

**İSTANBUL TOPKAPI ÜNİVERSİTESİ**

**MÜHENDİSLİK FAKÜLTESİ**

**YAZILIM MÜHENDİSLİĞİ**

**Makine Öğrenmesi**

Ferhat Uslu

21040301019

2023-2024 BAHAR YARIYILI

**Pima Indians Diabetes Dataset Analizi ve Modelleme Raporu**

Veri Seti Hakkında Bilgiler

Pima Indians Diabetes Dataset, Pima Kızılderilileri arasında 5 yıl içinde diyabetin başlangıcını tahmin etmeyi amaçlayan bir veri kümesidir. Bu veri seti, 768 gözlem, 8 girdi değişkeni ve 1 çıktı değişkeninden oluşmaktadır.

Metin dosyasından veri okuma işlemi gerçekleştirildi. Veri seti, 8 girdi değişkeni (ham öznitelikler) ve 1 çıktı değişkeni (hedef değişken) olmak üzere toplam 9 sütundan oluşmaktadır. Eksik veri olup olmadığını kontrol etmek adına veri seti üzerinde özet istatistikler incelendi.

Veri seti, %70 eğitim ve %30 test olarak iki bölüme ayrıldı. Bu bölme işlemi, modelin genelleştirilebilirliğini ve doğruluğunu değerlendirmek için yapılmıştır.

**Veri Seti Adı:** Pima Indians Diabetes Dataset  
**Problem Türü:** 5 yıl içinde diyabet başlangıcını tahmin etmek (2-sınıflı sınıflandırma)  
**Gözlem Sayısı:** 768  
**Öznitelik Sayısı:** 8 + 1 Hedef Değişken  
**Hedef Değişken:** Sınıf (0 veya 1)

**Kullanılan Algoritmalar ve Sonuçlar**

**Yorumlar**

**Ön Bilgi:**

Projede ortalama standart sapma, T-test, Normalizasyon, Accuracy (Doğruluk), Öklid - Minkowski karşılaştırması, Maxplotlar (plot fonksiyonu), Sensitivity-Specificity, Precision (Hassasiyet) sınıflandırması, Özdeğer ve Özvektörler, Regresyon işlemleri bulunmaktadır.

**Veri Okuma ve Ön İşleme:**

Veri metin dosyasından okunuyor. Öznitelikler ve hedef değişken ayrılıyor. Veri seti eğitim ve test setlerine ayrılıyor. Normalizasyon işlemi yapılıyor,

**Model Eğitimi ve Değerlendirme:**

Çoklu Doğrusal Regresyon, Multinominal Lojistik Regresyon ve Naive Bayes sınıflandırıcıları eğitiliyor. Her model için tahminler yapılıyor ve performans metrikleri (doğruluk, F1 skoru) hesaplanıyor. Karar Ağacı sınıflandırıcıları eğitiliyor ve görselleştiriliyor.

**Bileşen Analizi:**

PCA ve LDA kullanılarak veri seti boyutu azaltılıyor. PCA ve LDA bileşenleri görselleştiriliyor. Özniteliklerin önem sırası belirleniyor.

**Uzaklık Hesaplama:**

Öklid ve Minkowski uzaklıkları hesaplanıyor. İki vektör arasındaki uzaklıklar görselleştiriliyor.

**Veri İstatistikleri:**

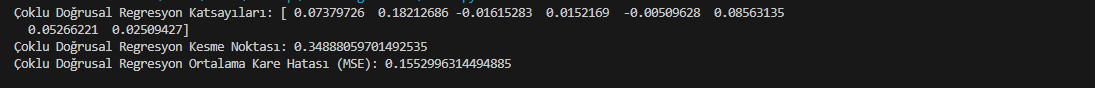
Özniteliklerin ortalama ve standart sapma değerleri hesaplanıyor.

**Genel Yorumlar:**

Kod oldukça kapsamlı ve birçok farklı model ve teknik kullanılarak veri seti üzerinde analiz yapılıyor. Performans metrikleri ve karışıklık matrisleri kullanılarak her bir modelin performansı değerlendiriliyor. PCA ve LDA gibi boyut azaltma teknikleriyle veri seti görselleştiriliyor. Uzaklık hesaplama ve öznitelik istatistikleri gibi ek bilgilendirici analizler de kodda yer alıyor.

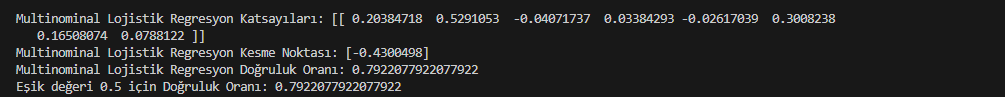
**Çoklu Doğrusal Regresyon Analizi**

Çoklu doğrusal regresyon modeli, eğitim veri seti üzerinde eğitildi ve test veri seti üzerinde tahminler yapıldı. Modelin performansı, Ortalama Kare Hatası kullanılarak değerlendirildi.



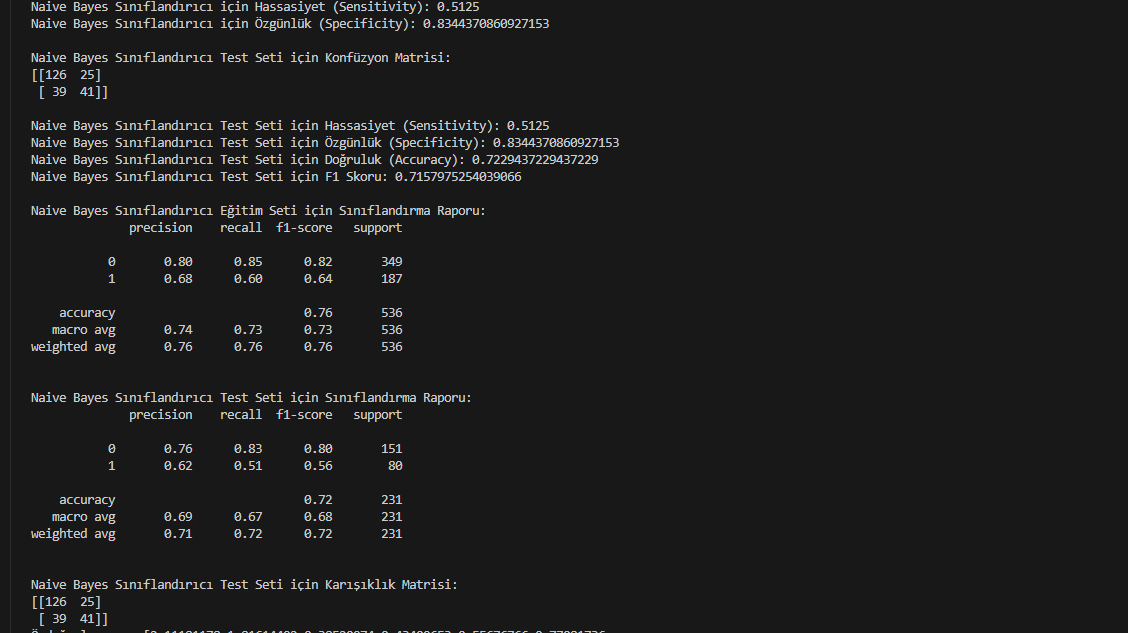
**Multinominal Lojistik Regresyon Analizi**

Multinominal lojistik regresyon modeli, eğitim veri seti üzerinde eğitildi ve test veri seti üzerinde tahminler yapıldı. Modelin doğruluğu, Accuracy Score metriği kullanılarak değerlendirildi.

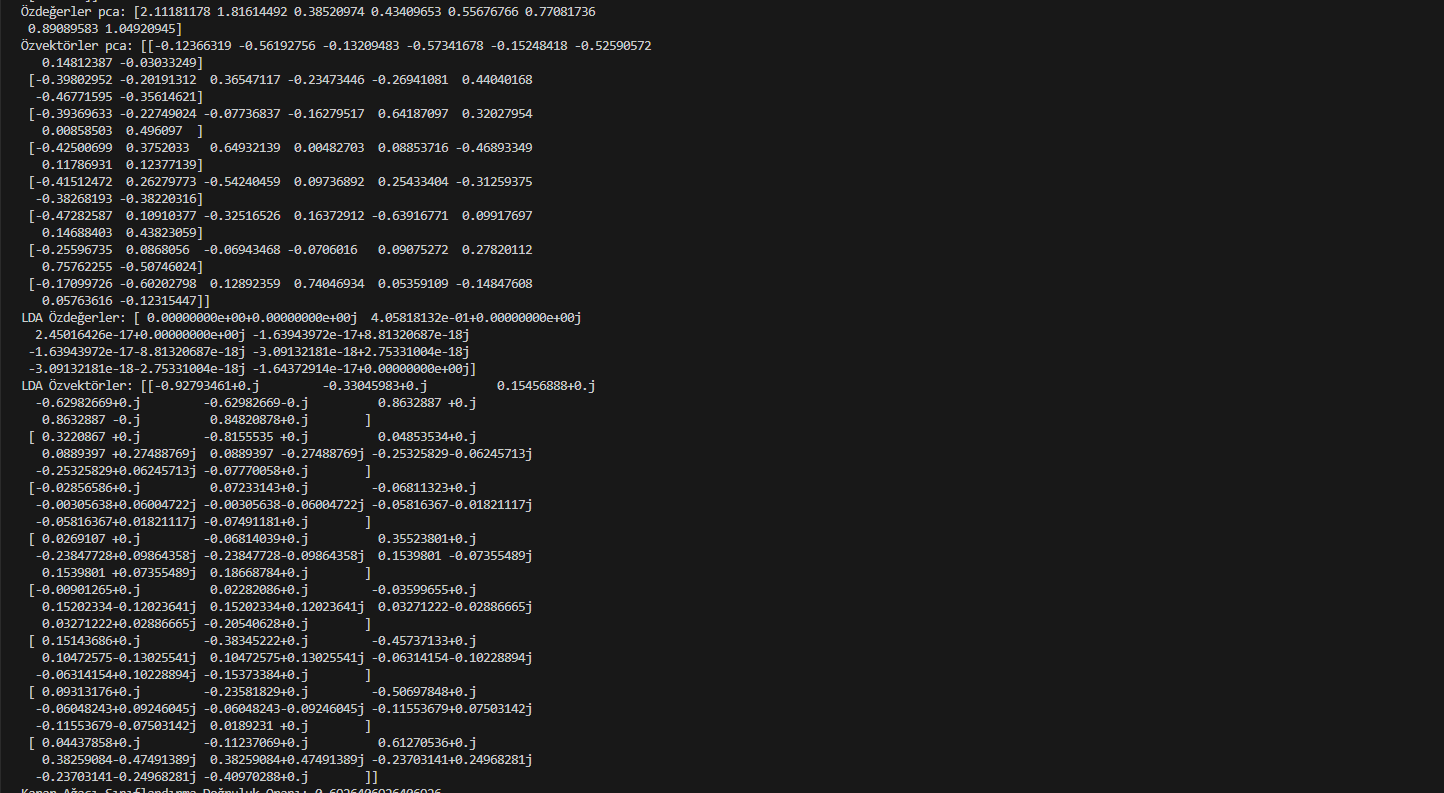


**Navie Bayes Sınıflandırıcısı**

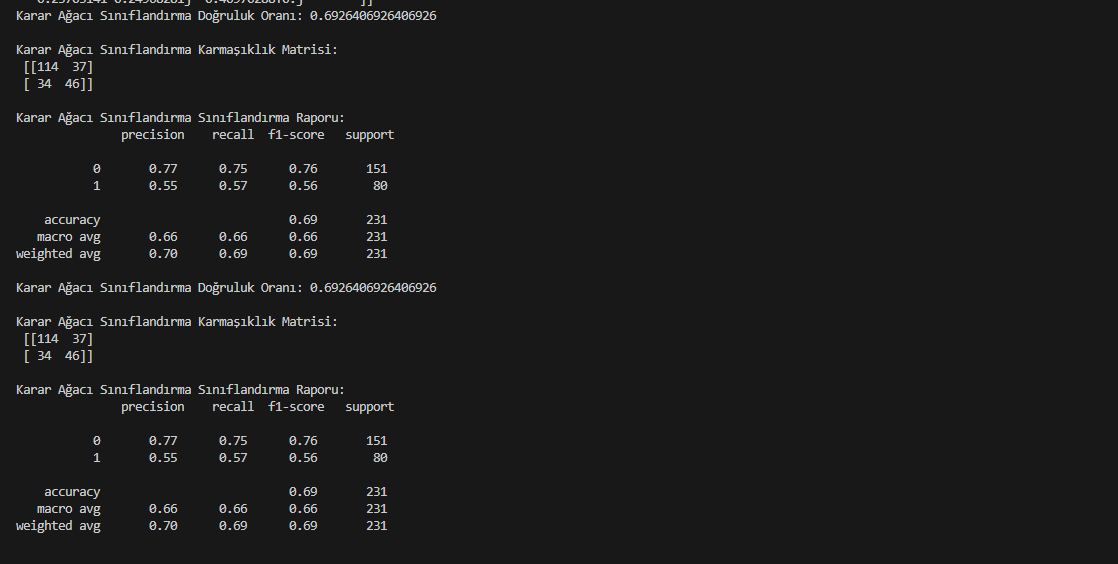
Naive Bayes sınıflandırıcısı, eğitim veri seti üzerinde eğitildi ve test veri seti üzerinde tahminler yapıldı. Modelin performansı, Hassasiyet, Özgünlük, Doğruluk, F1 Skoru gibi metriklerle değerlendirildi.



**Özdeğerler ve Özvektörler**

****

**Karar Ağacı Sınıflandırma Raporu- Karışıklık Matrisi**

****

**Özniteliklerin Önem Sırası (PCA-LDA)**

PCA ve LDA, boyut indirgeme teknikleri olarak kullanıldı. PCA ile öznitelikler arasındaki varyansın en yüksek olduğu bileşenler belirlendi. LDA ise sınıflar arası ayrımın en iyi olduğu bileşenler belirlendi. Her iki yöntemle elde edilen bileşenler, sınıflandırma sonuçlarına göre görselleştirildi.

**Ortalama:**

Özniteliklerin ortalaması, her bir özniteliğin veri setindeki genel eğilimini gösterir. Yüksek bir ortalama, o özniteliğin genellikle daha yüksek değerlere sahip olduğunu gösterebilir.

**Standart Sapma:**

Standart sapma, öznitelik değerlerinin ne kadar değişken olduğunu gösterir. Yüksek bir standart sapma, o özniteliğin değerlerinin genellikle ortalamanın üzerinde veya altında değişken olduğunu gösterir. Düşük bir standart sapma, öznitelik değerlerinin genellikle ortalamanın yakınında olduğunu gösterir. Bu bilgilere dayanarak, hangi özniteliklerin modelinizi etkileyeceğini ve hangi özniteliklerin daha az önemli olabileceğini belirleyebilirsiniz. Özellikle, yüksek standart sapma değerlerine sahip öznitelikler, modelinizi eğitirken daha fazla dikkate alınabilir.

**Özniteliklerin Ortalamaları:**

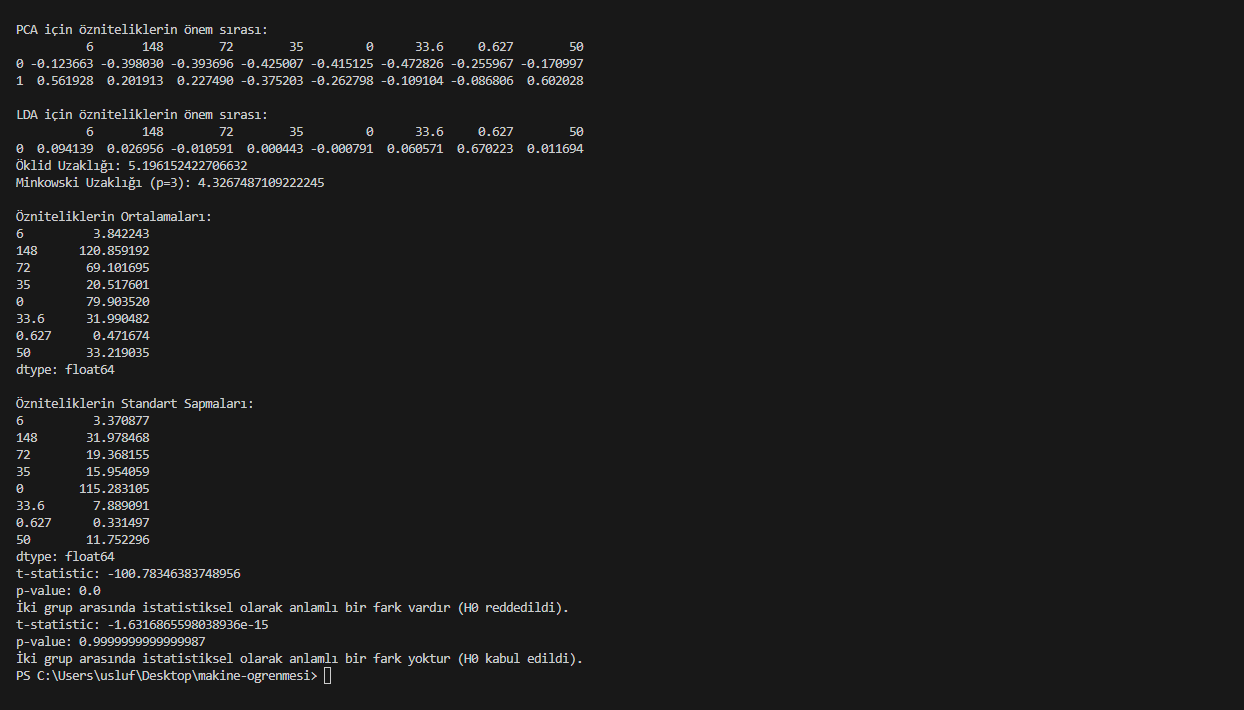
Ortalama değerler, her bir özniteliğin genel eğilimini göstermektedir

**Özniteliklerin Standart Sapmaları:**

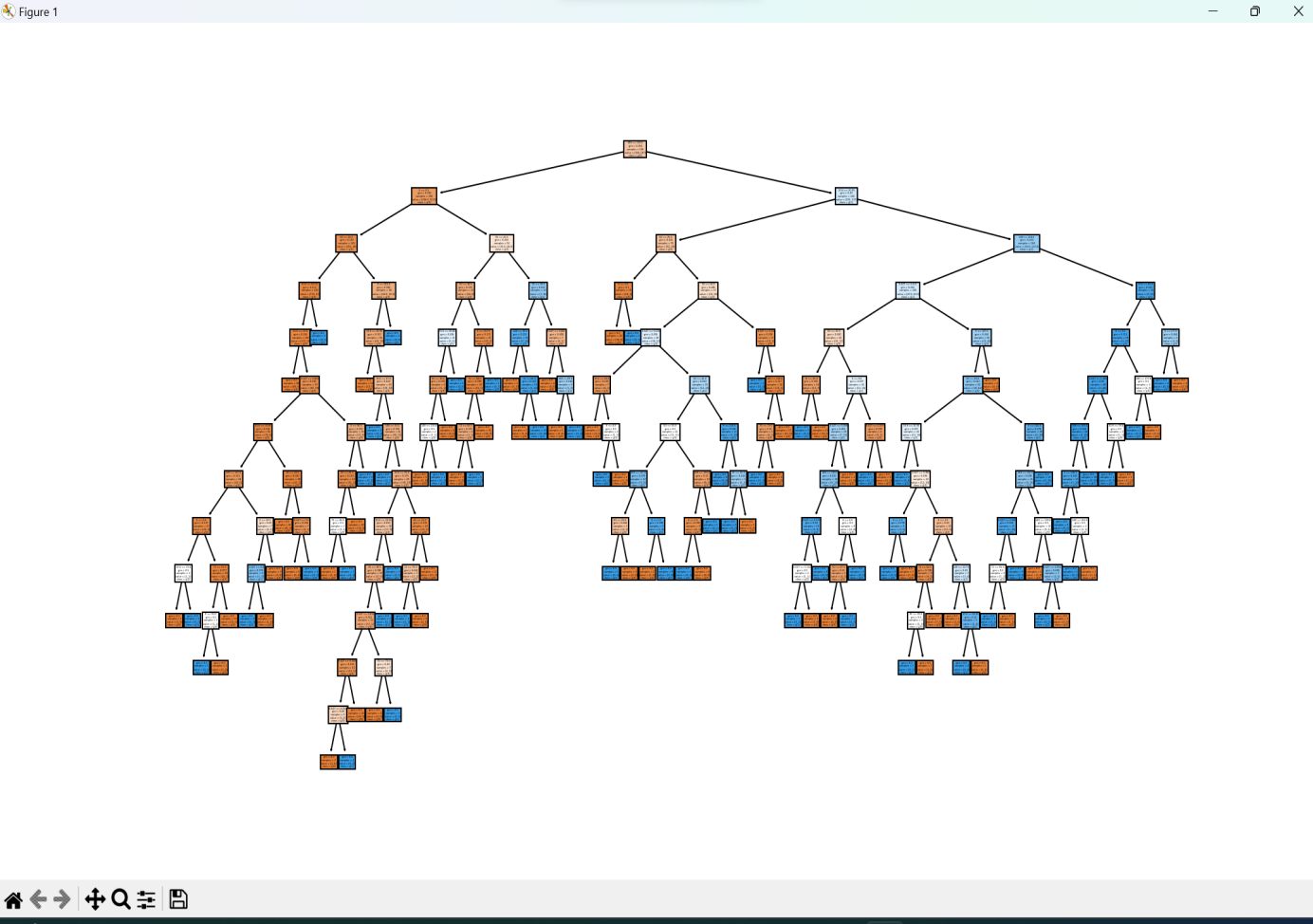
Standart sapma değerleri, öznitelik değerlerinin ne kadar değişken olduğunu göstermektedir.

**Datatipini Verme:**

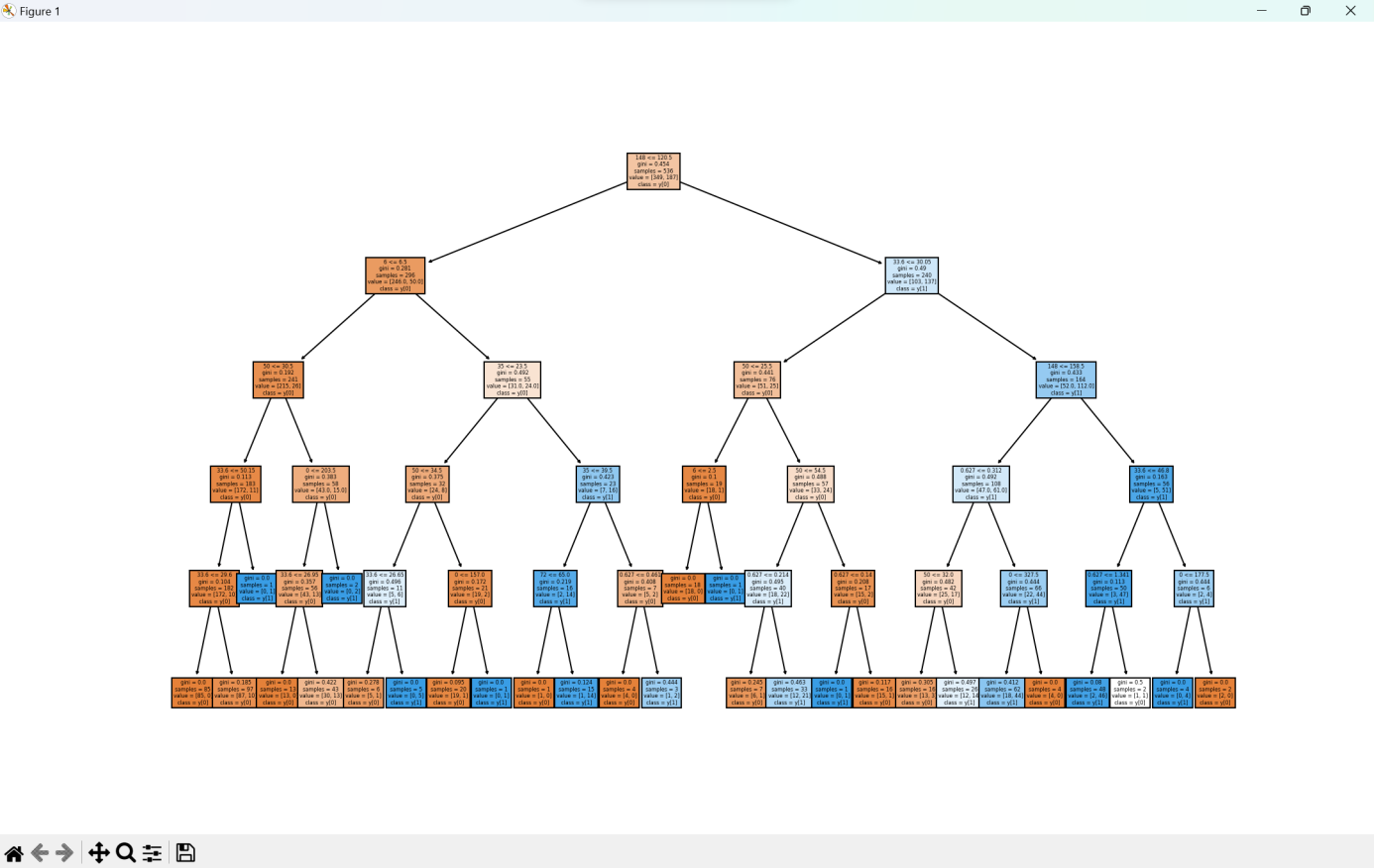
Ortalama ve standart sapma değerleri, bir özniteliğin genel dağılımını ve değişkenliğini gösterir. Bu değerlerin datatipini vermesi, bu özniteliklerin veri setinizde ne kadar yaygın ve değişken olduğunu anlamamıza yardımcı olur. Ancak, bu değerlerin datatipini verip vermemesi tamamen analizinize bağlıdır. Bu tür değerler genellikle veri bilimi projelerinde veya analizlerde kullanılır ve bu tür bilgilerin paylaşılması genellikle istenilen bir durumdur.

****

**Karar Ağacı Modeli**

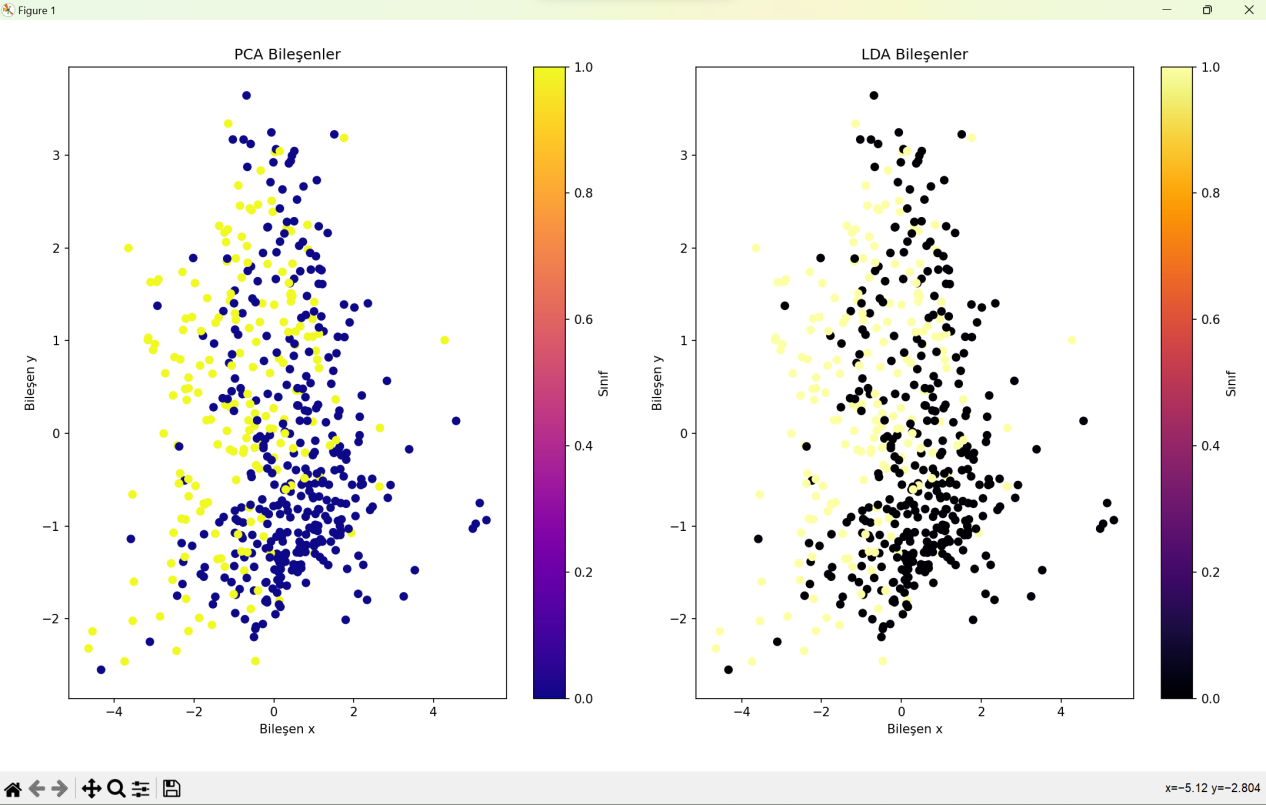
****

**Karar Ağacı Modeli (Budanmış Hali)**

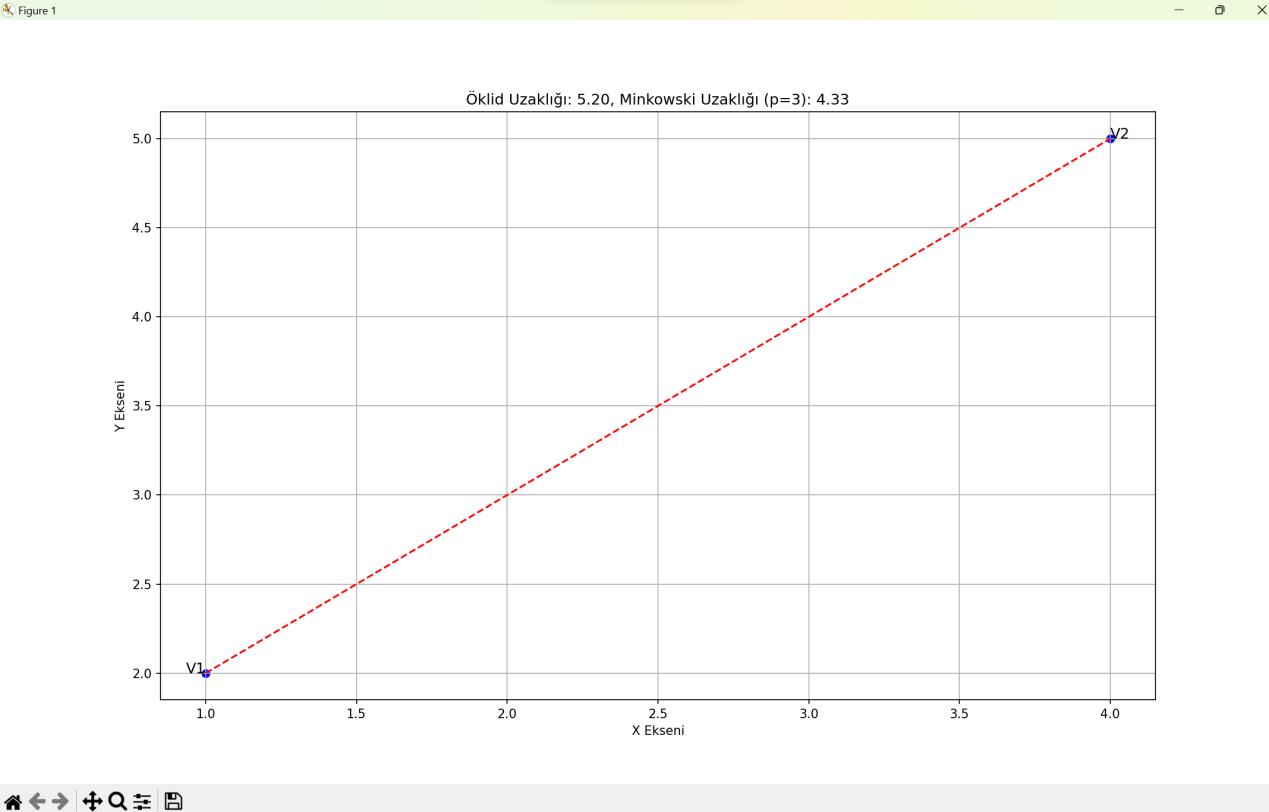
****

Karar ağacı modeli, önce standart bir şekilde eğitildi ve görselleştirildi. Daha sonra karar ağacı modeli budama (pruning) işlemi ile sınırlı bir derinliğe sahip olarak tekrar eğitildi. Modelin performansı, Accuracy Score, Karmaşıklık Matrisi ve Sınıflandırma Raporu gibi metriklerle değerlendirildi.

**PCA ve LDA Bileşenlerinin Görseli**

Minkowski

**Öklid-Minkowski Uzaklığı**

****