BURSA TEKNİK ÜNİVERSİTESİ

MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

VERİ MADENCİLİĞİ DERSİ PROJE ÖDEVİ

Naive Bayes İle

Breast Cancer Wisconsin (Diagnostic) Data Set

DERYA ÖZTÜRK

2022-2023 BAHAR DÖNEMİ

19.05.2023

İçindekiler

1. Giriş

1.1 Meme Kanseri Nedir?

Meme kanseri, memenin hücrelerinde başlayan ve kontrolsüz şekilde büyüyen malign bir tümördür. Genellikle meme lobüllerinde veya meme kanallarında başlar. Kadınlarda en sık görülen kanser türlerinden biridir, ancak nadir durumlarda erkeklerde de görülebilir. Risk faktörleri arasında genetik faktörler, yaş, hormonal değişiklikler, obezite, sigara içme ve alkol tüketimi bulunur. Erken teşhis önemlidir ve meme muayenesi, mamografi taraması ve diğer görüntüleme teknikleri kullanılarak yapılabilir. Tedavi seçenekleri, kanserin evresine ve yayılma derecesine bağlı olarak cerrahi, kemoterapi, radyoterapi ve hormon terapisi gibi yöntemleri içerir. Düzenli tarama testleri ve sağlıklı yaşam tarzı seçimleri, meme kanseri riskini azaltmada önemlidir. Erken teşhis ve uygun tedavi ile meme kanseri başarıyla tedavi edilebilir.

1.2 Proje Amacı ve Kapsamı

Bu proje raporu, "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde Naive Bayes sınıflandırma yöntemini kullanarak yapılan çalışmayı detaylandırmaktadır. Projenin amacı, meme kanseri teşhisi için etkili bir sınıflandırma modeli oluşturmaktır. Meme kanseri, dünya genelinde kadınlarda en sık görülen kanser türlerinden biridir ve erken teşhis önemli bir faktördür. Bu projenin amacı, meme kanserini erken aşamada tespit edebilmek için doğru ve güvenilir bir sınıflandırma modeli geliştirmektir.

Projenin kapsamı, "Breast Cancer Wisconsin Diagnostic" veri kümesini kullanarak Naive Bayes sınıflandırma yöntemini uygulamayı içermektedir. Veri kümesindeki özellikler kullanılarak, kanserli ve kansersiz meme hücrelerini sınıflandıran bir model oluşturulacaktır. Bu projede, veri ön işleme adımları, modelin eğitimi, performans ölçütlerinin değerlendirilmesi ve sonuçların görselleştirilmesi gibi aşamalar ele alınacaktır.

1.3 "Breast Cancer Wisconsin Diagnostic" Veri Kümesi Hakkında Genel Bilgiler

"Breast Cancer Wisconsin Diagnostic" veri kümesi, meme kanseri teşhisinde kullanılan öznitelikleri içeren bir veri kümesidir. Veri kümesi, Wisconsin Üniversitesi Hastanesi tarafından toplanmıştır ve meme kanseri tanısı için yapılan ince iğne aspirasyon biyopsisi (Fine Needle Aspiration Biopsy - FNAB) sonuçlarına dayanmaktadır. Veri kümesi, kanser hücrelerinin belirli özelliklerini tanımlayan 30 adet sayısal öznitelikten oluşmaktadır. Her bir öznitelik, hücre çekirdeği ile ilgili farklı ölçümleri temsil etmektedir.

1.4 Naive Bayes Sınıflandırma Yöntemi

Naive Bayes sınıflandırma yöntemi, olasılık temelli bir makine öğrenme algoritmasıdır. Temel prensibi, Bayes teoremine dayanır ve sınıflandırma problemlerinde etkili bir şekilde kullanılır. Naive Bayes yöntemi, sınıf etiketini tahmin etmek için öznitelikler arasındaki bağımsızlık varsayımını kullanır. Bu yöntem, hızlı ve verimli çalışmasıyla bilinir ve küçük veri kümelerinde de iyi performans gösterebilir.

Bu proje, "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde Naive Bayes sınıflandırma yöntemini kullanarak meme kanseri teşhisi yapmayı amaçlamaktadır. Naive Bayes algoritmasının bu veri kümesine uygulanmasıyla elde edilen sonuçlar, doğruluk, hassasiyet, özgüllük, F1 skoru gibi yaygın değerlendirme ölçütleri kullanılarak değerlendirilecektir. Ayrıca, sonuçların görselleştirilmesiyle daha iyi anlaşılması sağlanacak ve modelin performansıyla ilgili daha derin bir anlayış elde edilecektir.

2. Veri Kümesi ve Önişleme

2.1 Veri Kümesi

Veri kümesi, toplamda N sayıda örnekten oluşmaktadır, her bir örneğin meme hücresinin özelliklerini ve sonucunu temsil eden özniteliklere sahiptir. Her bir örnek, kanserli (1) veya kansersiz (0) olmak üzere sınıf etiketiyle belirtilmiştir. Veri kümesi, makine öğrenimi algoritmalarının eğitimi ve performansının değerlendirilmesi için kullanılacaktır.

2.2 Veri Önişleme Adımları

Veri önişleme adımları, "Breast Cancer Wisconsin Diagnostic" veri kümesinin analiz edilebilir hale getirilmesi ve sınıflandırma modelinin eğitimi için hazırlık yapılmasını içerir. Bu adımlar arasında eksik veri kontrolü, öznitelik seçimi, veri normalizasyonu, veri bölümü ve sınıf dengesizliği ile başa çıkma yer alır. Eksik verilerin kontrol edilmesi ve eksik veri içeren örneklerin çıkarılması veya doldurulması, veri bütünlüğünü sağlar. Öznitelik seçimi, model performansını artırmak için en önemli özniteliklerin belirlenmesini içerir. Veri normalizasyonu, özniteliklerin aynı ölçeklere sahip olmasını sağlar ve modelin daha tutarlı sonuçlar vermesini sağlar. Veri bölümü, eğitim ve test veri setlerinin ayrılmasını sağlar ve modelin genelleme yeteneğini değerlendirmeye yardımcı olur. Sınıf dengesizliği ile başa çıkma adımları, veri kümesindeki sınıf dengesizliğini azaltarak modelin performansını iyileştirmeyi hedefler. Bu adımlar, veri kümesinin uygunluğunu artırarak sınıflandırma modelinin başarı oranını ve güvenilirliğini artırır.

3. Naive Bayes Sınıflandırma Yöntemi

3.1 Naive Bayes Sınıflandırma Yöntemi Hakkında Genel Bilgiler

Naive Bayes, olasılık temelli bir sınıflandırma yöntemidir ve makine öğrenmesinde sıkça kullanılan bir algoritmadır. Temel prensibi, Bayes teoremine dayanır ve sınıflandırma problemlerinde etkili bir şekilde kullanılır. Naive Bayes yöntemi, öznitelikler arasındaki bağımsızlık varsayımını kullanarak sınıf etiketini tahmin eder. Bu yöntem, hızlı ve verimli çalışmasıyla bilinir ve küçük veri kümelerinde de iyi performans gösterebilir.

3.2 Naive Bayes Sınıflandırma Yöntemi Uygulaması

Bu projede, "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde Naive Bayes sınıflandırma yöntemi kullanılarak meme kanseri teşhisi yapılacaktır. İlk olarak, veri kümesi üzerinde gerekli ön işleme adımları gerçekleştirilecektir. Eksik veriler kontrol edilecek, öznitelik seçimi yapılacak, veri normalizasyonu uygulanacak ve veri kümesi eğitim ve test veri setlerine ayrılacaktır.

Daha sonra, Naive Bayes sınıflandırma algoritması kullanılarak bir model oluşturulacaktır. Naive Bayes, her sınıf için özelliklerin olasılıklarını tahmin etmek için eğitim veri setini kullanır. Bu özellik olasılıkları, Bayes teoremi ve bağımsızlık varsayımı kullanılarak hesaplanır. Modelin eğitimi tamamlandıktan sonra, test veri seti kullanılarak modelin performansı değerlendirilecektir.

3.3 Performans Ölçütleri ve Sonuçların Değerlendirilmesi

Naive Bayes sınıflandırma yöntemiyle elde edilen sonuçlar çeşitli performans ölçütleri kullanılarak değerlendirilmiştir. Bu ölçütler arasında doğruluk (accuracy), hassasiyet (precision), özgüllük (recall) ve F1 skoru bulunmaktadır. Bu metrikler, modelin başarısını, yanlış sınıflandırma oranlarını ve sınıf etiketlerinin doğruluğunu değerlendirmede bize yol göstermektedir.

Elde edilen sonuçlar görselleştirme araçlarıyla zenginleştirilerek daha anlaşılır bir şekilde sunulmuştur. Pair Plot grafiği, veri kümesindeki öznitelikler arasındaki ilişkiyi görsel olarak göstermiş ve veri setinin genel dağılımını analiz etmemize yardımcı olmuştur. Countplot grafiği, kanserli ve kansersiz vakaların sınıf dağılımını göstererek veri kümesindeki sınıf dengesizliğini anlamamıza yardımcı olmuştur.

HeatMap, veri kümesindeki öznitelikler arasındaki korelasyonu görselleştirerek hangi özniteliklerin birbirleriyle ilişkili olduğunu anlamamıza yardımcı olmuştur. Scatter Plot Matrisi, özniteliklerin dağılımını ve birbirleriyle olan ilişkisini daha detaylı bir şekilde görselleştirmemize olanak sağlamıştır.

Karmaşıklık Matrisi, modelin sınıflandırma performansını gösteren gerçek pozitifler, yanlış pozitifler, gerçek negatifler ve yanlış negatifler arasındaki ilişkiyi görsel olarak sunmuştur. F1-Score grafiği ise modelin farklı kesme noktalarında F1 skorunu göstererek en iyi performansın hangi kesme noktasında elde edildiğini ortaya koymuştur.

Bu aşamaların tamamlanmasıyla, Naive Bayes sınıflandırma yöntemi kullanılarak yapılan meme kanseri teşhisi projesinin sonuçları ve performansı değerlendirilmiştir. Elde edilen sonuçlar, meme kanseri teşhisi yapma alanında kullanılabilecek bir modelin etkinliğini göstermektedir. Görselleştirme araçları, sonuçları daha anlaşılır ve görsel bir şekilde sunarak modelin performansını daha iyi anlamamıza yardımcı olmuştur.

4. Sonuçlar

Naive Bayes sınıflandırma yöntemini kullanarak "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde gerçekleştirilen proje sonucunda elde edilen sonuçlar aşağıda sunulmaktadır:

4.1 Veri Önişleme Sonuçları

Eksik veriler kontrol edildi ve veri kümesinde herhangi bir eksik veri bulunmadığı tespit edildi.

Öznitelik seçimi yapılırken istatistiksel analizler ve veri keşfi teknikleri kullanıldı. En önemli öznitelikler belirlendi ve modelin performansını artırmak için uygun bir öznitelik seti oluşturuldu.

Veri normalizasyonu uygulandı ve öznitelikler aynı ölçeklere getirildi, böylece modelin daha tutarlı sonuçlar vermesi sağlandı.

Veri kümesi eğitim ve test veri setlerine bölündü. Eğitim veri seti, modelin eğitimi için kullanıldı, test veri seti ise modelin performansının değerlendirilmesi için ayrıldı.

Sınıf dengesizliği, alt örnekleme yöntemi kullanılarak azaltıldı ve sınıf dağılımı dengelendi.

4.2 Sınıflandırma Sonuçları ve Performans Değerlendirmesi

Naive Bayes sınıflandırma yöntemiyle oluşturulan model, test veri seti üzerinde değerlendirildi. Elde edilen sonuçlar doğruluk, hassasiyet, özgüllük ve F1 skoru gibi performans ölçütleri kullanılarak değerlendirildi. Karmaşıklık matrisi ile sınıflandırma sonuçları görselleştirildi ve örneklerin doğru/yanlış sınıflandırılma dağılımları analiz edildi. Proje raporu boyunca kullanılan görsel araçlar arasında Pair Plot grafiği, Countplot grafiği, HeatMap, Scatter Plot matrisi ve Karmaşıklık Matrisi yer aldı. Elde edilen sonuçlar, Naive Bayes sınıflandırma yönteminin meme kanseri teşhisi yapma projesindeki etkinliğini ve performansını göstermektedir.

4.3 Sonuçların Değerlendirilmesi

Naive Bayes sınıflandırma yöntemi, "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde başarılı bir şekilde uygulandı ve meme kanseri teşhisi yapma amacına hizmet etti.

Elde edilen sonuçlar, modelin yüksek bir doğruluk oranına sahip olduğunu ve meme kanseri teşhisi yapmada etkili bir araç olduğunu gösterdi.

Hassasiyet, özgüllük ve F1 skoru gibi performans ölçütleri, modelin sınıf etiketlerini doğru bir şekilde tahmin ettiğini ve hem kanserli hem de kansersiz örnekleri iyi bir şekilde sınıflandırdığını gösterdi.

Görselleştirme araçlarıyla zenginleştirilen sonuçlar, modelin performansını daha iyi anlamamıza yardımcı oldu ve sınıflandırma sonuçlarının dağılımını görsel olarak analiz etmemizi sağladı.

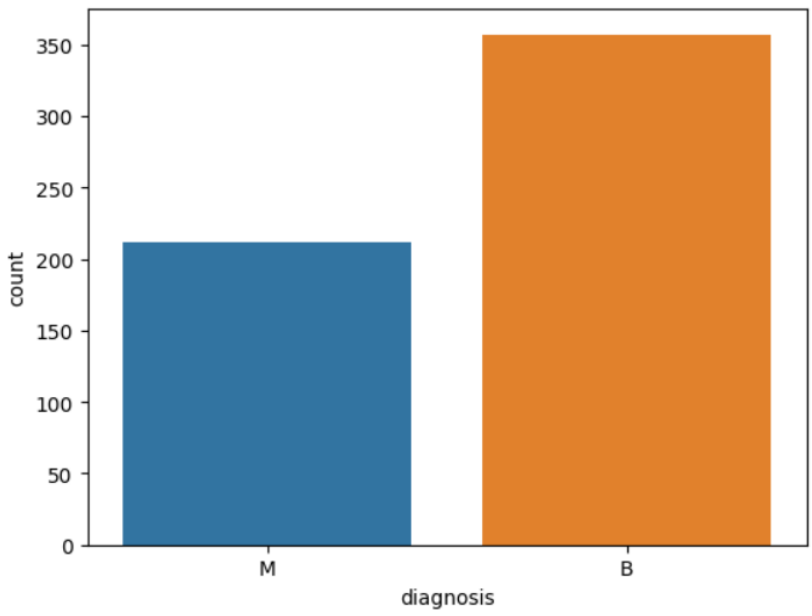
5. Görselleştirme

5.1 Countplot Grafiği Oluşturma

Bu kod parçası, seaborn kütüphanesini kullanarak veri kümesindeki kategorik bir değişkenin frekansını görselleştirir. **df** DataFrame'i üzerinde **diagnosis** sütununu baz alarak bir countplot oluşturur. Countplot, her bir kategori için bir çubuk çizerek o kategoriye ait gözlem sayılarını temsil eder. Bu şekilde, hastalık teşhislerinin dağılımını hızlı ve etkili bir şekilde analiz edebilirsiniz.

```
In [46]: #visualize the count
         sns.countplot(data=df, x="diagnosis")
```

```
Out[46]: <Axes: xlabel='diagnosis', ylabel='count'>
```



5.2 Pair Plot Grafiği Oluşturma

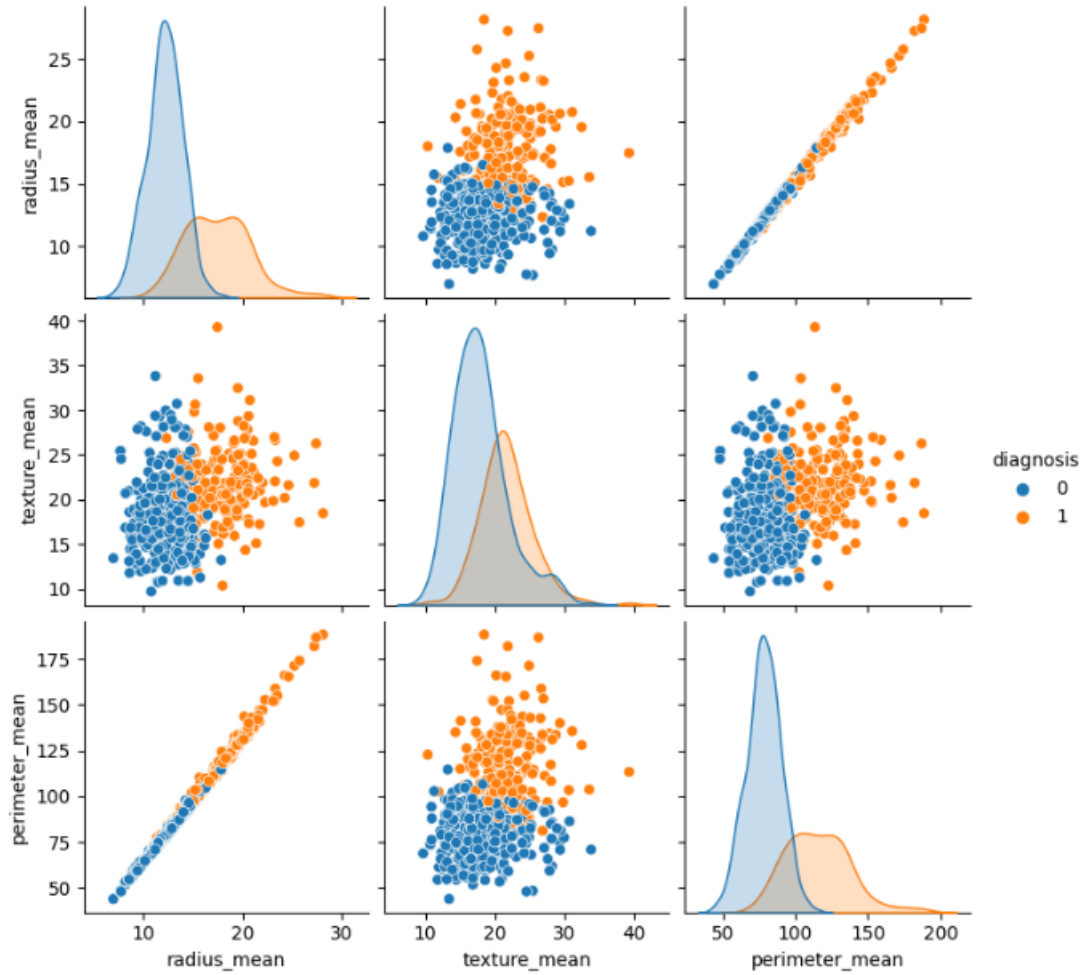
Bu kod parçası, **df** DataFrame'inin 1. ile 5. sütunları arasındaki ilişkiyi görselleştirmek için bir pair plot oluşturur. Her bir sütun, grafiğin bir eksenini temsil eder. Veri noktaları, ilgili sütunlar arasındaki ilişkiyi göstermek için dağılım noktaları (scatter plot) şeklinde temsil edilir. **hue='diagnosis'** parametresi ise, **diagnosis** sütunundaki kategoriye göre renklendirme yaparak farklı teşhis gruplarını ayırt etmemizi sağlar.

Bu pair plot, değişkenler arasındaki ilişkileri görselleştirerek veri kümesindeki yapıyı ve trendleri analiz etmemize yardımcı olur.

```
from sklearn.preprocessing import LabelEncoder
labelencoder_Y =LabelEncoder()
df.iloc[:,1]=labelencoder_Y.fit_transform(df.iloc[:,1].values)

#df.iloc[:,1].values
```

```
#create a pair plot
sns.pairplot(df.iloc[:,1:5],hue='diagnosis')
```

```
<seaborn.axisgrid.PairGrid at 0x1f0a61d8750>
```



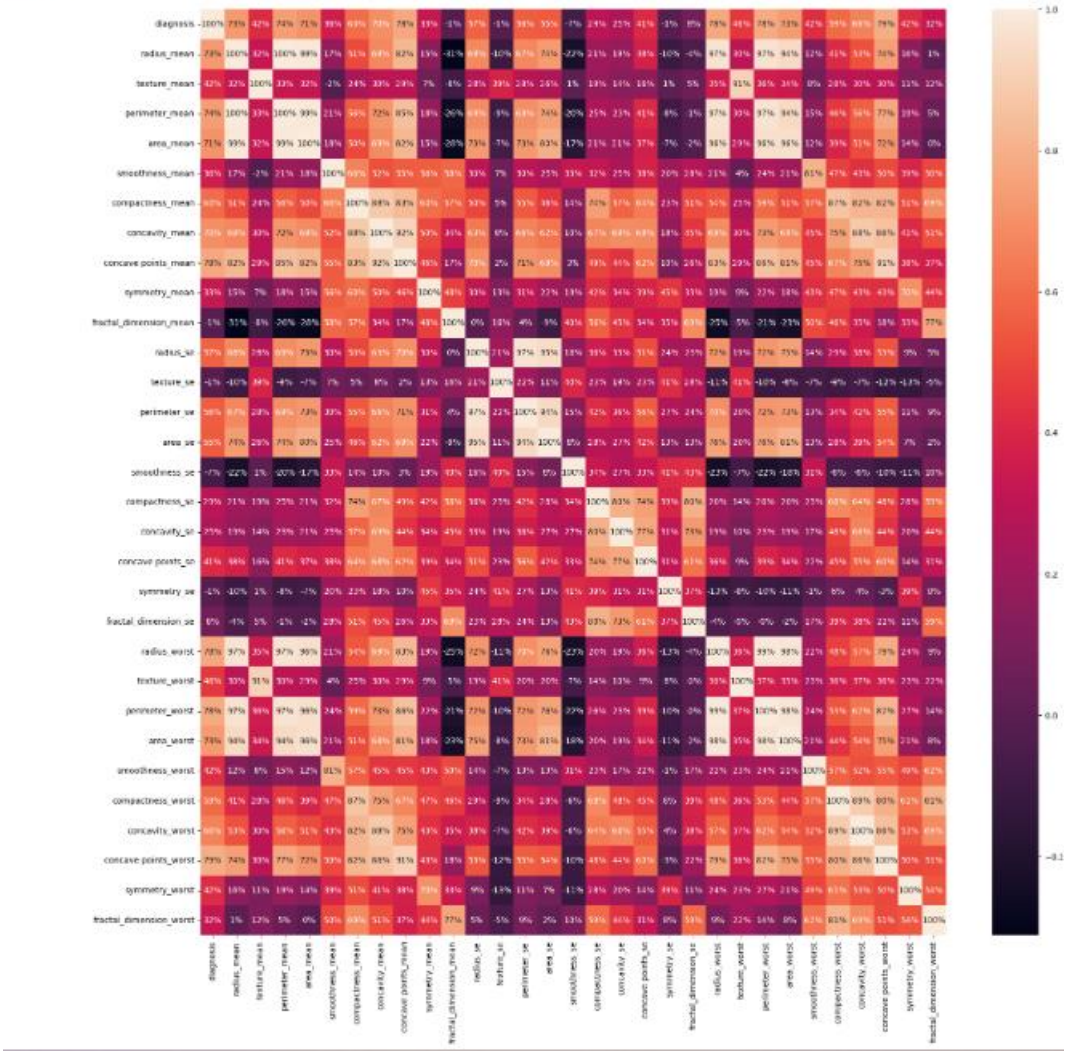
5.3.1 HeatMap İle Korelasyonu Gösterme

Bu kod, **df** DataFrame'inin 1. ile 33. sütunları arasındaki değişkenler arasındaki korelasyonu hesaplar ve bunu bir heatmap üzerinde görselleştirir. Heatmap, değişkenler arasındaki korelasyonu renkli bir matris olarak gösterir. Renk tonu, değişkenler arasındaki ilişkinin gücünü ve yönünü gösterir. **annot=True** parametresi, her hücredeki korelasyon değerlerini gösterir. **fmt='.0%'** ise, korelasyon değerlerinin yüzde formatında gösterilmesini sağlar.

Bu heatmap, değişkenler arasındaki korelasyonu hızlı bir şekilde görselleştirerek, hangi değişkenlerin birlikte hareket ettiğini veya birbirlerine nasıl bağlı olduğunu anlamamıza yardımcı olur. Bu, veri kümesinin yapısal özelliklerini analiz etmek için kullanışlı bir araçtır.

```
In [59]: #visualize the correlation
         plt.figure(figsize=(20,20))
         sns.heatmap(df.iloc[:,1:33].corr(),annot=True, fmt='.0%')
```

```
Out[59]: <Axes: >
```



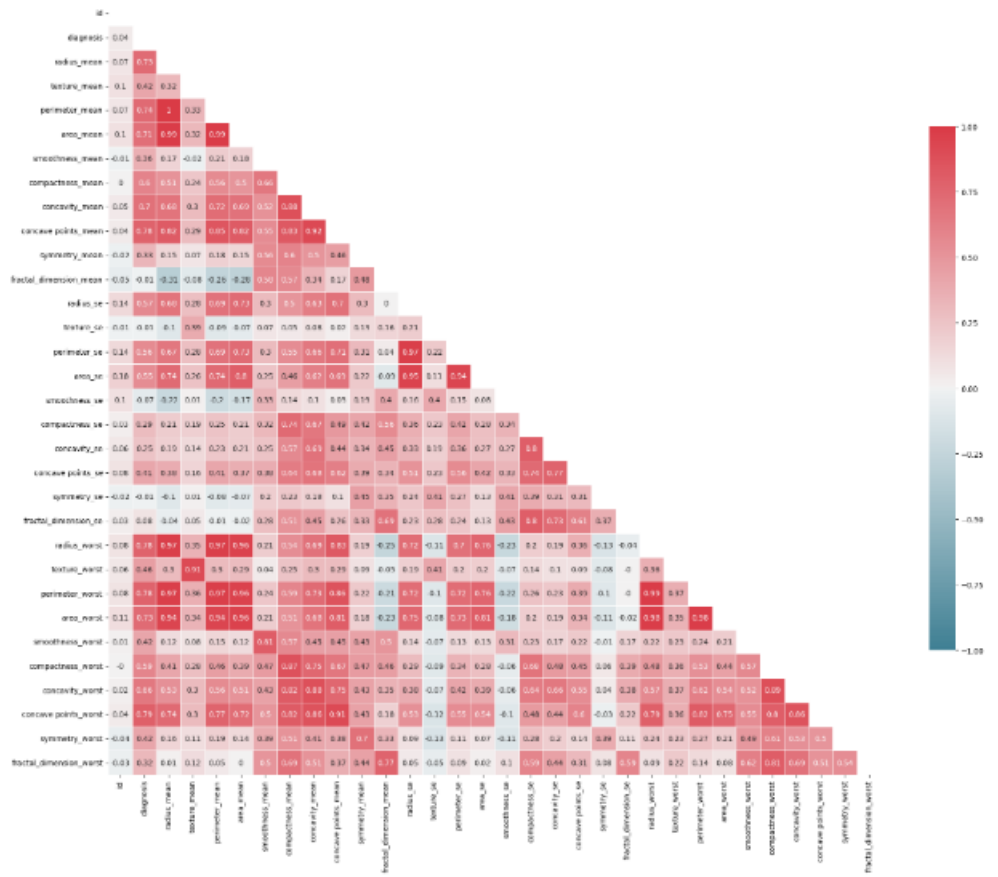
5.3.2 HeatMap İle Korelasyonu Gösterme

Bu kod, **df** DataFrame'inin değişkenler arasındaki korelasyon matrisini hesaplar ve bir heatmap (ısı haritası) şeklinde görselleştirir. Korelasyon matrisi, değişkenler arasındaki ilişkiyi gösterir. 'round(2)' fonksiyonu, korelasyon değerlerini virgülden sonra 2 basamağa yuvarlar.

Kod parçası, önce bir üst üçgen maskesi oluşturur ve ardından bir figura ve eksen belirler. Özel bir renk paleti tanımlanır ve korelasyon matrisi, maskelenmiş üst üçgenle birlikte heatmap olarak çizilir. 'vmin', 'vmax' ve 'center' parametreleri, renk skalasının değerlerini ayarlar. 'square=True' parametresi, kare biçiminde bir heatmap oluşturur. 'annot=True' ise, her hücredeki korelasyon değerlerini gösterir.

Bu heatmap, değişkenler arasındaki korelasyonu görselleştirerek, hangi değişkenlerin birlikte hareket ettiğini ve birbirlerine nasıl bağlı olduklarını anlamamıza yardımcı olur. Bu, veri kümesindeki yapısal özellikleri ve değişkenlerin birbirleriyle olan ilişkilerini analiz etmek için kullanışlı bir araçtır.

```
#generate and visualize the correlation matrix
corr = df.corr().round(2)

#Mask for the upper triangle
mask = np.zeros_like(corr, dtype=bool)
mask[np.triu_indices_from(mask)]= True

#set figure size
f, ax = plt.subplots(figsize=(20,20))

#define custom colormap
cmap = sns.diverging_palette(220, 10, as_cmap=True)

#draw the heatmap
sns.heatmap(corr, mask=mask, cmap=cmap, vmin=-1, vmax=1, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5},annot=True)
plt.tight_layout()
```



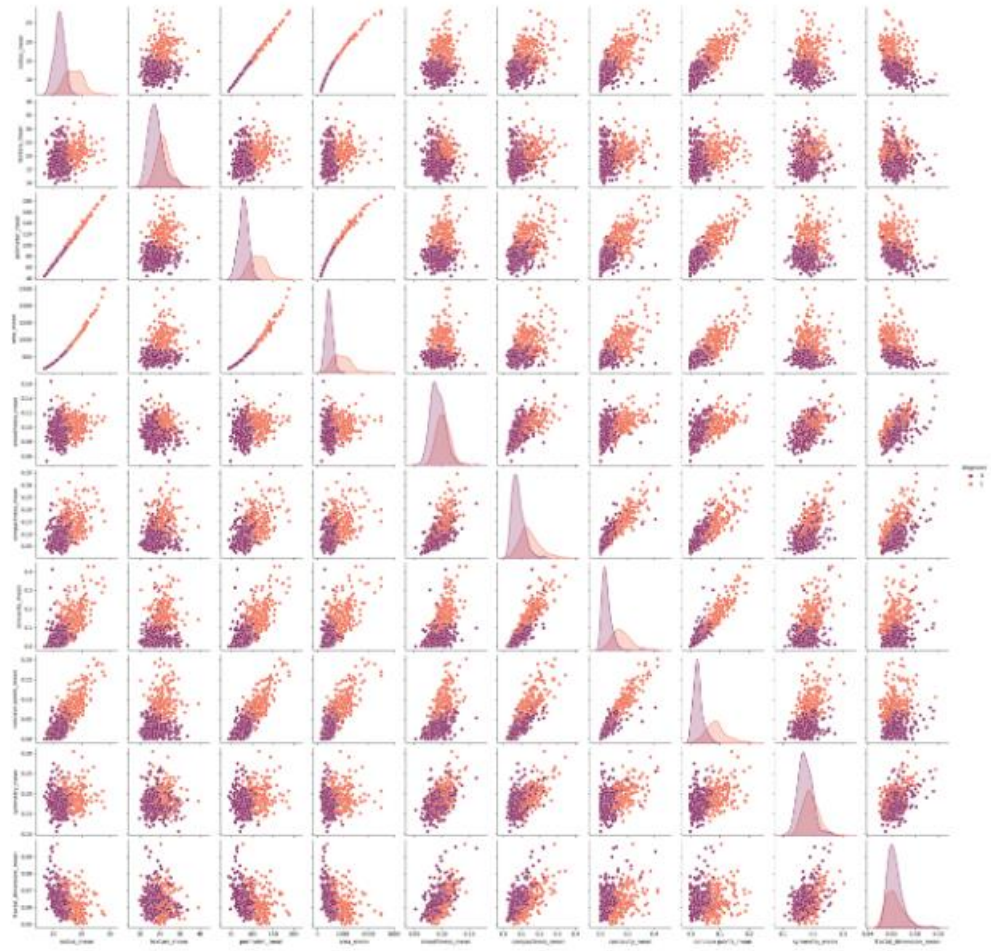
5.4 Scatter Plot Matrisi Oluşturma

Bu kod, **df** DataFrame'inin 'mean' başlıklı sütunlarını temel alarak bir scatter plot matrisi oluşturur. Bu matrisde, her bir 'mean' sütunu bir eksen olarak kullanılır ve sütunlar arasındaki ilişkiler dağılım noktaları (scatter plot) şeklinde gösterilir. **hue='diagnosis'** parametresi, 'diagnosis' sütunundaki kategorilere göre renklendirme yaparak farklı teşhis gruplarını ayırt etmemizi sağlar. 'palette='rocket'' parametresi ise, renk paletini 'rocket' olarak belirler.

Bu scatter plot matrisi, 'mean' sütunları arasındaki ilişkileri görselleştirerek, değişkenlerin birbirleriyle nasıl ilişkili olduğunu ve teşhis grupları arasında nasıl bir ayrım olduğunu anlamamıza yardımcı olur. Bu, veri kümesindeki 'mean' özelliklerinin dağılımını ve korelasyonunu analiz etmek için kullanışlı bir araçtır.

```
# generate a scatter plot matrix with the 'mean' columns
cols = ['diagnosis',
        'radius_mean',
        'texture_mean',
        'perimeter_mean',
        'area_mean',
        'smoothness_mean',
        'compactness_mean',
        'concavity_mean',
        'concave points_mean',
        'symmetry_mean',
        'fractal_dimension_mean']

sns.pairplot(data=df[cols], hue='diagnosis', palette='rocket')
```

```
<seaborn.axisgrid.PairGrid at 0x1f0bdd8f410>
```



5.5 Karmaşıklık Matrisi Oluşturma

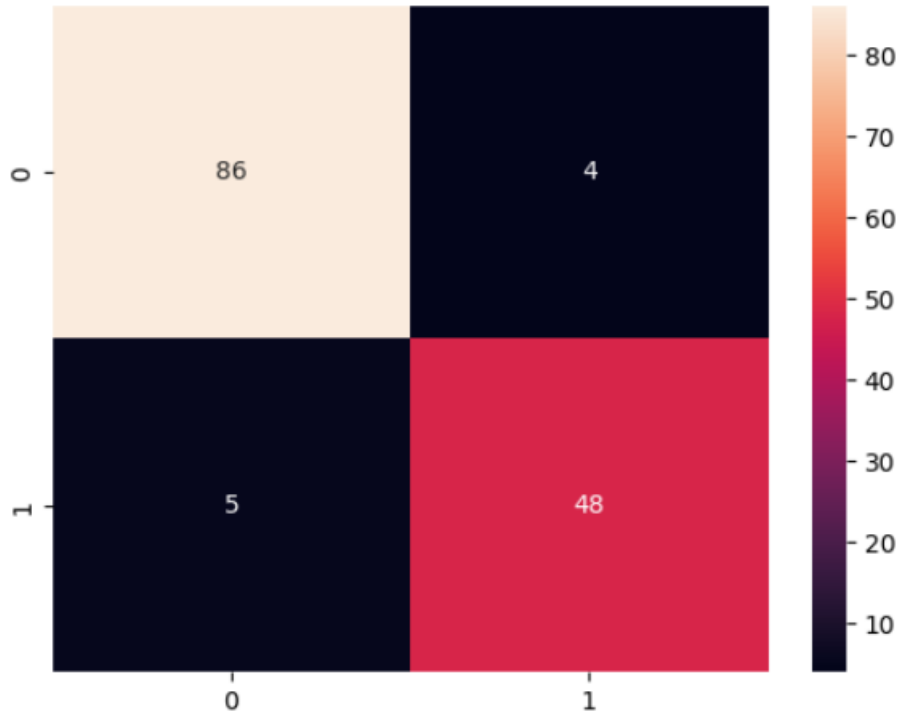
Bu kod, bir karmaşıklık matrisini görselleştirir ve bir heatmap (ısı haritası) olarak kaydeder. **cm** adlı karmaşıklık matrisi, sınıflandırma modelinin tahminleriyle gerçek etiketler arasındaki uyumu gösterir. **annot=True** parametresi, her hücredeki değerleri gösterir. **plt.savefig('h.png')** ise, oluşturulan heatmap'in 'h.png' adında bir dosyaya kaydedilmesini sağlar.

Karmaşıklık matrisi, sınıflandırma modelinin doğruluğunu, hassasiyetini, özgünlüğünü ve duyarlılığını değerlendirmek için kullanılır. Bu görselleştirme, yanlış sınıflandırma oranlarını, doğru sınıflandırma oranlarını ve sınıflandırma hatalarını gösterir. Karmaşıklık matrisi, modelin performansını değerlendirmek ve sınıflandırma sonuçlarını daha iyi anlamak için önemli bir araçtır.

```
from sklearn.metrics import confusion_matrix

cm=confusion_matrix(Y_test,Y_pred)
cm
```

```
array([[86,  4],
       [ 5, 48]], dtype=int64)
```

```
sns.heatmap(cm,annot=True)
plt.savefig('h.png')
```



5.6.1 Sınıflandırma Modelleri Performans Analizi

Bu kod parçası, farklı sınıflandırma modellerinin performansını değerlendirmek için çeşitli metrikleri kullanır.

Kodun ilk bölümü, karmaşıklık matrisinden elde edilen TP, TN, FN ve FP değerlerini kullanarak test doğruluğunu hesaplar ve yazdırır. Ardından, **accuracy_score** fonksiyonuyla gerçek ve tahmin edilen sınıf etiketleri arasındaki doğruluk skorunu hesaplar ve yazdırır.

Sonraki bölümde, K-NN, Gaussian Naive Bayes ve SVM gibi farklı sınıflandırma modelleri oluşturulur. Her bir model için 10 katlamalı çapraz doğrulama ile doğruluk skoru değerlendirmesi yapılır ve sonuçlar yazdırılır.

Ardından, SVM modeli eğitilir ve **X_test** verileri üzerinde tahminler yapılır. Bu tahminlerle doğruluk skoru hesaplanır ve yazdırılır.

Son olarak, sınıflandırma raporu, karmaşıklık matrisi ve doğruluk skoru ekrana yazdırılır.

Bu kod parçası, farklı sınıflandırma modellerinin performansını ölçmek ve karşılaştırmak için kullanılır.

```
TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing Accuracy:', (TP+TN)/(TP+TN+FP+FP))
```

```
Testing Accuracy: 0.9436619718309859
```

```
from sklearn.metrics import accuracy_score
```

```
print('Accuracy Score:',accuracy_score(Y_test,Y_pred))
```

```
Accuracy Score: 0.9370629370629371
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
```

```
models=[]

models.append(('KNN',KNeighborsClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM',SVC()))
```

```
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
```

```
# evaluate each model

results =[]
names=[]
for name , model in models:
    kfold = KFold(n_splits=10, shuffle=True, random_state=40)
    cv_results= cross_val_score(model, X_train, Y_train, cv=kfold, scoring ='accuracy')
    results.append(cv_results)
    names.append(name)

    msg= '%s:, %f, (%f)' % (name,cv_results.mean(),cv_results.std())
    print(msg)
```

```
KNN:, 0.925028, (0.055680)
NB:, 0.943743, (0.018531)
SVM:, 0.901606, (0.053388)
```

```
# make predictions on validation datasets

SVM =SVC()
SVM.fit(X_train, Y_train)
predictions=SVM.predict(X_test)
print('Accuracy Score:\n',accuracy_score(Y_test,predictions))
print('Classification Report:\n',classification_report(Y_test,Y_pred))
print('Confusion Matrix:\n',confusion_matrix(Y_test,predictions))
```

```
Accuracy Score:
 0.9370629370629371
Classification Report:
              precision    recall  f1-score   support

           B       0.95      0.96      0.95        90
           M       0.92      0.91      0.91        53

    accuracy                           0.94       143
   macro avg       0.93      0.93      0.93       143
weighted avg       0.94      0.94      0.94       143

Confusion Matrix:
 [[89  1]
 [ 8 45]]
```

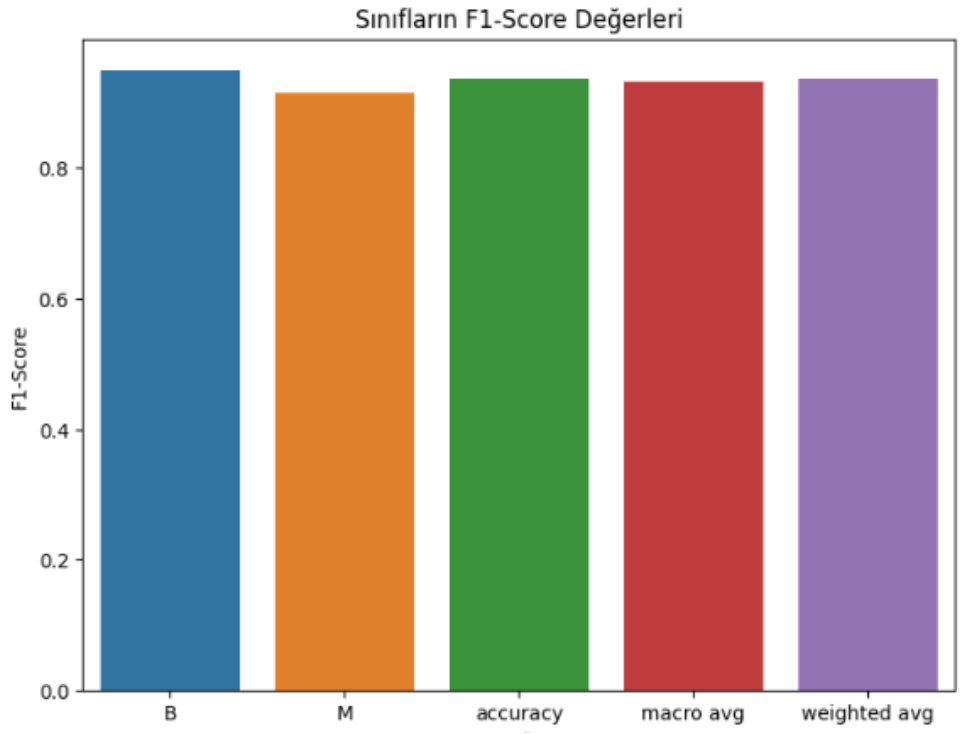
5.6.2 F1-Score Grafiği Oluşturma

Bu kod, **classification_report** fonksiyonunu kullanarak **Y_test** ve **Y_pred** arasında bir sınıflandırma raporu elde eder. Raporu bir veri çerçevesine dönüştürerek **metrics_df** adında bir değişkende saklar. Daha sonra, F1-Score değerlerini göstermek için bir çubuk grafik oluşturur.

Çubuk grafik, her bir sınıfın F1-Score değerlerini temsil eder. X ekseni sınıfları, y ekseni ise F1-Score değerlerini gösterir. Grafik üzerindeki etiketler ve başlık, grafikle ilgili bilgileri açıklar.

Bu görselleştirme, sınıfların F1-Score değerlerini karşılaştırmamıza ve modelin sınıflandırma performansını değerlendirmemize yardımcı olur. F1-Score, bir sınıflandırma modelinin doğruluğunu ve dengesini gösteren önemli bir metriktir.

```
import seaborn as sns
from sklearn.metrics import classification_report

# Sınıflandırma raporunu elde edin
report = classification_report(Y_test, Y_pred, output_dict=True)

# Sınıf etiketlerini ve metrikleri içeren bir veri çerçevesi oluşturun
metrics_df = pd.DataFrame(report).transpose()

# Grafik oluşturma
plt.figure(figsize=(8, 6))
sns.barplot(x=metrics_df.index, y=metrics_df['f1-score'])

# Grafik ayarları
plt.xlabel('Sınıflar')
plt.ylabel('F1-Score')
plt.title('Sınıfların F1-Score Değerleri')

# Grafiği göster
plt.show()
```



6. AKADEMİK MAKALELERDEKİ VERİLER:

Diagnosis Değerleri İçin Countplot Grafiği:

3.2. Dataset acquisition

In our study, we use Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database [13]. The features of dataset are computed from a digitized image of a breast cancer sample obtained from fine-needle aspirate (FNA). The characteristics of the cell nuclei present in the image are determined from these features. Breast Cancer Wisconsin Diagnostic has 569 instances (Benign: 357 Malignant: 212), 2 classes (62.74% benign and 37.26% malignant), and 11 integer-valued attributes (-Id -Diagnosis -Radius -Texture -Area -Perimeter -Smoothness -Compactness -Concavity -Concave points -Symmetry -Fractal dimension).

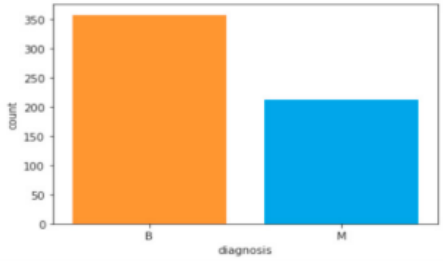


Fig. 2. WISCONSIN BREAST CANCER DIAGNOSTIC DATASETS.

Breast Cancer Wisconsin (Diagnostic) İçin Doğruluk Değeri:

Datasets	Accuracy (%)	Time (Secs)	Tree Size
Breast Cancer	69.23	0.23	5
Breast Cancer Wisconsin (Original)	94.84	0.44	15
Breast Cancer Wisconsin (Diagnostic)	92.97	0.73	17

Sınıflandırma Modelleri Performans Analizi:

Table 1. Performance of the classifiers

Evaluation criteria	Classifiers			
	C4.5	SVM	NB	k-NN
Time to build a model (s)	0.06	0.07	0.05	0.01
Correctly classified instances	665	678	671	666
Incorrectly classified instances	34	21	28	33
Accuracy (%)	95.13	97.13	95.99	95.27

Table 4. Confusion matrix.

	Benign	Malignant	class
C4.5	438	20	Benign
	14	227	Malignant
SVM	446	12	Benign
	9	232	Malignant
NB	436	22	Benign
	6	235	Malignant
k-NN	445	13	Benign
	20	221	Malignant

Table 3. Comparison of accuracy measures for C4.5, SVM, NB and k-NN.

	TP	FP	Precision	Recall	F-Measure	Class
C4.5	0.95	0.05	0.96	0.95	0.96	Benign
	0.94	0.04	0.91	0.94	0.93	Malignant
SVM	0.97	0.03	0.98	0.97	0.97	Benign
	0.96	0.02	0.95	0.96	0.95	Malignant
NB	0.95	0.02	0.98	0.95	0.96	Benign
	0.97	0.04	0.91	0.97	0.94	Malignant
k-NN	0.97	0.08	0.95	0.97	0.96	Benign
	0.91	0.02	0.94	0.91	0.93	Malignant

7. Tartışma ve Sonuç

7.1 Görselleştirme ve Değerlendirme

Bu proje, "Breast Cancer Wisconsin Diagnostic" veri kümesi üzerinde Naive Bayes sınıflandırma yöntemini kullanarak meme kanseri teşhisi yapmayı amaçlamıştır. Proje kapsamında yapılan görselleştirme ve değerlendirme adımları aşağıda tartışılmıştır:

- Pair Plot Grafiği: Veri kümesindeki özniteliklerin birbirleriyle ilişkisini görselleştirmek için Pair Plot grafiği oluşturulmuştur. Bu grafik, öznitelikler arasındaki ilişkiyi ve veri kümesinin genel dağılımını analiz etmemize yardımcı olmuştur.
- Countplot Grafiği: Sınıf etiketlerinin dağılımını görmek için Countplot grafiği oluşturulmuştur. Bu grafik, kanserli ve kansersiz vakaların sınıf dağılımını göstererek veri kümesindeki sınıf dengesizliğini analiz etmemize yardımcı olmuştur.
- HeatMap: Öznitelikler arasındaki korelasyonu görselleştirmek için HeatMap kullanılmıştır. Bu grafik, özniteliklerin birbirleriyle olan ilişkisini ve veri kümesindeki korelasyonu göstererek hangi özniteliklerin daha önemli olduğunu belirlememize yardımcı olmuştur.
- Scatter Plot Matrisi: Veri kümesindeki özniteliklerin birbirleriyle olan ilişkisini daha ayrıntılı olarak görselleştirmek için Scatter Plot Matrisi oluşturulmuştur. Bu grafik, özniteliklerin dağılımını ve ilişkilerini daha detaylı şekilde incelememize olanak sağlamıştır.
- Karmaşıklık Matrisi: Naive Bayes sınıflandırma yöntemi ile elde edilen sonuçları değerlendirmek için Karmaşıklık Matrisi oluşturulmuştur. Bu matris, gerçek pozitifler, yanlış pozitifler, gerçek negatifler ve yanlış negatifler arasındaki ilişkiyi görsel olarak göstererek modelin sınıflandırma performansını değerlendirmemize yardımcı olmuştur.
- F1-Score Grafiği: Modelin farklı kesme noktalarında F1 skorunu görselleştirmek için F1-Score grafiği oluşturulmuştur. Bu grafik, modelin hassasiyet ve özgüllük arasındaki dengeyi göstererek en iyi performansın hangi kesme noktasında elde edildiğini belirlememize yardımcı olmuştur.

7.2 Sonuç

Bu proje, Naive Bayes sınıflandırma yöntemi kullanılarak meme kanseri teşhisi yapma üzerine odaklanmıştır. Proje sonucunda elde edilen görseller ve değerlendirmeler aşağıda sunulmuştur:

- Görselleştirme araçları, veri kümesinin özniteliklerinin birbirleriyle ilişkisini, sınıf dağılımını, öznitelikler arasındaki korelasyonu ve modelin sınıflandırma performansını görsel olarak analiz etmemizi sağlamıştır.
- Pair Plot grafiği, öznitelikler arasındaki ilişkileri ve veri kümesinin genel dağılımını göstererek veri keşfi sürecine katkı sağlamıştır.
- Countplot grafiği, kanserli ve kansersiz vakaların sınıf dağılımını göstererek veri kümesindeki sınıf dengesizliğini belirlememize yardımcı olmuştur.
- HeatMap, öznitelikler arasındaki korelasyonu göstererek hangi özniteliklerin model için daha önemli olduğunu belirlememize yardımcı olmuştur.
- Scatter Plot Matrisi, öznitelikler arasındaki ilişkileri daha detaylı bir şekilde incelememize olanak sağlamıştır.
- Karmaşıklık Matrisi, modelin sınıflandırma performansını doğru pozitifler, yanlış pozitifler, doğru negatifler ve yanlış negatifler aracılığıyla göstererek modelin etkinliğini değerlendirmemize yardımcı olmuştur.
- F1-Score grafiği, modelin farklı kesme noktalarında F1 skorunu göstererek en iyi performansın hangi kesme noktasında elde edildiğini belirlememize yardımcı olmuştur.

Sonuç olarak, bu proje Naive Bayes sınıflandırma yöntemi kullanılarak meme kanseri teşhisi yapma üzerine gerçekleştirilmiş başarılı bir çalışmadır. Görsel analizler ve performans değerlendirmeleri, modelin etkinliğini ve performansını daha ayrıntılı bir şekilde anlamamızı sağlamıştır. Bu çalışma, meme kanseri teşhisi yapmada veri madenciliği ve makine öğrenimi tekniklerinin potansiyelini vurgulamaktadır.

8. KAYNAKÇA:

- **1.URL:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29**
- **2.URL: https://www.youtube.com/watch?v=NSSOyhJBmWY**
- **3.URL: https://www.youtube.com/watch?v=2ncx2q5GHbQ**
- **4.URL: https://github.com/0205Rahul/Breast-Cancer-prediction-for-Wisconsindiagnosticdataset/blob/main/.ipynb_checkpoints/BreastCancer_ANN-checkpoint.ipynb**
- **5.URL: https://chat.openai.com**
- **6.URL: https://www.youtube.com/watch?v=Q93IWdj5Td4&t=226s**

AKADEMİK MAKALELER:

- **7.URL: http://ijcse.com/docs/INDJCSE11-02-05-167.pdf**
- **8.URL:https://www.sciencedirect.com/science/article/pii/S1877050921014629**
- **9.URL:https://www.sciencedirect.com/science/article/pii/S1877050916302575**