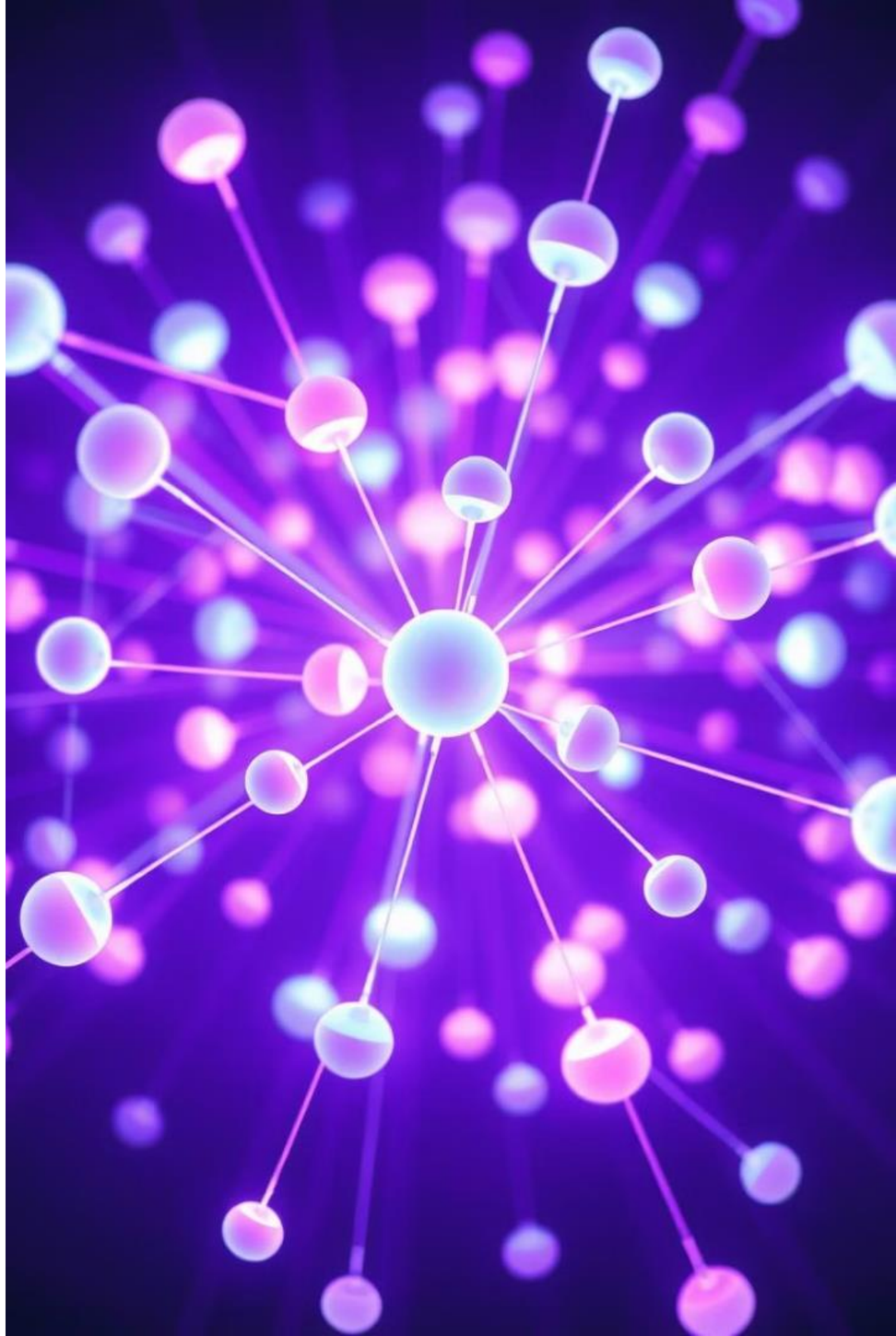
Breast Cancer Wisconsin Veri Seti ile Kanser Teşhisinde Boyut İndirme Tekniklerinin Etkisi

Breast Cancer Wisconsin veri setini kullanarak farklı boyut indirgeme tekniklerinin sınıflandırma algoritmalarının performansına etkisini incelemeyi hedefliyoruz.

Çeşitli yöntemleri uygulayarak, boyut indirgeme tekniklerinin kanser teşhisinde önemli bir rol oynadığını ortaya koymayı hedefliyoruz.



Veri Seti: Breast Cancer Wisconsin (Diagnostic)

Özellikler

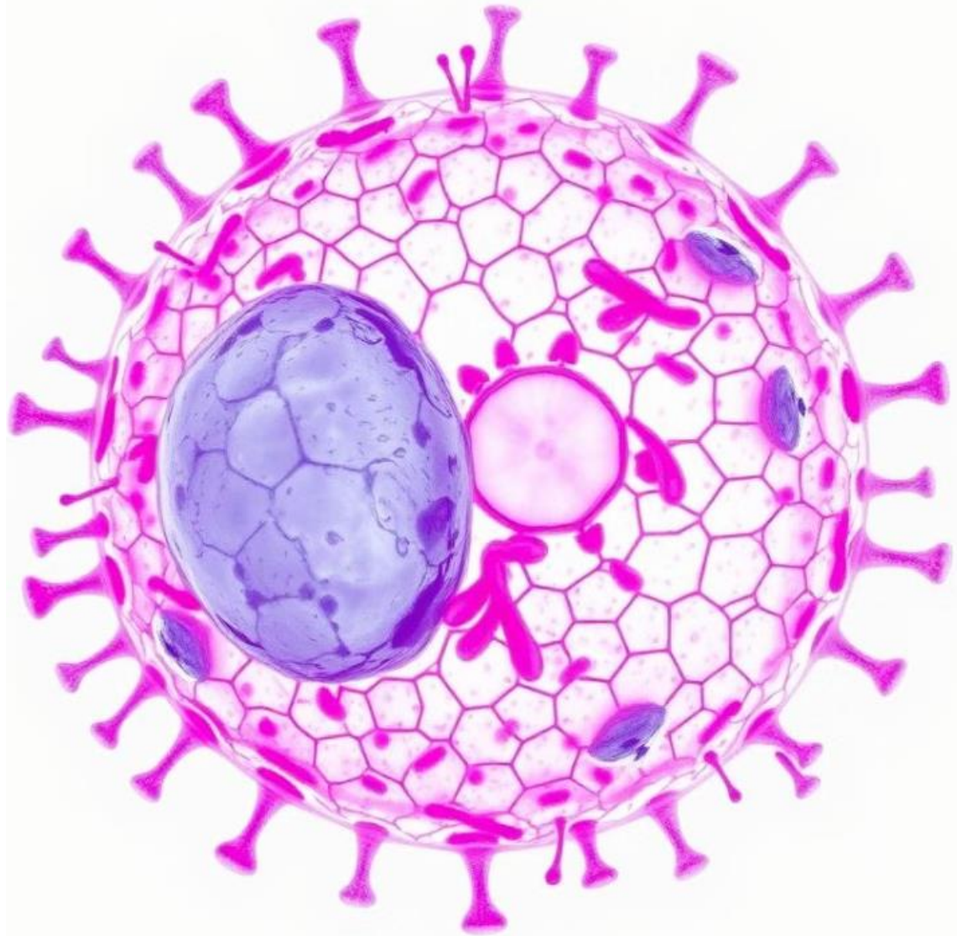
Veri seti, hücre çekirdeği çapı, doku, çevre, alan, pürüzsüzlük, sıkışıklık, dışbükeylik, dışbükey noktalar, simetri ve fraktal boyut gibi 30 özelliği içerir.

Sınıf Etiketleri

Her örnek, iyi huylu (B) veya kötü huylu (M) olmak üzere iki sınıftan birine aittir. Bu bilgiler, sınıflandırma algoritmaları için eğitim verisi olarak kullanılır.

Önem

Veri seti, meme kanseri teşhisinin zorluğunu ve farklı boyut indirgeme teknikleri kullanılarak bu teşhisin nasıl iyileştirilebileceğini göstermektedir.



Veri Ön İşleme

1 Ölçeklendirme

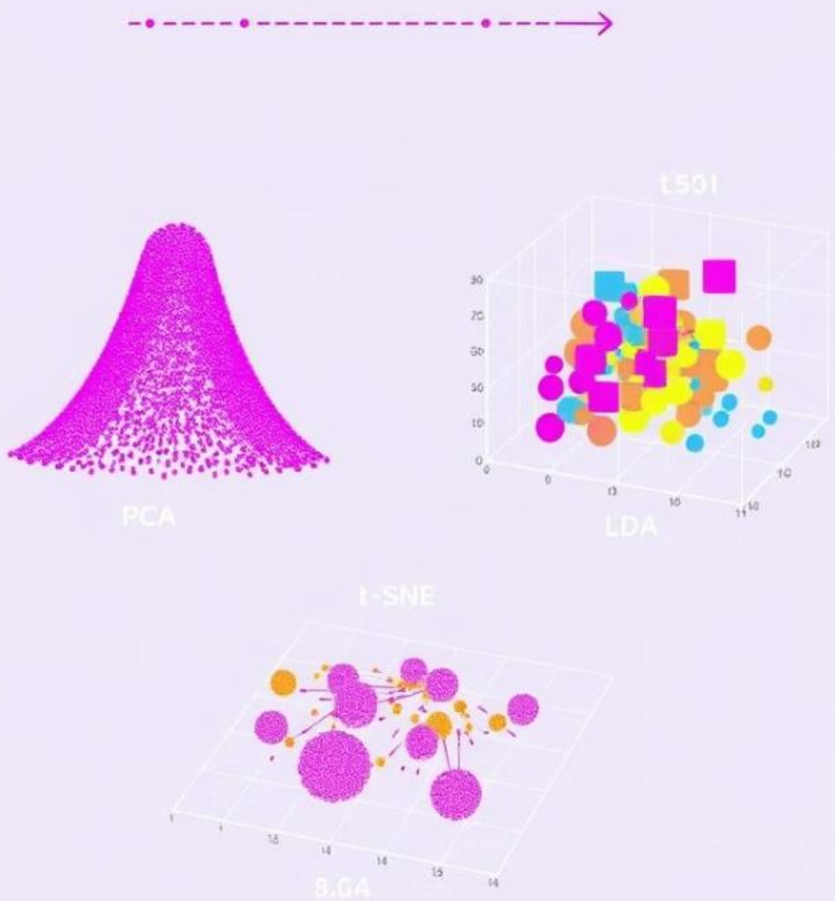
StandardScaler kullanarak tüm özellikler standartlaştırılmıştır. Bu, tüm özelliklerin aynı ölçekte olmasını sağlayarak sınıflandırma algoritmalarının performansını iyileştirir.

2 Özellik Seçimi

Analiz için gerekli olmayan ID sütunu veri setinden kaldırılmıştır. Bu, modelin sadece ilgili özelliklere odaklanmasını sağlar.

3 Etiket Kodlama

Teşhis etiketi 0 (İyi Huylu) ve 1 (Kötü Huylu) olacak şekilde kodlanmıştır. Bu, sınıflandırma algoritmaları için etiketleri sayısal hale getirir.



Boyut İndirgeme Teknikleri

PCA Principal Component Analysis

Veri setindeki varyansı en üst düzeye çıkararak boyut indirgeme işlemi yapar. Doğrusal ilişkileri etkili bir şekilde yakalar, ancak doğrusal olmayan yapılar için uygun olmayabilir.

LDA Linear Discriminant Analysis

Sınıflar arasındaki farklılığı maksimize ederek sınıflandırma için optimize edilmiş boyut indirgeme sağlar. Sınıf bilgisine dayalı olarak çalışır, ancak varsayımlarına bağlı olarak performansı değişebilir.

t-SNE t-Distributed Stochastic Neighbor Embedding

Görselleştirme için mükemmeldir, yüksek boyutlu verileri düşük boyutlu bir uzaya koruyarak karmaşık yapıları ve kümelenmeleri ortaya çıkarır. Hesaplama maliyeti ve parametrelere duyarlılığa dikkat etmek gerekir.

Principal Component Analysis (PCA)

İşlem

PCA, verinin varyansını en fazla açıklayan temel bileşenleri bulur. Bu temel bileşenler, orijinal verinin düşük boyutlu bir temsilini oluşturur.

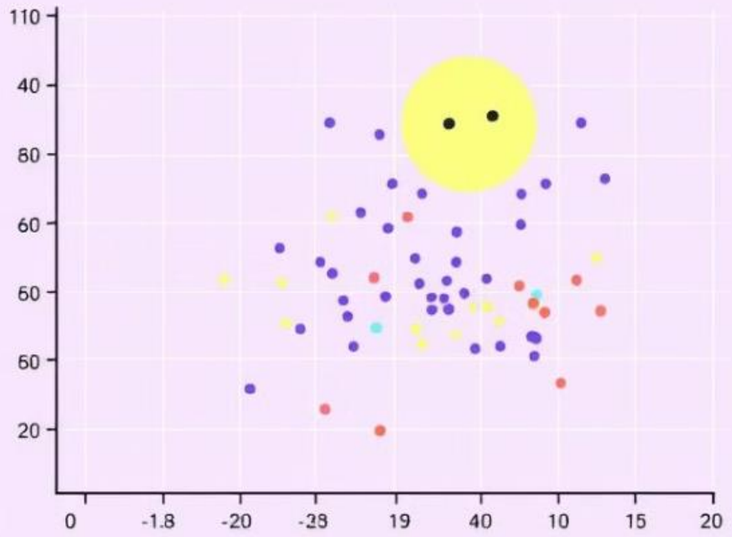
PCA, orijinal özellikleri lineer olarak dönüştürerek yeni bir özellik uzayı oluşturur. Bu dönüşüm, verinin temel yapısını koruyarak boyutları azaltmayı amaçlar.

Avantajlar

PCA, hızlı, basit ve veri kümesindeki varyansın büyük bir kısmını korumak için etkilidir. Ayrıca, veri gürültüsünü azaltmaya yardımcı olur.

Dezavantajlar

PCA, sınıf bilgisi hakkında bilgi sağlamaz. Bu nedenle, sınıflandırma problemlerinde en iyi seçim olmayabilir.



Linear Discriminant Analysis (LDA)

1

İşlem

LDA, sınıflar arasındaki ayrımı maksimize eden doğrusal dönüşümleri belirler. Bu dönüşümler, sınıflandırma algoritmalarının daha iyi performans göstermesine yardımcı olur.

2

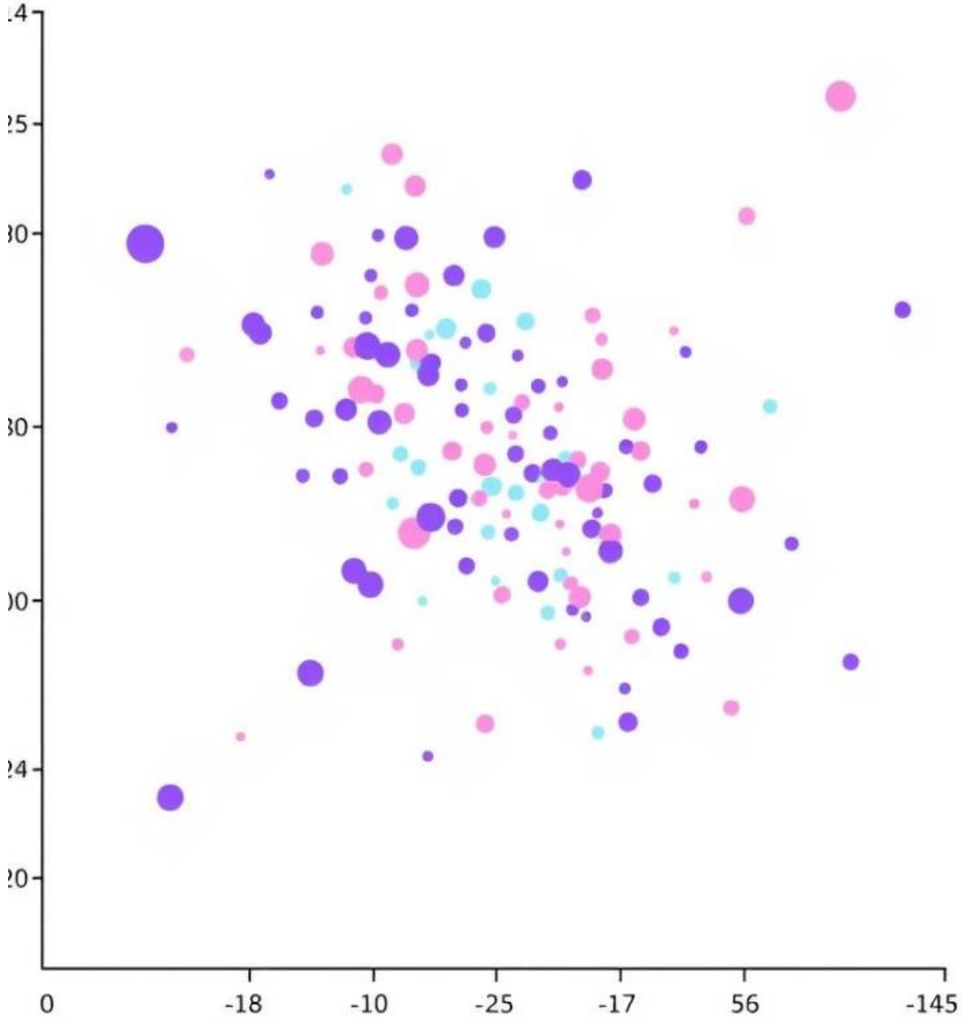
Avantajlar

LDA, sınıf bilgisini dikkate alarak, sınıflandırma problemleri için PCA'dan daha etkilidir. Ayrıca, veri kümesinde küçük bir varyans olsa bile iyi performans gösterir.

3

Dezavantajlar

LDA, doğrusal olarak ayrıştırılabilir veri için en iyi sonucu verir. Dolayısıyla, doğrusal olmayan veri kümeleri için uygun olmayabilir.



t-Distributed Stochastic Neighbor Embedding (t-SNE)

1

İşlem

t-SNE, benzer veri noktalarını birbirine yakın, farklı noktaları ise birbirinden uzak yerleştirerek yüksek boyutlu verileri görselleştirmek için kullanılır.

2

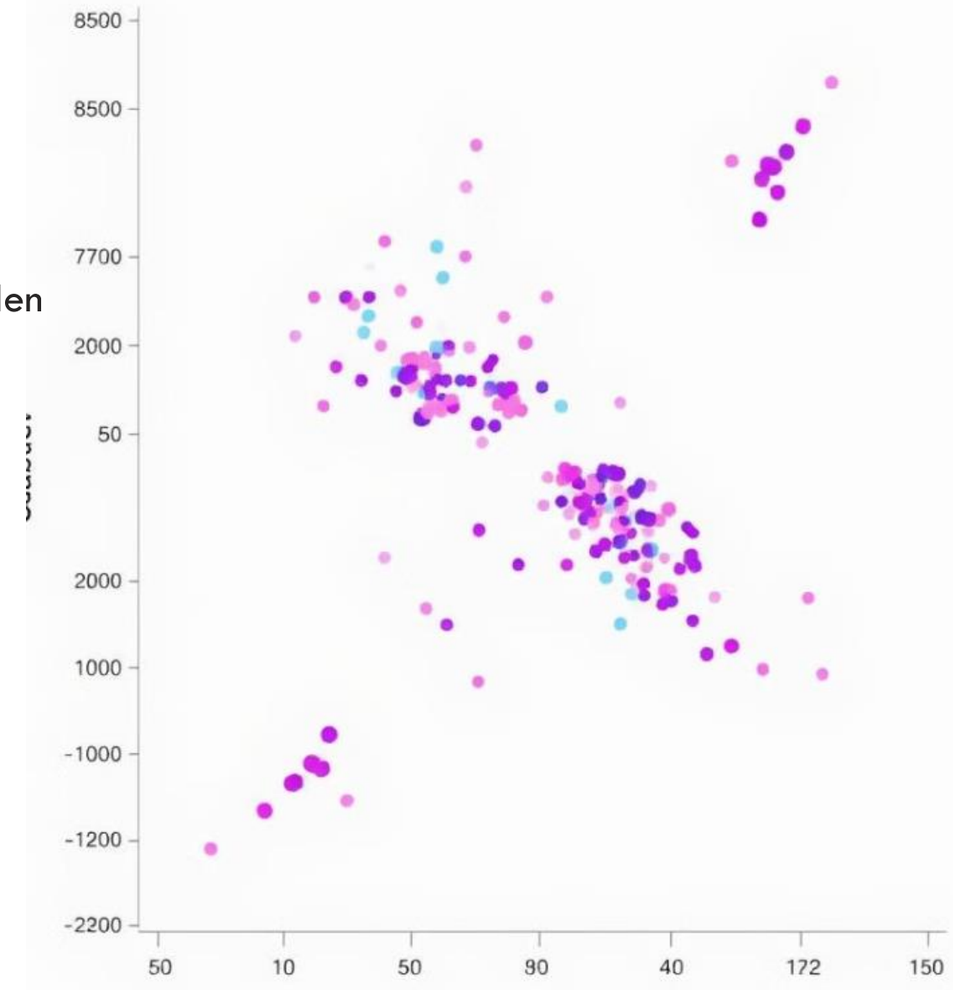
Avantajlar

t-SNE, özellikle büyük veri kümeleri için, verideki karmaşık yapılar ve kümelenmeleri ortaya çıkarma konusunda oldukça etkilidir.

3

Dezavantajlar

t-SNE, hesaplama açısından pahalı olabilir ve sonuçlar başlangıç parametrelerine duyarlı olabilir. Küme sayısı önceden belirtilmelidir.



Boyut İndirgeme: Neden Gerekli?

Hesaplama Maliyeti

Yüksek boyutlu veri setleri, analiz ve modelleme için çok fazla hesaplama gücü gerektirir. Bu, büyük veri kümeleri için eğitim sürelerinin uzamasına ve işlem maliyetlerinin artmasına yol açabilir.

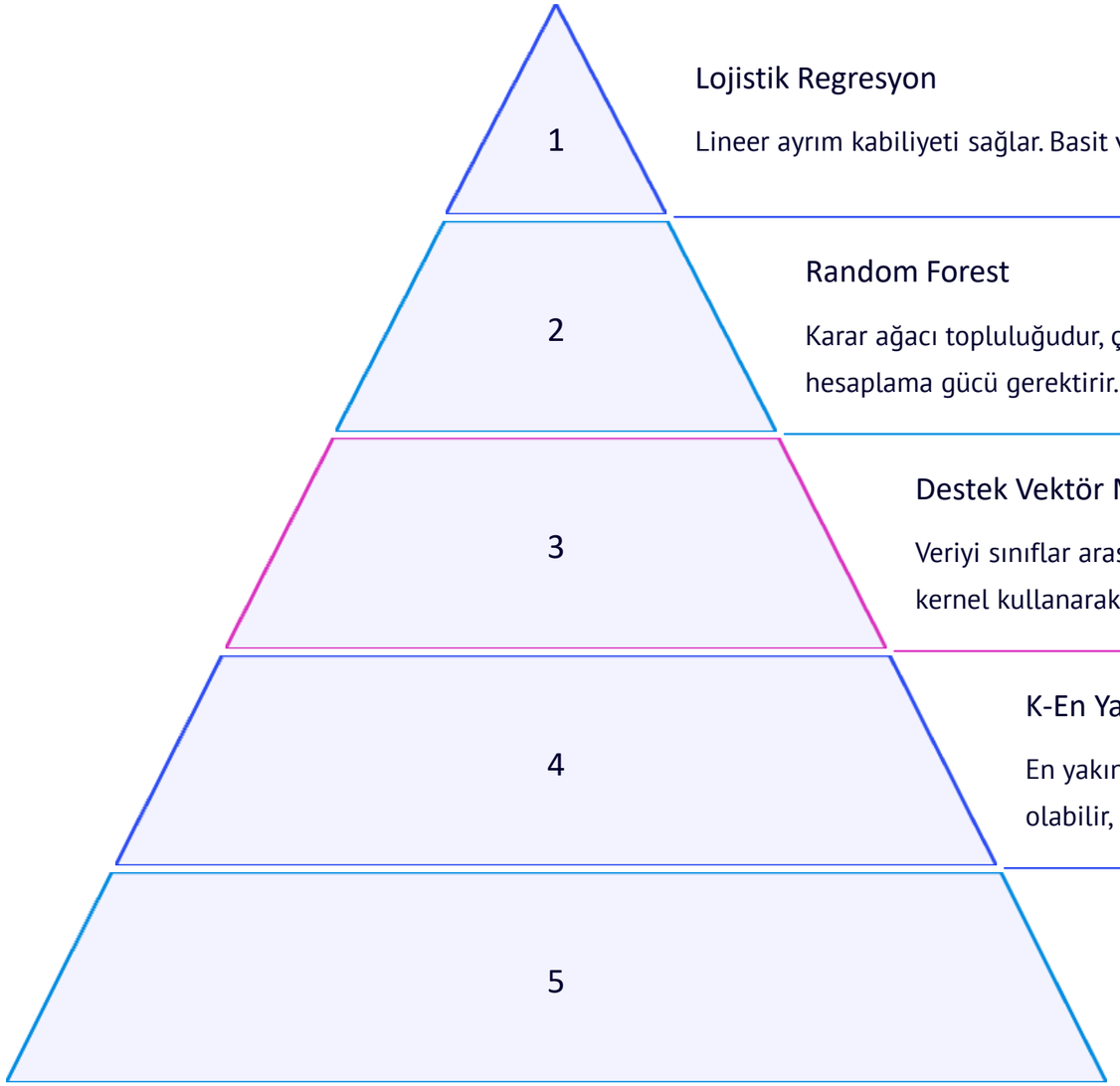
Aşırı Öğrenme

Yüksek boyutlu veri setlerinde, model verileri çok iyi öğrenebilir ve yeni verilere genelleme yeteneğini kaybedebilir. Bu da aşırı öğrenme (overfitting) olarak bilinen bir sorun yaratır.

Görselleştirme Zorluğu

Yüksek boyutlu verileri görselleştirmek oldukça zordur. İnsan beyni 3 boyuttan fazlasını görselleştiremez, bu nedenle yüksek boyutlu veri setleri görsel olarak anlaşılmaz hale gelir.

Sınıflandırma Algoritmaları



Lojistik Regresyon

Lineer ayrım kabiliyeti sağlar. Basit ve yorumlanabilirdir, ancak lineer olmayan verilerde performansı düşük olabilir.

Random Forest

Karar ağacı topluluğudur, çoklu karar ağacı kullanarak tahmin yapar. Aşırı uyuma karşı dayanıklıdır, ancak daha fazla hesaplama gücü gerektirir.

Destek Vektör Makineleri (SVM)

Veriyi sınıflar arasındaki uzaklığı maksimize eden bir hiper düzleme oturtur. Lineer olmayan sınırlarda kernel kullanarak çalışabilir. Büyük veri setlerinde eğitim süresi uzun olabilir.

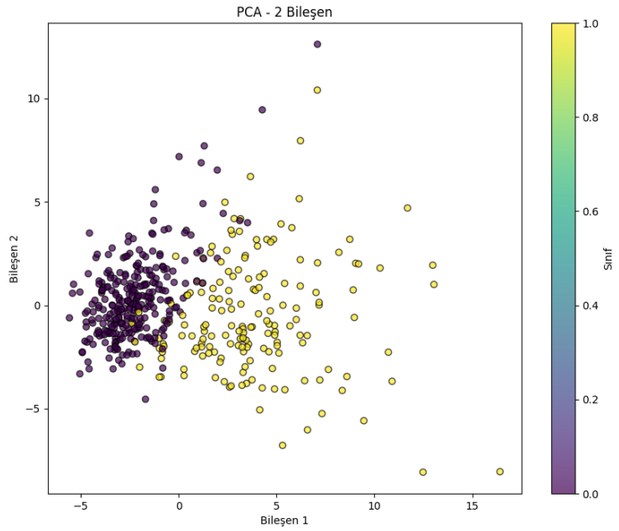
K-En Yakın Komşu (KNN)

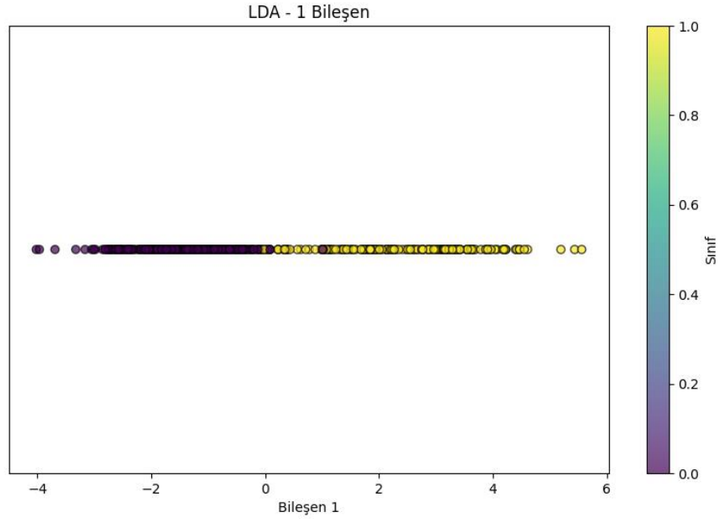
En yakın komşuluk bilgilerini kullanarak tahmin yapar. Verinin dağılımına bağlı olarak etkili olabilir, ancak büyük veri setlerinde hesaplama maliyeti artar.

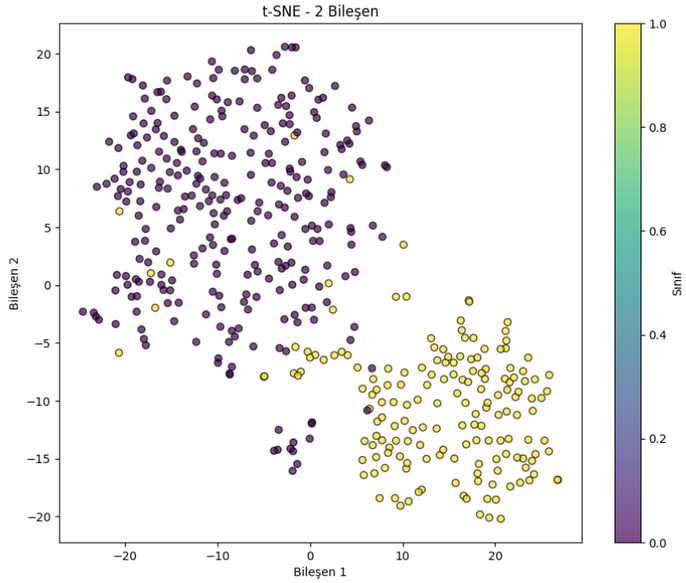
Naive Bayes

Özelliklerin bağımsız olduğu varsayılarak Bayes Teoremi'ni kullanır. Hızlı ve etkili bir modeldir, ancak bu bağımsızlık varsayımı her zaman gerçekleşmez.

Veri Analizi ve Görselleştirme







PCA Scatter Plot

PCA ile indirgenmiş verilerin dağılımı, veri setindeki tüm özelliklerin maksimum varyansı koruyarak iki boyutta nasıl görüntülendiğini göstermektedir. Bu, verinin yapısını daha iyi anlamamıza yardımcı olur.

LDA Scatter Plot

LDA'nın sınıflar arası farklılığı gösterdiği görselleştirme, iki sınıf arasındaki ayrımı daha net bir şekilde ortaya koymaktadır. Bu, sınıflandırma modellerinin performansını iyileştirmek için hangi özelliklerin en önemli olduğunu belirlememize yardımcı olur.

t-SNE Scatter Plot

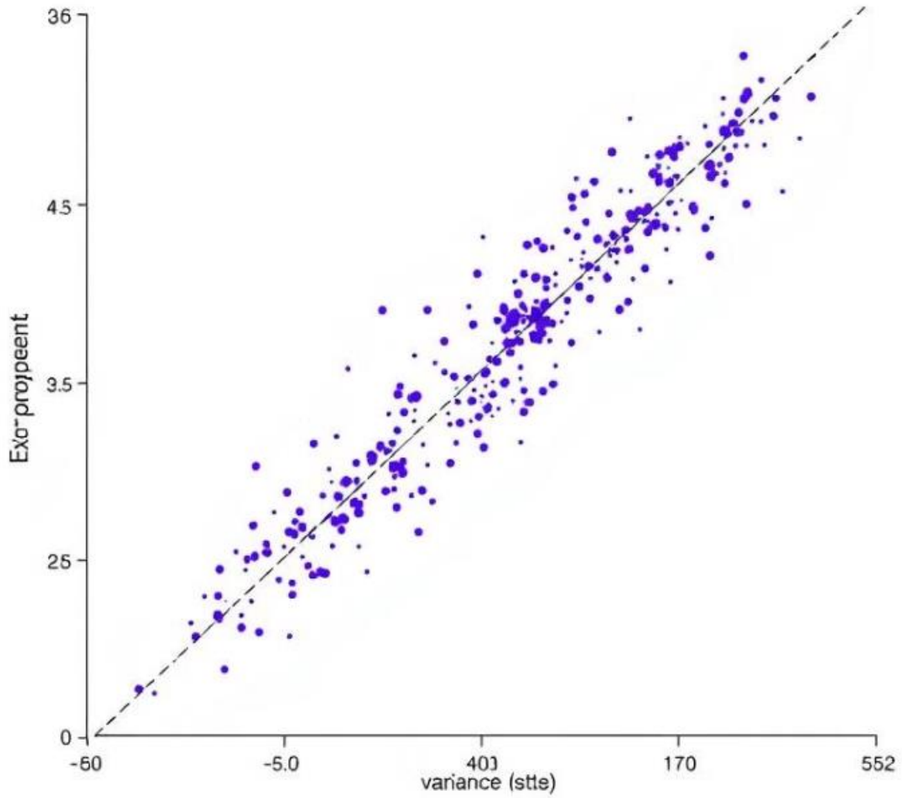
Verilerin karmaşık yapısının iki boyutlu bir yansıması, sınıflar arasındaki benzerliklerin ve farklılıkların daha iyi görülmesini sağlamaktadır. Bu, verinin kümelenme yapısını ve sınıflar arasındaki ilişkileri görselleştirmemize yardımcı olur.

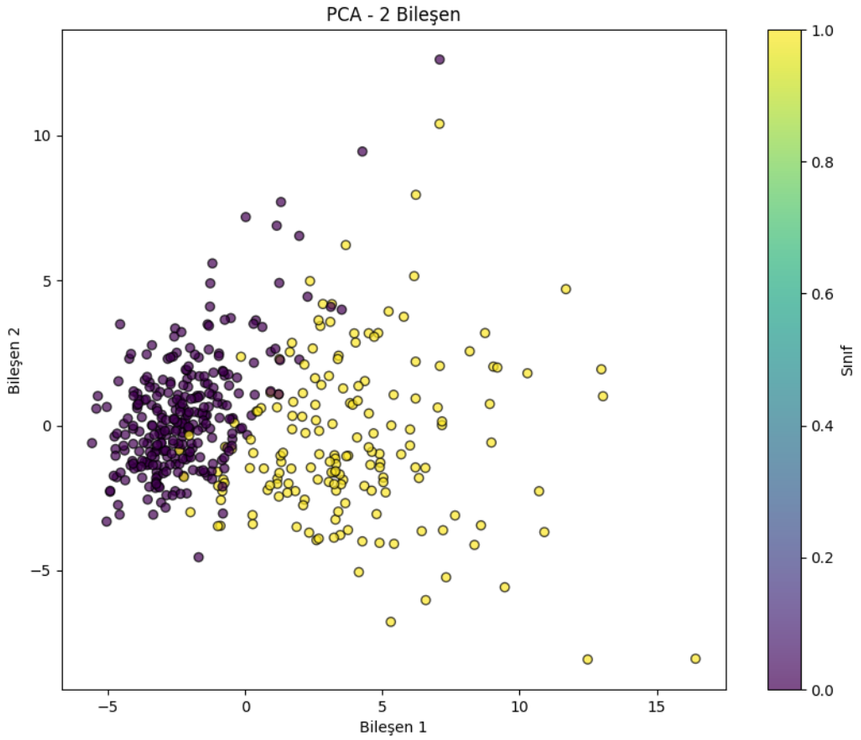
Boyut İndirgeme Tekniklerinin Karşılaştırılması

Model	Reduction Method	Precision	Recall	F1-Score
Logistic Regression	LDA	0.97	0.97	0.97
Logistic Regression	No Reduction	0.97	0.96	0.96
Logistic Regression	PCA	0.95	0.92	0.93
Logistic Regression	t-SNE	0.39	0.45	0.39
Random Forest	LDA	0.95	0.96	0.95
Random Forest	No Reduction	0.98	0.96	0.97
Random Forest	PCA	0.94	0.93	0.93
Random Forest	t-SNE	0.28	0.3	0.29
SVM	LDA	0.97	0.97	0.97
SVM	No Reduction	0.97	0.95	0.96
SVM	PCA	0.95	0.93	0.94
SVM	t-SNE	0.31	0.46	0.37
KNN	LDA	0.93	0.95	0.94
KNN	No Reduction	0.96	0.95	0.96
KNN	PCA	0.95	0.93	0.94
KNN	t-SNE	0.24	0.23	0.23
Naive Bayes	LDA	0.99	0.98	0.98
Naive Bayes	No Reduction	0.92	0.91	0.91
Naive Bayes	PCA	0.91	0.88	0.89
Naive Bayes	t-SNE	0.32	0.36	0.33

Boyut İndirgeme Tekniklerinin Performansı

Grafik, PCA bileşenleri tarafından açıklanan varyansın kümülatif oranını göstermektedir. 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.



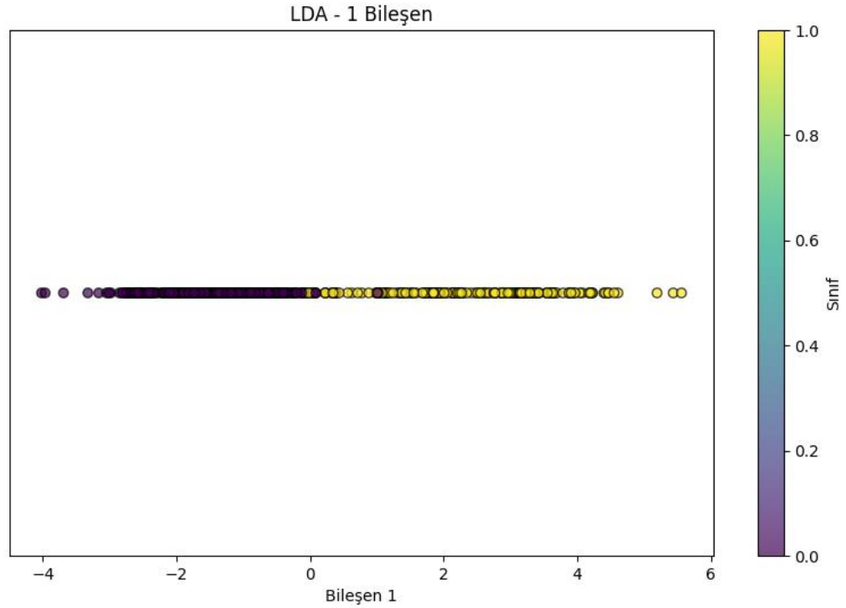


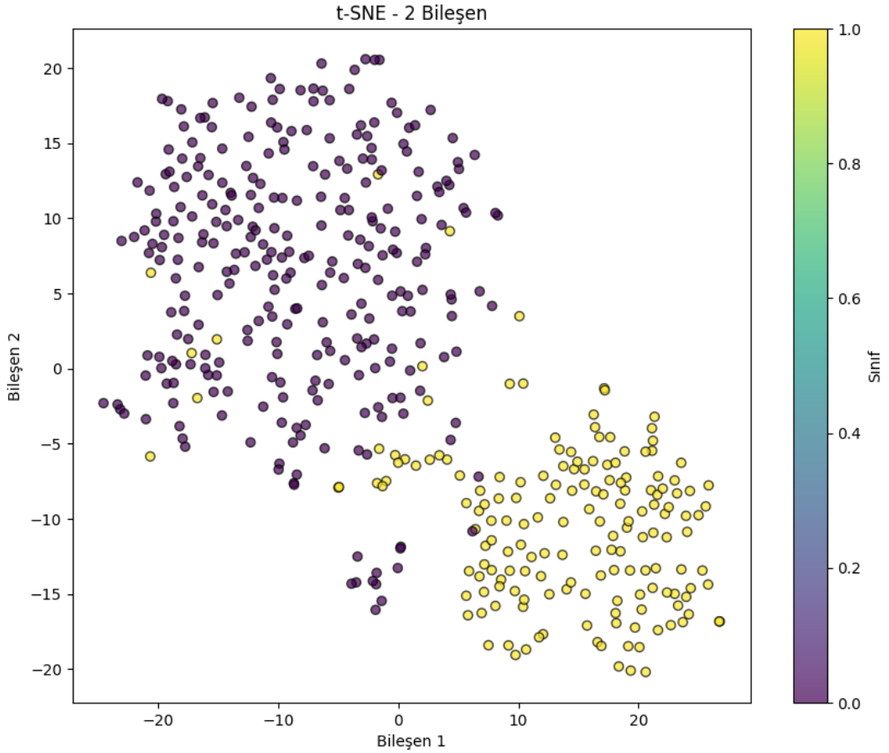
PCA Görselleştirme

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.

LDA Görselleştirme

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

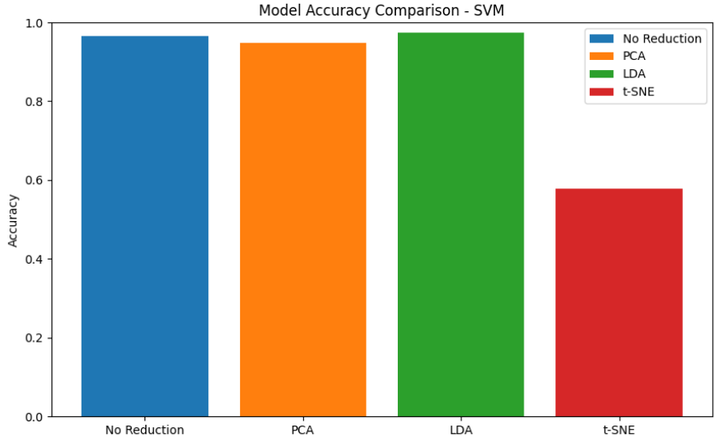


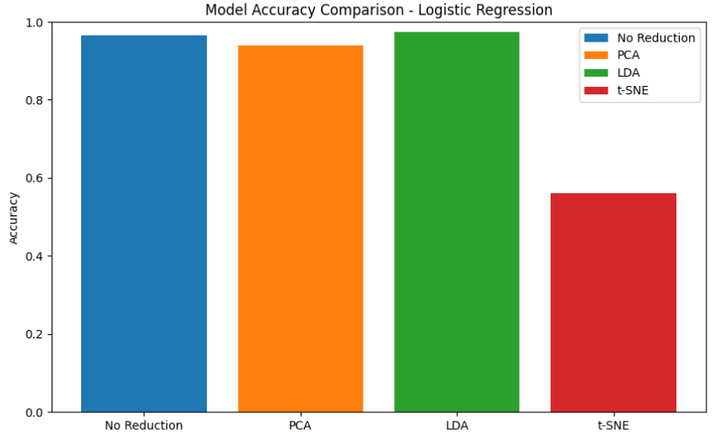


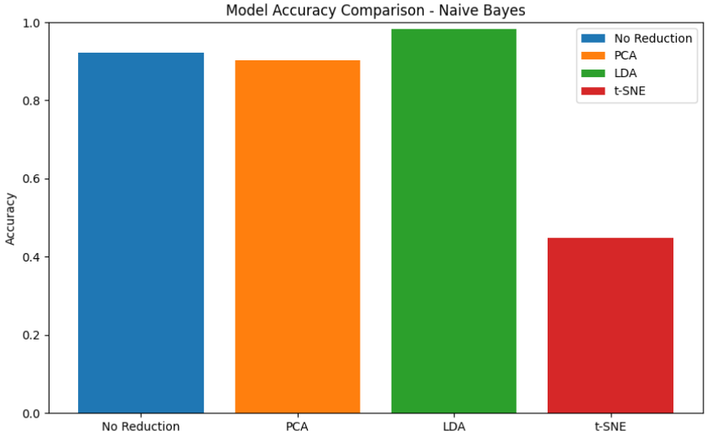
t-SNE Görselleştirme

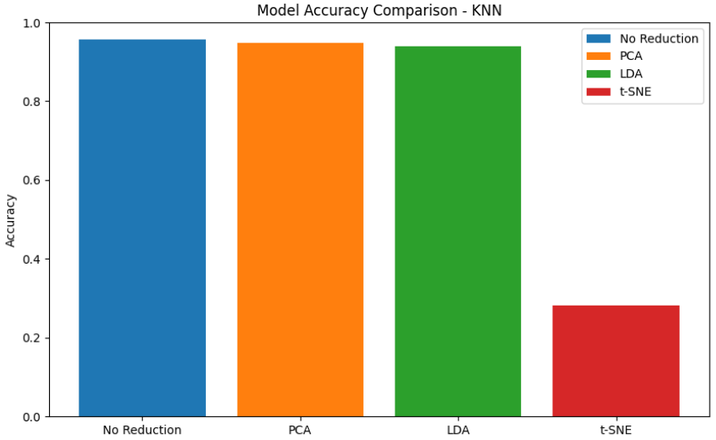
Bu scatter plot, t-SNE kullanılarak iki bileşene indirgenen verinin görselleştirilmesidir. t- SNE, verilerin karmaşık yapısını koruyarak düşük boyutta temsil etmeye çalışır. Bu grafikte, Benign ve Malignant sınıflar arasında belirgin kümeler olduğu görülmektedir, ancak bu kümelerin kenarları arasındaki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

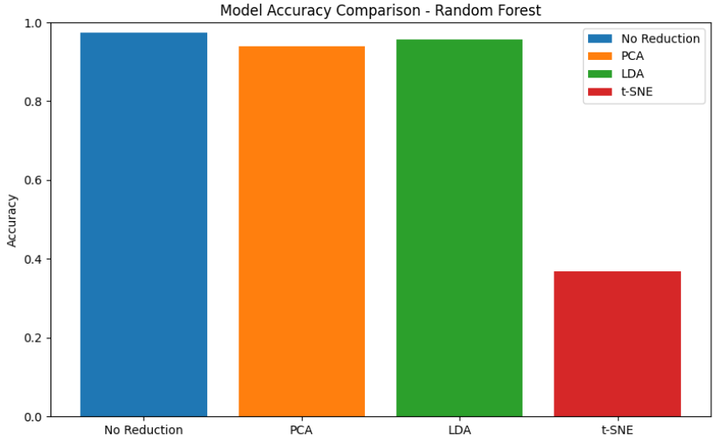
Boyut İndirgeme Tekniklerinin Accuracy Karşılaştırılması





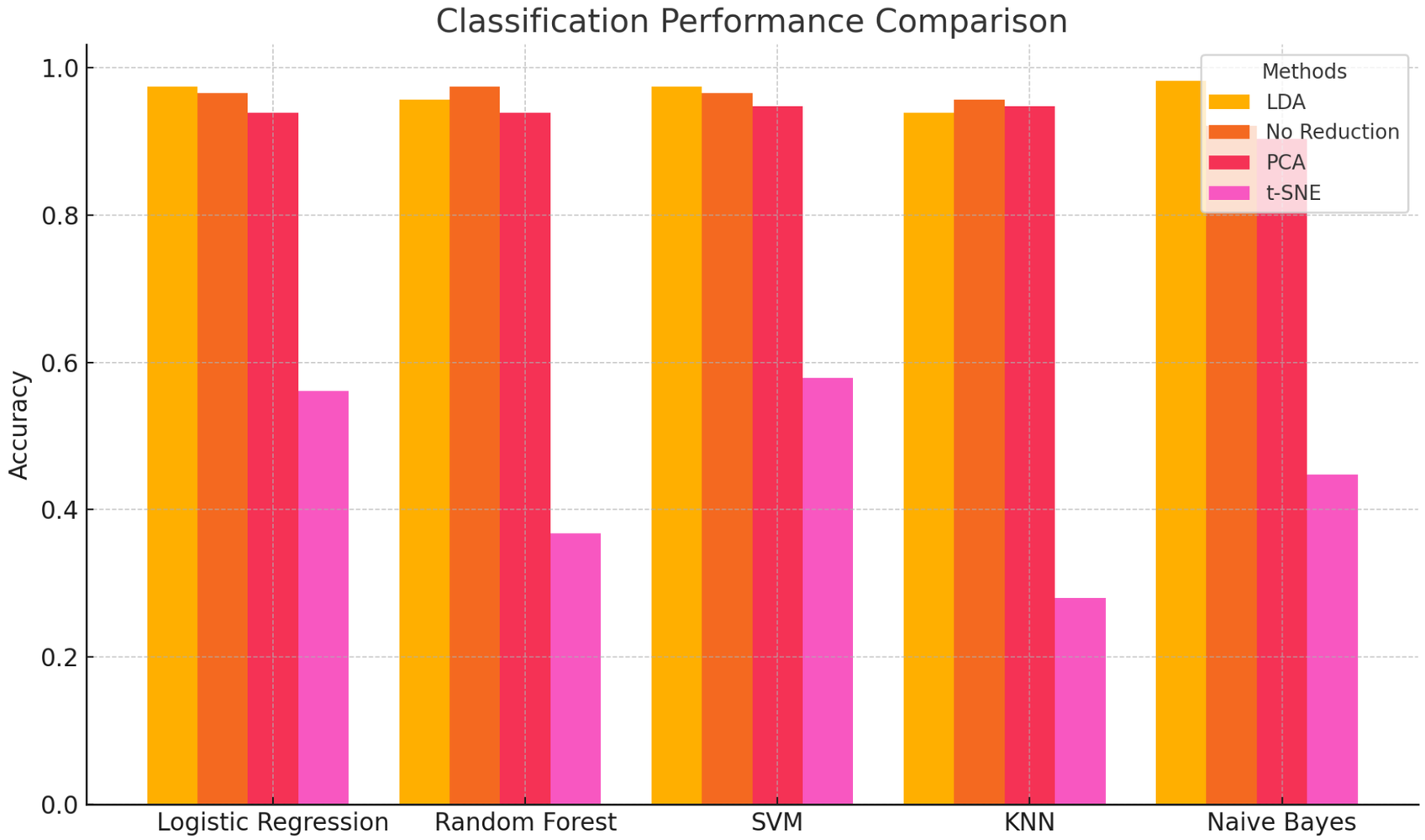






Sınıflandırma Performansının Karşılaştırılması

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.



Sınıflandırma Performansının Karşılaştırılması

Yapılan karşılaştırmada, **LDA (Linear Discriminant Analysis)**, genel olarak en yüksek doğruluk değerlerini sağlamış ve sınıflandırma performansında en başarılı yöntem olarak öne çıkmıştır. Özellikle **Naive Bayes** ve **Logistic Regression** modelleriyle birlikte kullanıldığında, doğruluk oranı %97'nin üzerine çıkmıştır.

No Reduction ve **PCA**, LDA'nın hemen gerisinde yer almış ve makul bir performans göstermiştir. Ancak, **t-SNE**, tüm modeller için belirgin şekilde daha düşük doğruluk oranları sunmuş ve sınıflandırma görevleri için uygun bir yöntem olmadığını göstermiştir.

Bu sonuçlar, boyut indirgeme tekniklerinin seçiminin sınıflandırma performansında önemli bir etkiye sahip olduğunu ve LDA'nın bu bağlamda daha avantajlı olduğunu göstermektedir.

Sonuçlar ve Tartışma

Bu çalışmada, boyut indirgeme tekniklerinin göğüs kanseri sınıflandırma modellerinin performansına etkisini inceledik. LDA, sınıflar arasındaki ayrımı maksimize etmesi nedeniyle en iyi performansı sergiledi. Sınıflandırma modellerinde en yüksek doğruluk oranına bu yöntemle ulaştık. PCA, verideki en büyük varyansı koruyarak boyut indirgeme işlemini başarıyla gerçekleştirdi. t-SNE, verilerin görsel analizinde benzersiz bir araç olarak öne çıktı, ancak sınıflandırma doğruluğunu artırma konusunda diğer yöntemler kadar etkili olmadı.

Use comentialit reduelsire antivans taɒles in thimer altiotiodly, lacvary Meal astonctalletiion in as accury. of Ligints

	PCA	T-SNEEU	t-SMP	LDAE	LDA	
aL	PC%	90%	SV%	SM%	Rıandom	Boosti
:anesst	99%	96%	98%	25%	94%	96%
ectoured	96%	99%	25%	23%	76%	96%
anfeastion	56%	35%	70%	15%	54%	34%
enfeastiod	63%	05%	35%	28%	33%	96%
enfeastiod	56%	75%	77%	76%	35%	35%
enfeastiod	42%	10%	33%	76%	33%	34%
arriaty	43%	18%	35%	18%	33%	33%
enfeascood	95%	16%	38%	272	53%	33%
enfeascod	29%	27%	25%	28%	387	39%
eclution	24%	25%	28%	23%	55%	36%
enfeastiod	34%	07%	83%	20%	21%	20%
enfeascood	36%	15%	83%	14%	53%	35%
enfeascod	1.9%	31%	55%	68%	33%	66%
.unfeascood	49%	97%	95%	27%	33%	38%
enfeascood	114%	17%	94%	37%	76%	75%
enfeastood	27%	12%	85%	20%	33%	35%
anfegscood	55%	07%	15%	98%	69%	35%

Provie Accivine bly inthrsfcads ewy contoltedearning.com.

PCA Neden Önemli ve Başarılı?

Varyansı Maksimize Etme: PCA, veri setindeki özelliklerin varyansını en iyi açıklayan bileşenleri seçerek boyut indirgeme işlemi yapar. Bu, daha az sayıda bileşenle veri setinin büyük bir kısmını temsil etmeyi sağlar.

Gürültüyü Azaltma: PCA, verideki gürültüyü (önemsiz varyansları) filtreleyerek sınıflandırma modellerine daha temiz bir veri sağlar.

Doğrusal Özelliklerde Etkili: PCA, doğrusal ilişkilere sahip veri setlerinde iyi sonuç verir ve hesaplama maliyetlerini düşürür.

No Reduction(Boyut İndirgeme Olmadan) Performans:

Avantajlar: Tüm veri özelliklerini kullandığı için bilgi kaybı yoktur. Özellikle küçük veri setlerinde faydalı olabilir.

Dezavantajlar: Yüksek boyutlu verilerde işlem maliyeti artar. Gereksiz özelliklerin varlığı overfitting'e yol açabilir.

Genel T-SNE LDA Neden Daha İyi?

Sınıflar Arası Ayrımı Maksimize Etme: LDA, sınıflar arasındaki farkı optimize ederek sınıflandırma modellerinde daha yüksek doğruluk sağlar.

Doğrusal Ayrım: Verilerin doğrusal olarak ayrılabilir olduğu durumlarda en etkili yöntemdir.

Küçük Varyanslı Verilerde Avantaj: LDA, az sayıda özelliğe sahip veri setlerinde bile iyi performans gösterir.

Genel T-SNE Neden Başarısız?

Sınıflandırmaya Odaklı Değil: t-SNE, daha çok görselleştirme amaçlıdır ve sınıflandırma performansına katkısı sınırlıdır.

Parametre Hassasiyeti: Yanlış parametre seçimi, sınıflandırma doğruluğunu olumsuz etkileyebilir.

Doğrusal Olmayan İlişkiler: Sadece veriyi görsel olarak temsil etmeyi amaçladığından sınıflandırma algoritmalarıyla uyumlu değildir.

Genel Değerlendirme

LDA

En yüksek doğruluk oranlarını sağlayarak sınıflandırma performansında en başarılı yöntem oldu.

Sınıflar arasındaki ayrımı maksimize etmesi, sınıflandırma modelleri için büyük bir avantaj sundu.

No Reduction

Bilgi kaybı yaşanmadığı için özellikle küçük veri setlerinde etkiliydi.

Ancak, boyut arttıkça işlem maliyetleri ve overfitting riski nedeniyle sınırlamalar gösterdi.

PCA

Verideki en büyük varyansı koruyarak boyut indirgeme işleminde başarılı oldu.

Ancak, sınıf bilgisini dikkate almaması nedeniyle sınıflandırma görevlerinde LDA kadar etkili değildi.

t-SNE

Görselleştirme için mükemmel bir araç olmasına rağmen sınıflandırma doğruluğu açısından zayıf performans sergiledi. Parametrelere hassasiyeti ve hesaplama maliyetleri, sınıflandırma için uygun olmamasına neden oldu.