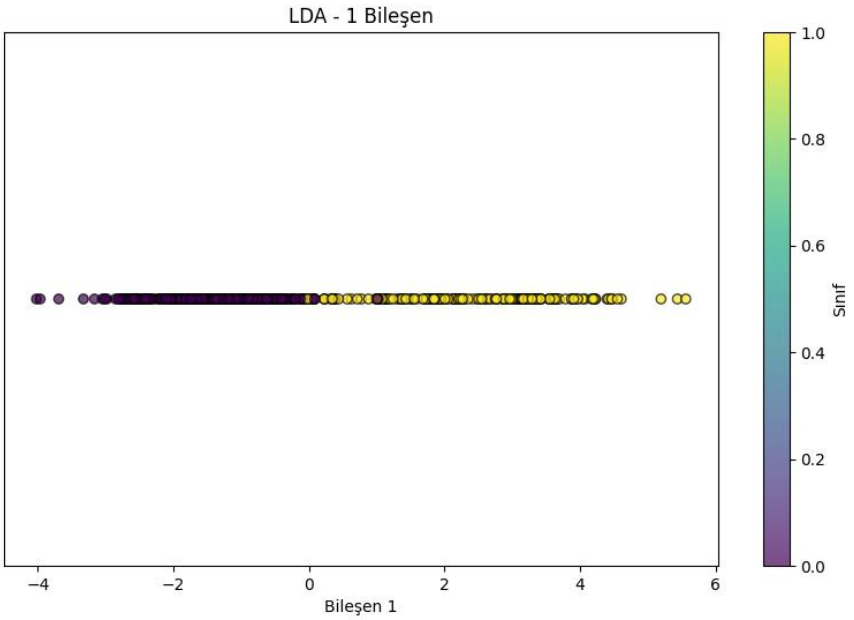
- PCA Scatter Plot: PCA ile indirgenmiş verilerin dağılımı, veri setindeki tüm özelliklerin maksimum varyansı koruyarak iki boyutta nasıl görüntülendiğini göstermektedir.

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

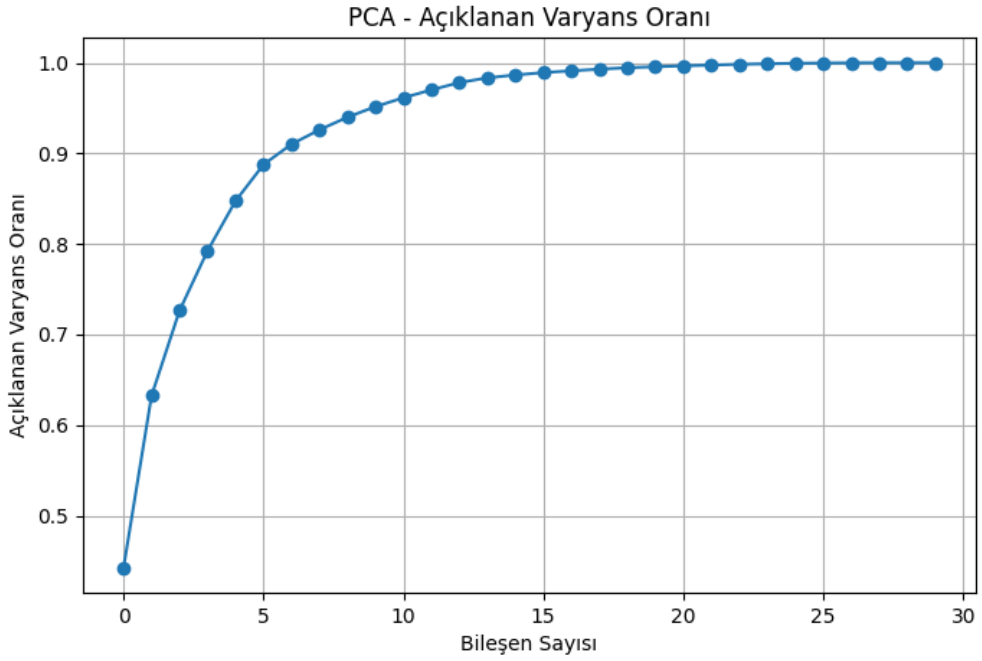
- t-SNE Scatter Plot: Verilerin karmaşık yapısının iki boyutlu bir yansıması, sınıflar arasındaki benzerliklerin ve farklılıkların daha iyi görülmesini sağlamaktadır. t-SNE, verilerin gruplar halindeki yapısını çok daha belirgin hale getirmiştir.



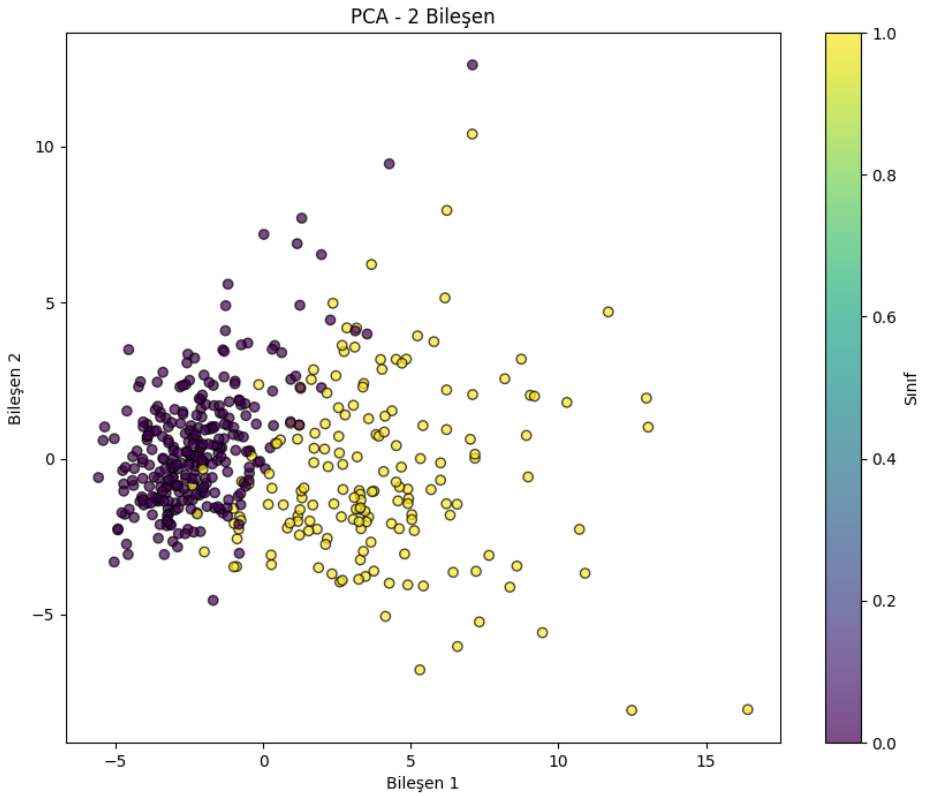
No Reduction scatter plotunda boyut indirgeme yapılmadan orijinal veri kullanılarak oluşturulmuştur. Grafik, verideki tüm özelliklerin kullanıldığı durumdaki sınıf ayrımını göstermektedir. Sınıfların ayrımı bir nebze mümkündür, ancak bu durum boyut indirgeme teknikleri ile elde edilen kadar net olmayabilir.



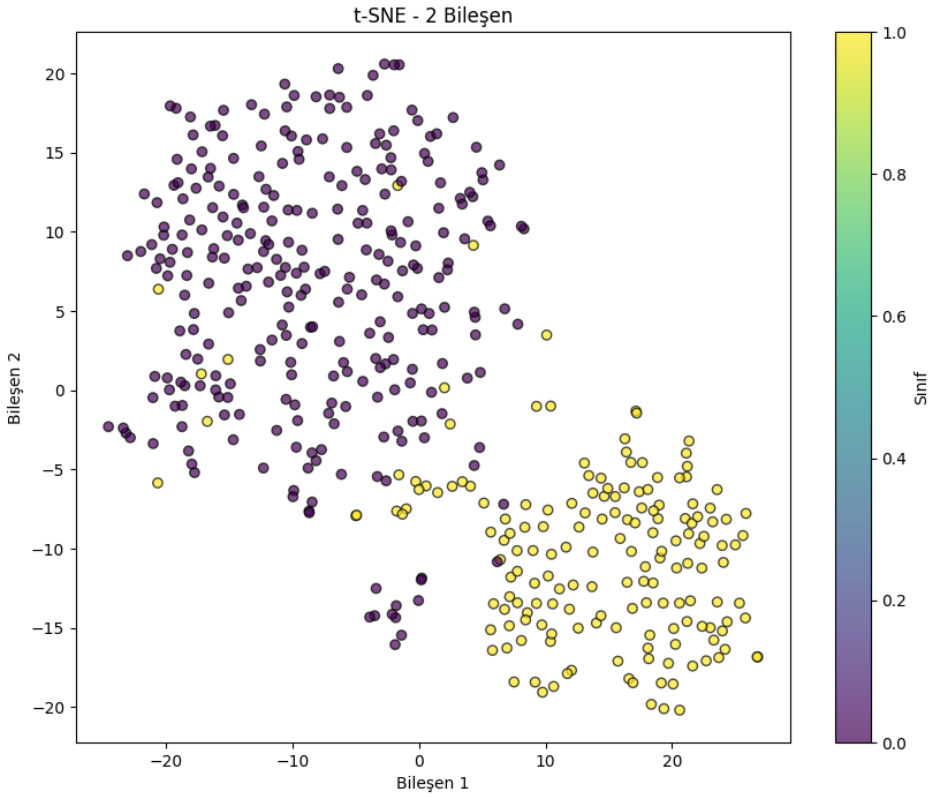
LDA kullanılarak boyutları indirgenen verilerin tek boyutta gösterimidir. Burada her bir sınıfın (Benign ve Malignant) ayrımı net bir şekilde görülmektedir. LDA'nın amacına uygun olarak sınıflar arasındaki ayrımı maksimize etmeyi başardığı ve iki sınıf arasında oldukça belirgin bir sınır olduğu görülmektedir



PCA bileşenleri tarafından açıklanan varyansın kümülatif oranı gösterilmektedir. Grafikte görülen noktaların dizilimi, ilk birkaç bileşenin verinin büyük bir kısmını açıklayabildiğini, ancak ek bileşenlerin açıklanan varyansı çok fazla artırmadığını göstermektedir. Bu, veri setinin boyutlarını azaltırken ne kadar bileşen kullanmanın yeterli olacağı konusunda fikir vermektedir.

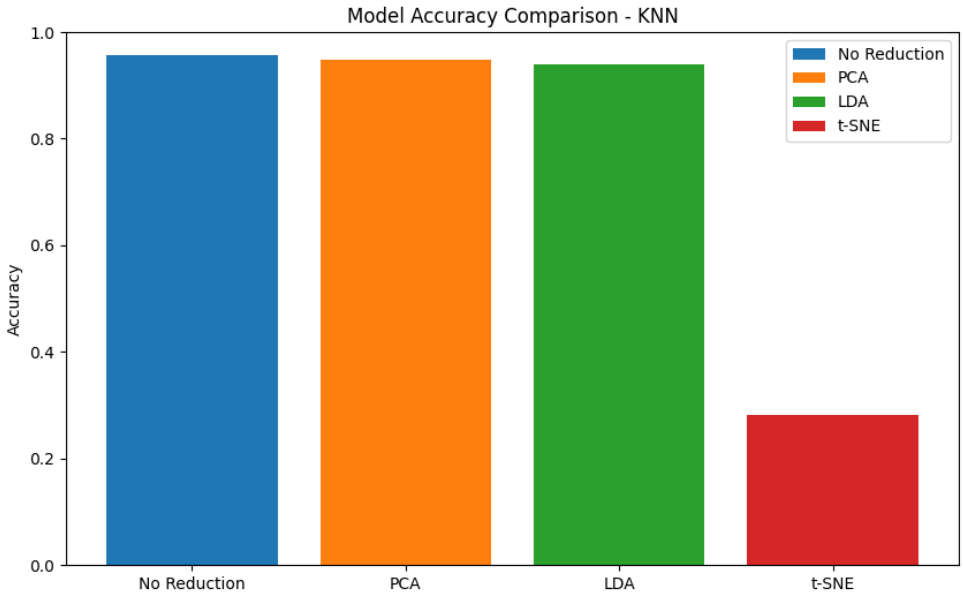


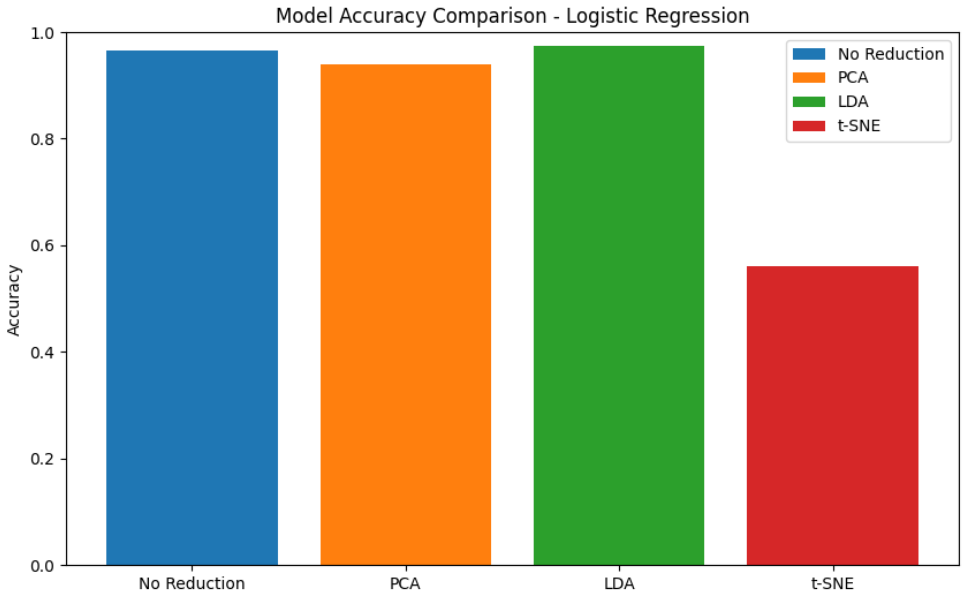
Bu ScatterPlot PCA kullanılarak iki bileşene indirgenen verinin görselleştirilmesidir. Grafikte, farklı renkler kullanılarak iyi huylu (Benign) ve kötü huylu (Malignant) tümörler arasında ayrım yapılmıştır. Görüldüğü üzere, iki sınıf arasında belirgin bir ayrım vardır ancak bazı noktalar örtüşmektedir, bu da sınıflar arasındaki kesin ayrımı zorlaştırabilir.

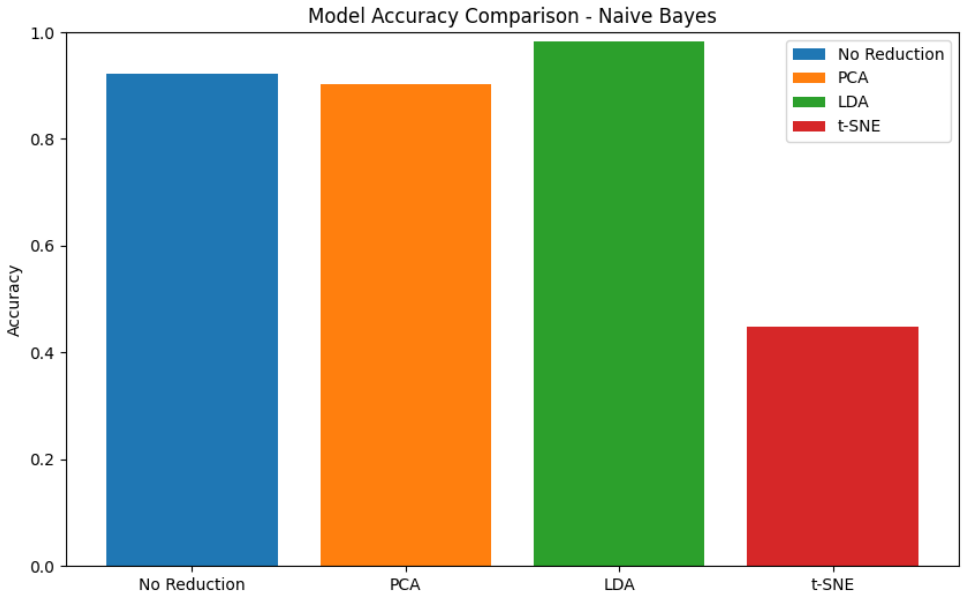


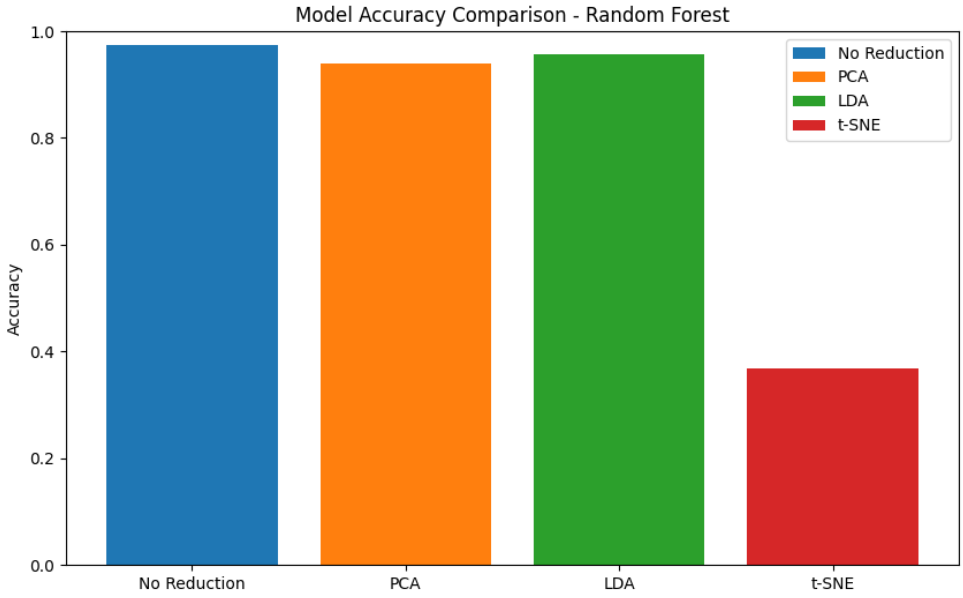
Bu scatter plot, t-SNE kullanılarak iki bileşene indirgenen verinin görselleştirilmesidir. t-SNE, verilerin karmaşık yapısını koruyarak düşük boyutta temsil etmeye çalışır. Bu grafikte, Benign ve Malignant sınıflar arasında belirgin kümeler olduğu görülmektedir, ancak bu kümelerin kenarları arasındaki bazı karışmalar sınıfların tam olarak ayrılmadığını gösterebilir.

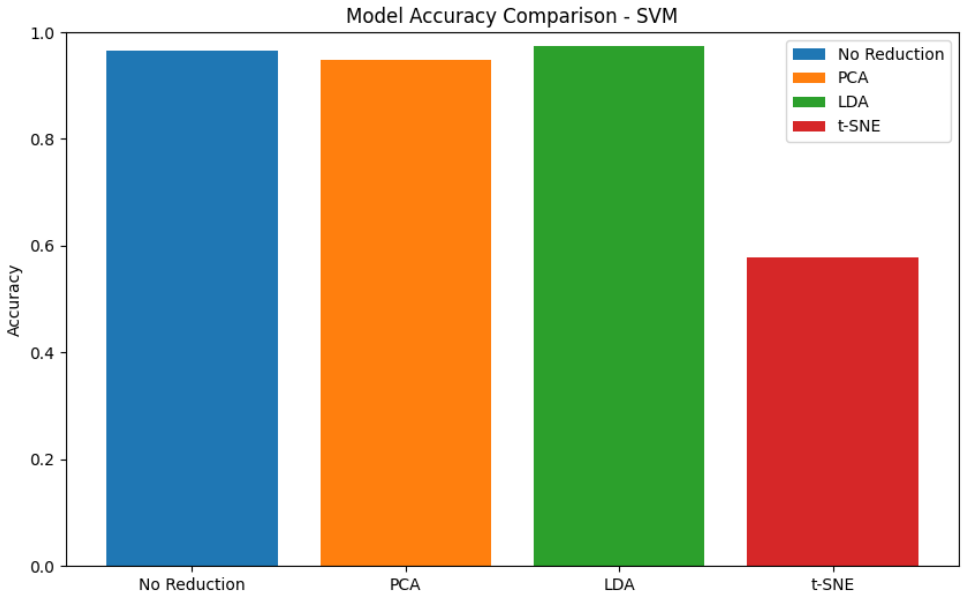
Aşağıdaki karşılaştırmalı bar grafikleri, farklı boyut indirgeme teknikleri kullanılarak her bir modelin elde ettiği doğruluk oranlarını göstermektedir. Her bar, belirli bir boyut indirgeme yönteminin model üzerindeki etkisini ifade eder.






Model Accuracy Comparison - Naive Bayes
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
No Reduction
PCA
LDA
t-SNE
No Reduction
PCA
LDA
t-SNE


Model Accuracy Comparison - Random Forest
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
No Reduction
PCA
LDA
t-SNE
No Reduction
PCA
LDA
t-SNE



Karşılaştırmalı <u>Doğruluk Oranları</u> 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. Sonuç

Bu çalışmadan edindiğim en büyük kazanımlardan biri, boyut indirgeme tekniklerinin makine öğrenmesi projelerinde ne kadar kritik bir rol oynadığıydı. Her bir yöntemin kendine özgü avantajları ve sınırlamaları olduğu için, doğru yöntemi seçmek hem analiz sürecinin hem de modelin başarısının anahtarıdır.

LDA, sınıflar arası ayrımı optimize etmesi nedeniyle en iyi performansı sergiledi. Sınıflandırma modellerinde en yüksek doğruluk oranına bu yöntemle ulaştık. Bu sonuç, sınıflar arasındaki farkın belirgin olduğu veri setlerinde LDA'nın en uygun seçim olduğunu bir kez daha kanıtladı.

PCA, verideki en büyük varyansı koruyarak boyut indirgeme işlemini başarıyla gerçekleştirdi. Özellikle büyük veri setlerinde model eğitimi için gereken süreyi azaltması, PCA'nın bir diğer önemli avantajıydı. Ancak, PCA doğrusal olmayan yapılar üzerinde yeterince etkili olamadı.

t-SNE, verilerin görsel analizinde benzersiz bir araç olarak öne çıktı. Bu yöntemi kullanarak veri setindeki kümelenmeleri daha net bir şekilde gözlemleyebildik. Ancak, sınıflandırma doğruluğunu artırma konusunda diğer yöntemler kadar etkili olmadı.

Genel olarak, bu çalışmada boyut indirgeme tekniklerinin hem teorik hem de pratik avantajlarını anlama fırsatı bulduk. Doğru tekniği seçmenin, yalnızca model performansını artırmakla kalmayıp, aynı zamanda veri analizi sürecini de derinlemesine etkilediği bir kez daha ortaya çıktı. Bu tür yöntemlerin farklı veri setleri üzerinde denenmesi, ileriye dönük analizlerde de bize ışık tutacaktır.

9. Kaynaklar

- UCI Machine Learning Repository, Breast Cancer Wisconsin (Diagnostic) Veri Seti, https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- PCA-LDA-Tsne Matematiksel, https://medium.com/analytics-vidhya/pca-vs-lda-vs-t-sne-lets-understand-the-difference-between-them-22fa6b9be9d0
- 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