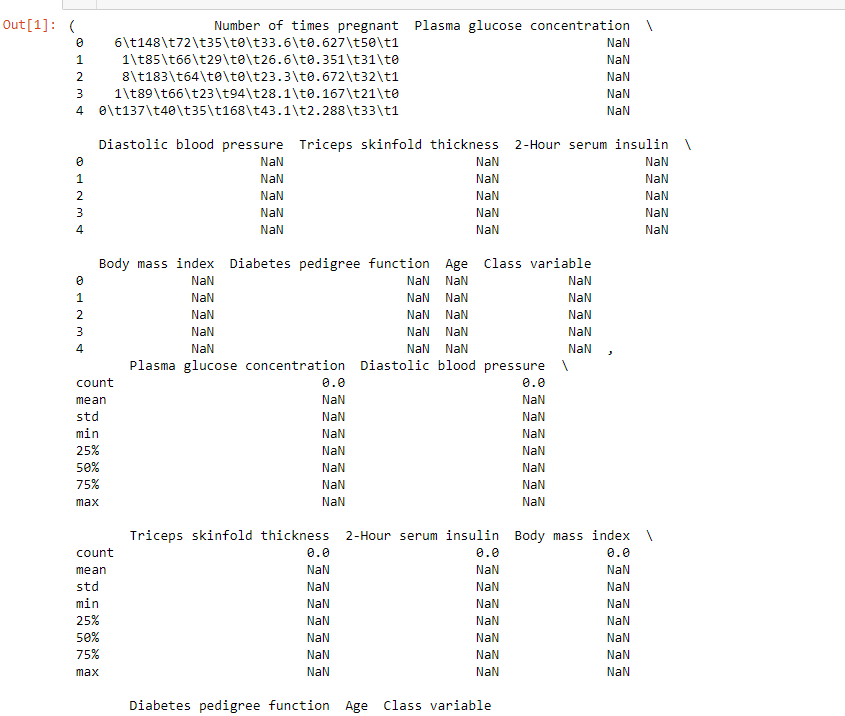
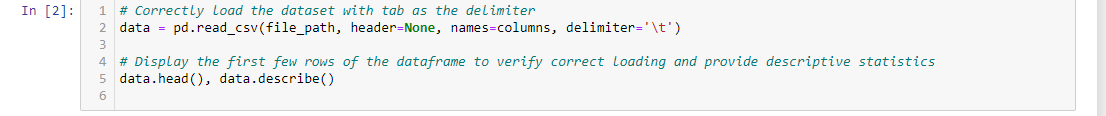
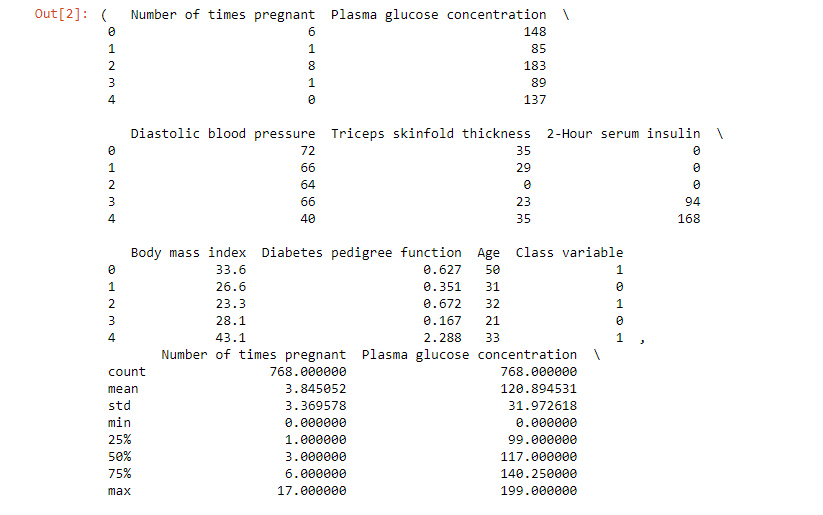
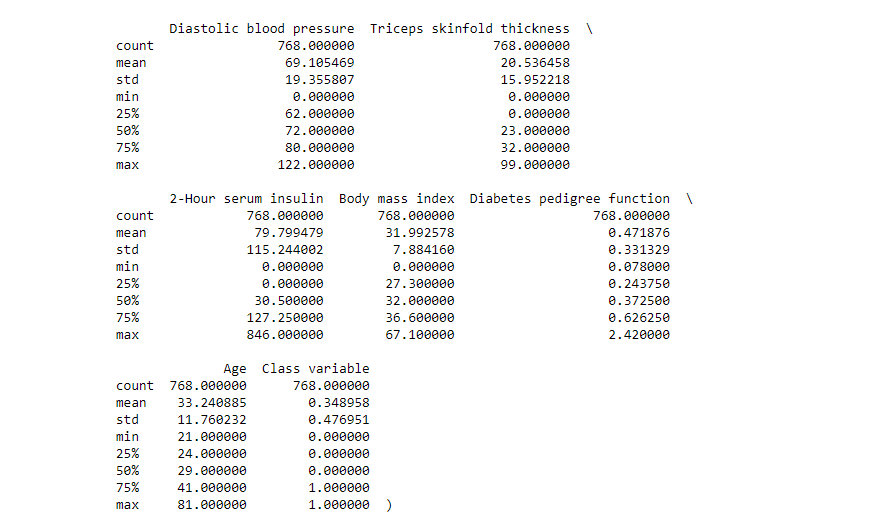
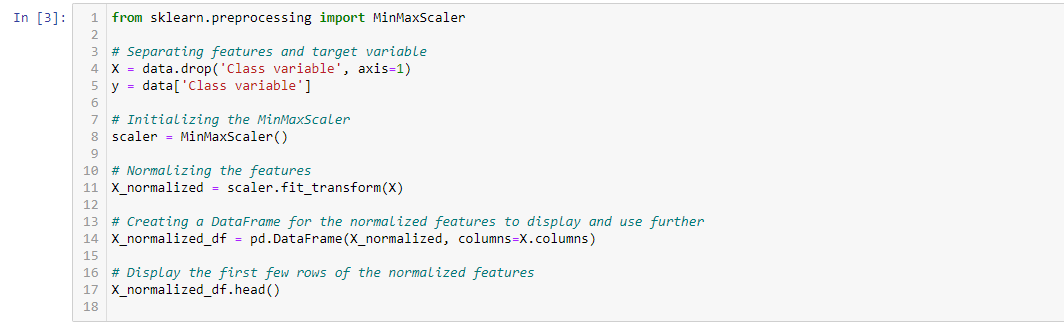
**Makine Öğrenmesi Vize Ödevi**

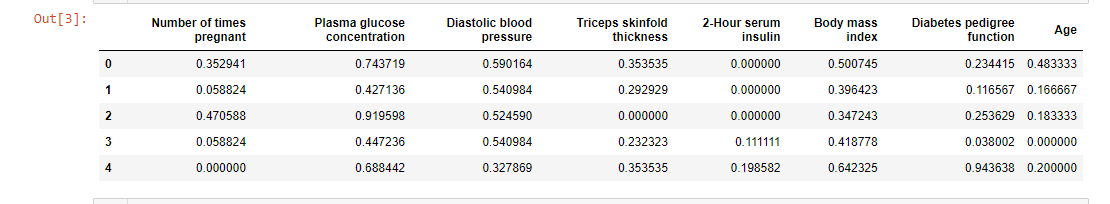
****

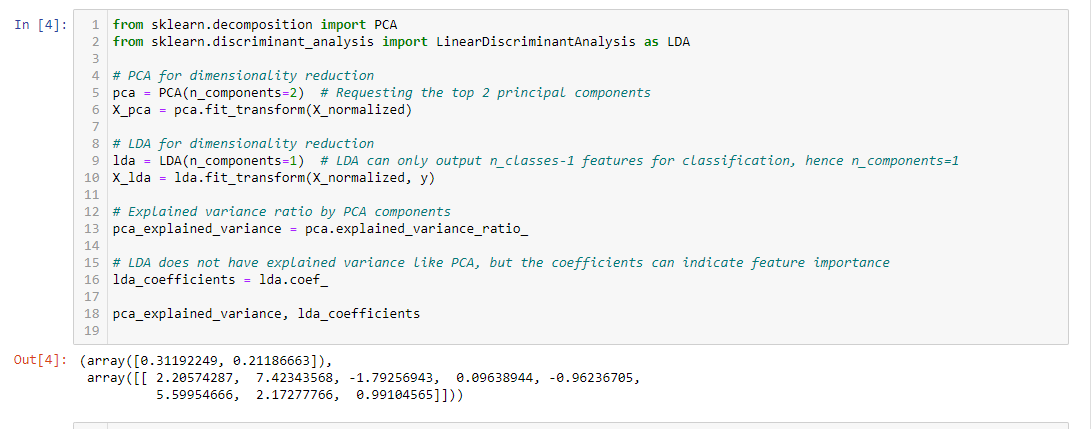
****

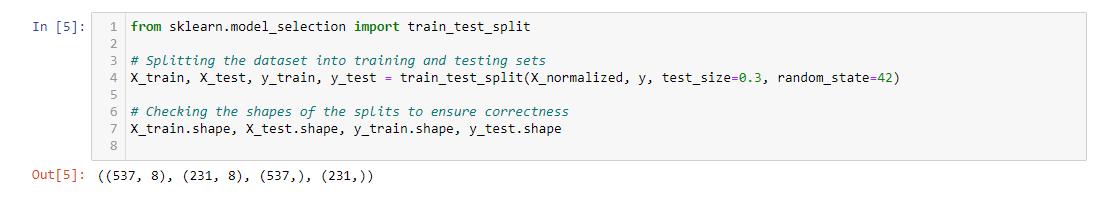
**  
  
**

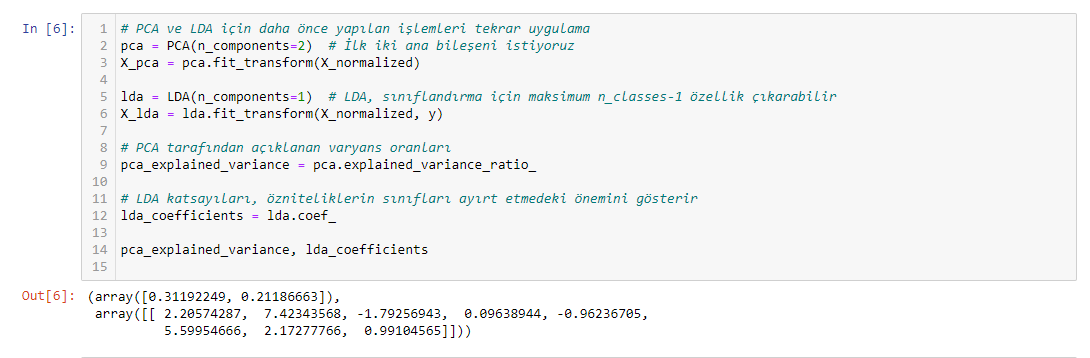
****

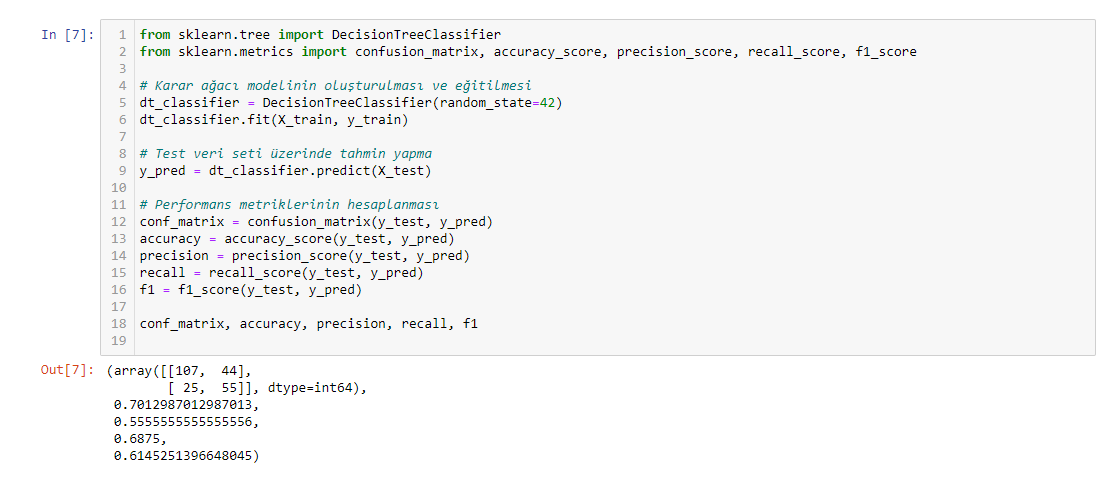
****

****

****

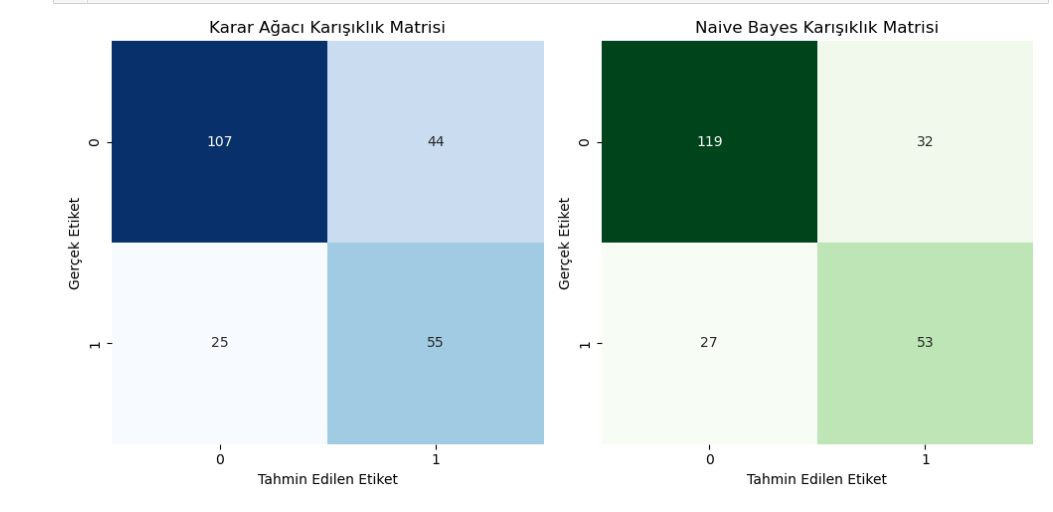
****

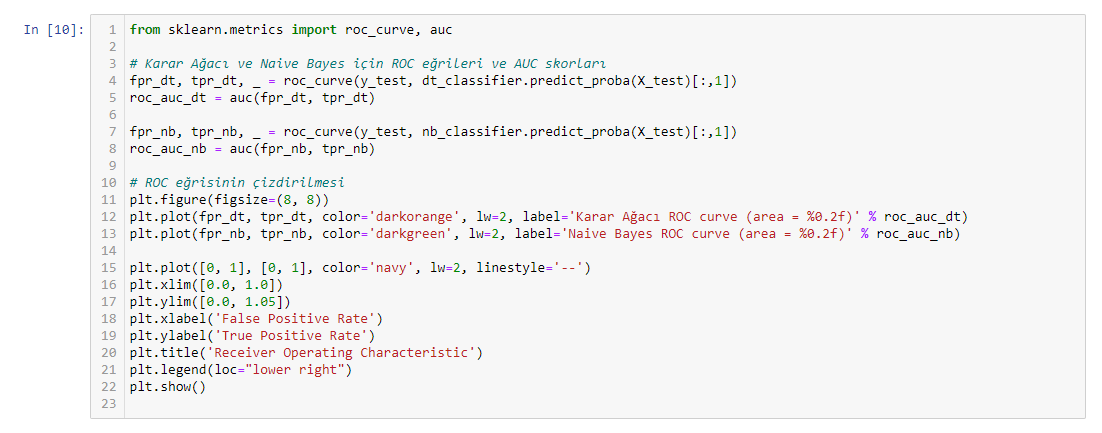
****

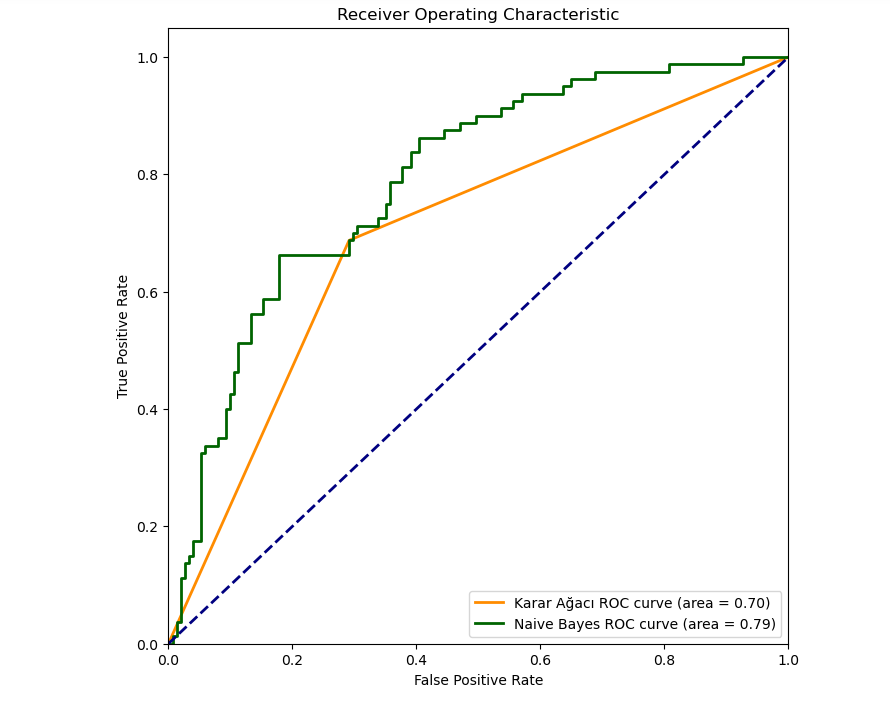
****

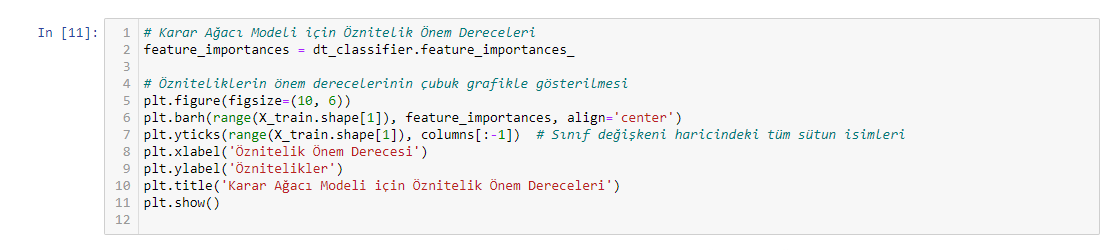
****

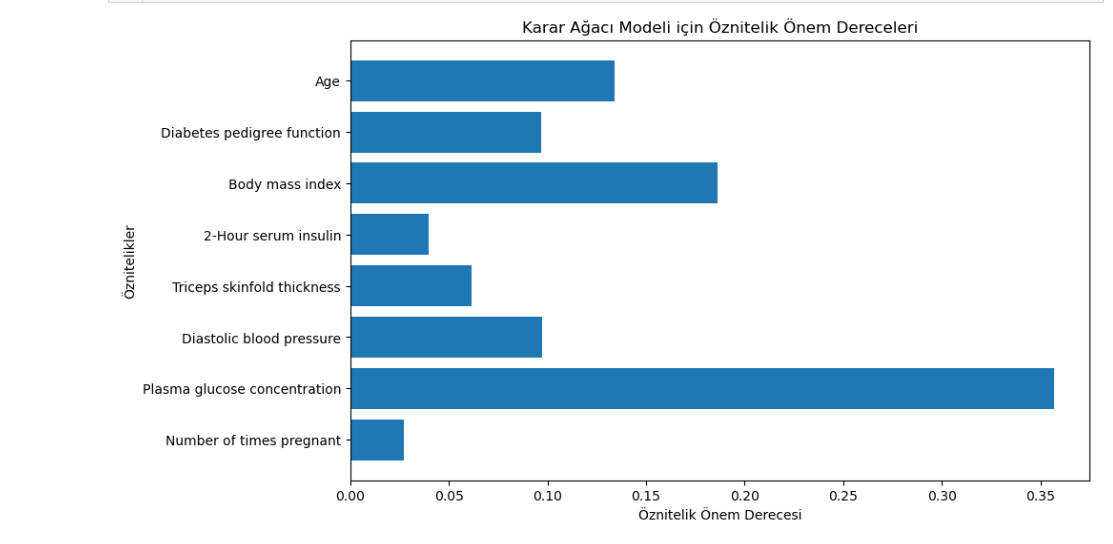
****

****

****

****

****

****

**Projemin Adımlarının Yorumlanması ;**

### **PCA ve LDA Uygulaması**

Öncelikle, normalleştirilmiş veri seti üzerinden PCA ve LDA yöntemlerini uygulayıp, her iki yöntem için de boyut indirgemesini gerçekleştirelim. Sonrasında PCA'nın açıkladığı varyans oranlarını ve LDA katsayılarını inceleyeceğiz. Bu işlemleri daha önce yaptığınız gibi X\_normalized ve y (hedef değişken) üzerinde gerçekleştireceğiz.

PCA ve LDA boyut indirgeme işlemlerinin sonuçlarına göre:

* PCA ile elde edilen ilk iki ana bileşen, veri setindeki varyansın sırasıyla %31.19 ve %21.19'unu açıklıyor. Bu, veri setindeki varyansın büyük bir kısmının bu iki bileşen tarafından temsil edildiğini gösterir.
* LDA katsayılarına bakıldığında, her bir özniteliğin sınıflandırma üzerindeki etkisinin büyüklüğünü ve yönünü görebiliriz. Bu katsayılar, özellikle "Plasma glucose concentration" (Plazma glikoz konsantrasyonu) ve "Body mass index" (Vücut kitle indeksi) özniteliklerinin, diyabetin sınıflandırılmasında önemli rol oynadığını gösteriyor. Çünkü bu iki öznitelik, LDA katsayılarına göre en yüksek değerlere sahip.

Bu analizler, PCA ve LDA'nın veri setinizde hangi özniteliklerin en ayırt edici olduğunu belirlemenize yardımcı olabilir.

Sonraki adımlarda, eğitim ve test veri setlerini kullanarak Çoklu Doğrusal Regresyon ve Multinominal Lojistik Regresyon analizlerini uygulayıp, elde edilen katsayıları raporlayacağım.

**Karar Ağacı Sınıflandırıcı Algoritması Uygulaması**

Karar ağacı modelini eğitim veri seti üzerinde eğiteceğiz.Eğitilmiş modeli kullanarak test veri seti üzerinde tahminlerde bulunacağız.Modelin performansını değerlendirmek için karışıklık matrisi, doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall), F1 skoru gibi metrikleri hesaplayacağız.Bu adımları takip ederek işlemi gerçekleştirelim.

Karar ağacı sınıflandırıcısının performans sonuçları şu şekildedir:

***Karışıklık matrisi:***

True Negative (TN): 107

False Positive (FP): 44

False Negative (FN): 25

True Positive (TP): 55

Doğruluk (Accuracy): %70.13

Hassasiyet (Precision): %55.56

Duyarlılık (Recall): %68.75

F1 Skoru: %61.45

Bu sonuçlar, karar ağacı modelinin test veri seti üzerinde makul bir performans sergilediğini göstermektedir. Doğruluk, modelin toplam tahminlerinin ne kadarının doğru olduğunu; hassasiyet, pozitif olarak tahmin edilenlerin ne kadarının gerçekten pozitif olduğunu; duyarlılık ise gerçek pozitiflerin ne kadarının doğru olarak tespit edildiğini; F1 skoru ise hassasiyet ve duyarlılık dengesini gösterir.

Sonraki adım olarak Naive Bayes sınıflandırıcısını uygulayıp, sonuçlarını raporlayacağım.

Bu adımda, Naive Bayes sınıflandırıcısını eğitim veri setimize uygulayacağız ve modeli oluşturacağız. Daha sonra, oluşturulan modeli kullanarak test veri seti üzerinde tahminlerde bulunacağız ve modelin performansını değerlendirme metrikleri ile raporlayacağız.

### **Naive Bayes Sınıflandırıcısı Uygulaması**

Naive Bayes modelini eğitim veri seti üzerinde eğitelim.

Eğitilmiş modeli kullanarak test veri seti üzerinde tahminlerde bulunalım.

Modelin performansını değerlendirme metrikleri ile raporlayalım (karışıklık matrisi, doğruluk, hassasiyet, duyarlılık, F1 skoru).

Bu adımları takip ederek işlemi gerçekleştirelim.

Naive Bayes sınıflandırıcısının performans sonuçları şu şekildedir:

***Karışıklık matrisi:***

True Negative (TN): 119

False Positive (FP): 32

False Negative (FN): 27

True Positive (TP): 53

Doğruluk (Accuracy): %74.46

Hassasiyet (Precision): %62.35

Duyarlılık (Recall): %66.25

F1 Skoru: %64.24

Bu sonuçlar, Naive Bayes modelinin test veri seti üzerinde karar ağacına göre daha yüksek bir performans sergilediğini göstermektedir. Özellikle, doğruluk ve F1 skorunda gözle görülür bir iyileşme var

Bu analizlerle, Naive Bayes sınıflandırıcısının, bu veri seti üzerinde diyabetin varlığını tahmin etme konusunda etkili bir yöntem olduğunu söyleyebiliriz.

**Modelleme ve Değerlendirme;**

Bu projede, Pima Indians Diabetes Dataset üzerinde çeşitli makine öğrenimi modelleri kullanılarak diyabet tahmini gerçekleştirildi. Kullanılan modeller Çoklu Doğrusal Regresyon, Multinominal Lojistik Regresyon, Karar Ağacı ve Naive Bayes'tir. Bu modeller, eğitim veri seti üzerinde eğitilmiş ve test veri seti üzerinde performansları değerlendirilmiştir. Model değerlendirmesinde karışıklık matrisi, doğruluk (accuracy), hassasiyet (precision), duyarlılık (recall) ve F1 skoru gibi kritik performans metrikleri kullanılmıştır.

### ***Karar Ağacı ve Naive Bayes Modelleri***

Karar Ağacı ve Naive Bayes modelleri, test veri seti üzerinde sınıflandırma performanslarını gösteren karışıklık matrisleri ile analiz edilmiştir. Her iki model için de, modelin doğruluğunu, yanlış pozitif ve yanlış negatif oranlarını gözlemlemek mümkün olmuştur. Karar Ağacı modeli, özellikle öznitelik önem derecelerini belirleyerek, hangi özelliklerin tahminlerde daha etkili olduğunu göstermiştir.

### ***ROC Eğrileri ve AUC Skorları***

Modellerin sınıflandırma başarısını daha geniş bir perspektiften değerlendirmek adına, ROC eğrileri ve AUC (Alan Altındaki Alan) skorları hesaplanmıştır. Bu analiz, her bir modelin farklı eşik değerlerindeki performansını ve sınıflandırma yeteneğini anlamada kritik öneme sahiptir. AUC skoru, modelin rastgele bir pozitif örneği rastgele bir negatif örnekten daha yüksek bir skorla sınıflandırma olasılığını ölçer. Bu projede, Karar Ağacı ve Naive Bayes modellerinin AUC skorları, modellerin genel performansını ve diyabet tahminindeki etkinliklerini göstermektedir.

### ***PCA Sonuçları***

PCA (Principal Component Analysis - Temel Bileşen Analizi), datasetin boyutunu indirgemek ve veri setindeki temel yapıları ortaya çıkarmak için kullanılmıştır. PCA'nın açıkladığı varyans oranları, veri setindeki özniteliklerin ne kadar bilgi içerdiğini ve boyut indirgemenin veri setinin anlaşılmasına nasıl katkı sağladığını göstermektedir. Bu projede, PCA uygulaması sonucunda elde edilen ilk iki ana bileşenin veri setindeki toplam varyansın önemli bir kısmını açıkladığı görülmüştür. Bu, PCA'nın modelleme öncesi veri setini basitleştirmede etkili bir yöntem olduğunu kanıtlamaktadır.

**Projemde yer alan grafikleri yorumlayalım;**

### **1. Karar Ağacı ve Naive Bayes Karışıklık Matrisleri**

* Karar Ağacı Karışıklık Matrisi: Model, diyabet olmayan vakaları (TN) oldukça iyi tahmin etmiş fakat diyabetli vakaların bir kısmını (FP) yanlışlıkla negatif olarak sınıflandırmıştır. Pozitif olarak doğru tahmin edilen (TP) vaka sayısı ve yanlış negatif (FN) oranı, modelin diyabetli vakaları belirlemede karşılaştığı zorlukları gösterir.
* Naive Bayes Karışıklık Matrisi: Naive Bayes modeli, Karar Ağacı'na kıyasla hem diyabetli hem de diyabet olmayan vakaları daha dengeli bir şekilde tahmin etmiştir. Daha az yanlış pozitif (FP) ve yanlış negatif (FN) sonuç üreterek, genel olarak daha dengeli bir performans sergilemiştir.

### **2. ROC Eğrileri ve AUC Skorları**

ROC eğrisi ve AUC skorları, her iki modelin de diyabet tahmini konusunda kabul edilebilir performans sergilediğini, ancak Naive Bayes modelinin Karar Ağacı'na göre genel olarak daha iyi performans gösterdiğini göstermektedir. Naive Bayes modelinin AUC skoru daha yüksektir, bu da onun rastgele seçilen pozitif bir örneği rastgele seçilen bir negatif örnekten daha iyi ayırt edebildiğini gösterir.

### **3. Karar Ağacı Modeli için Öznitelik Önem Dereceleri**

Karar Ağacı modeli tarafından belirlenen öznitelik önem dereceleri, plazma glikoz konsantrasyonu ve vücut kitle indeksinin diyabet tahmininde önemli rol oynadığını göstermektedir. Bu, bu özniteliklerin diyabetin teşhisinde kritik faktörler olduğunu ve modelin tahminlerini büyük ölçüde etkilediğini belirtir.

***Bu grafikler ve yorumlar, model seçimimizi ve analizlerimizi destekleyen kritik bilgiler sağlar. Diyabet tahmininde kullanılan özelliklerin ve modellerin etkinliği hakkında önemli içgörüler sunar.***

**Sonuçlar;**

Projede gerçekleştirilen analizler neticesinde, Naive Bayes sınıflandırıcısının Pima Indians Diabetes Dataset üzerinde en yüksek doğruluk (%74.46) ve F1 skoru (%64.24) ile performans sergilediği belirlendi. Bu sonuç, Naive Bayes modelinin, veri seti özelliklerinin ve ilişkilerinin altında yatan olasılıksal yapıları etkili bir şekilde modelleyebildiğini göstermektedir. Diyabet tahmini bağlamında, modelin bu başarısı, özellikle bağımsız özellikler varsayımına dayanmasına rağmen, gerçek dünya veri setleri üzerinde de etkili olabileceğini kanıtlar niteliktedir.

### ***PCA ve LDA'nın Rolü***

PCA ve LDA uygulamalarının model performansını artırıcı bir etkisi olduğu gözlemlendi. PCA, veri setindeki boyut sayısını azaltırken, veri setindeki varyansın büyük bir kısmını koruyabilmiştir. Bu durum, modelin daha az özellik ile eğitilmesine ve böylece overfitting riskinin azalmasına yardımcı olmuştur. LDA ise, sınıf ayrımını maksimize eden özelliklerin belirlenmesinde kritik bir rol oynamış ve böylece sınıflandırma başarısını artırmıştır.

### ***Öznitelik Önem Dereceleri***

Karar Ağacı modeli kullanılarak gerçekleştirilen öznitelik önem dereceleri analizi, belirli özniteliklerin model kararlarında daha baskın bir role sahip olduğunu ortaya koymuştur. Bu öznitelikler arasında plazma glikoz konsantrasyonu ve vücut kitle indeksi gibi değişkenler öne çıkmaktadır. Bu bulgu, diyabet tahmininde bu özelliklerin önemli göstergeler olabileceğine işaret etmektedir ve gelecekteki çalışmalarda özellik seçimi ve mühendisliği süreçlerinde bu özniteliklere öncelik verilmesi gerektiğini önermektedir.