Decision Tree & Random Forest V8

November 21, 2021

Replace All zero features with mean
compute_class_weight
RandomOverSampler

[1]:
```
import numpy as np # Import numpy for data preprocessing
import pandas as pd # Import pandas for data frame read
import matplotlib.pyplot as plt # Import matplotlib for data visualisation
import seaborn as sns # Import seaborn for data visualisation
import plotly.express as px # Import plotly for data visualisation
from sklearn.model_selection import train_test_split # Import train_test_split␣
↪for data split
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree␣
↪Classifier
from sklearn.ensemble import RandomForestClassifier # Import Random Forest␣
↪Classifier
from sklearn.model_selection import train_test_split # Import train_test_split␣
↪function
from sklearn import metrics #Import scikit-learn metrics module for accuracy␣
↪calculation
from sklearn import tree # Import export_graphviz for visualizing Decision Trees

from sklearn.utils.class_weight import compute_class_weight
from imblearn.over_sampling import RandomOverSampler # Up-sample or Down-sample
```

0.1 Data read

[2]:
```
df = pd.read_csv("data/diabetes.csv") # Data read
```

[3]:
```
df.head() # print data
```

[3]:
```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
```

```
0                  0.627   50         1
1                  0.351   31         0
2                  0.672   32         1
3                  0.167   21         0
4                  2.288   33         1
```

[4]:
```
df.isna().sum() # check for null value
```

[4]:
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

[5]:
```
df.describe()
```

[5]:
```
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000
```

[6]:
```
df.corr()
```

[6]:
```
                          Pregnancies   Glucose  BloodPressure  SkinThickness  \
Pregnancies                  1.000000  0.129459       0.141282      -0.081672
Glucose                      0.129459  1.000000       0.152590       0.057328
BloodPressure                0.141282  0.152590       1.000000       0.207371
```

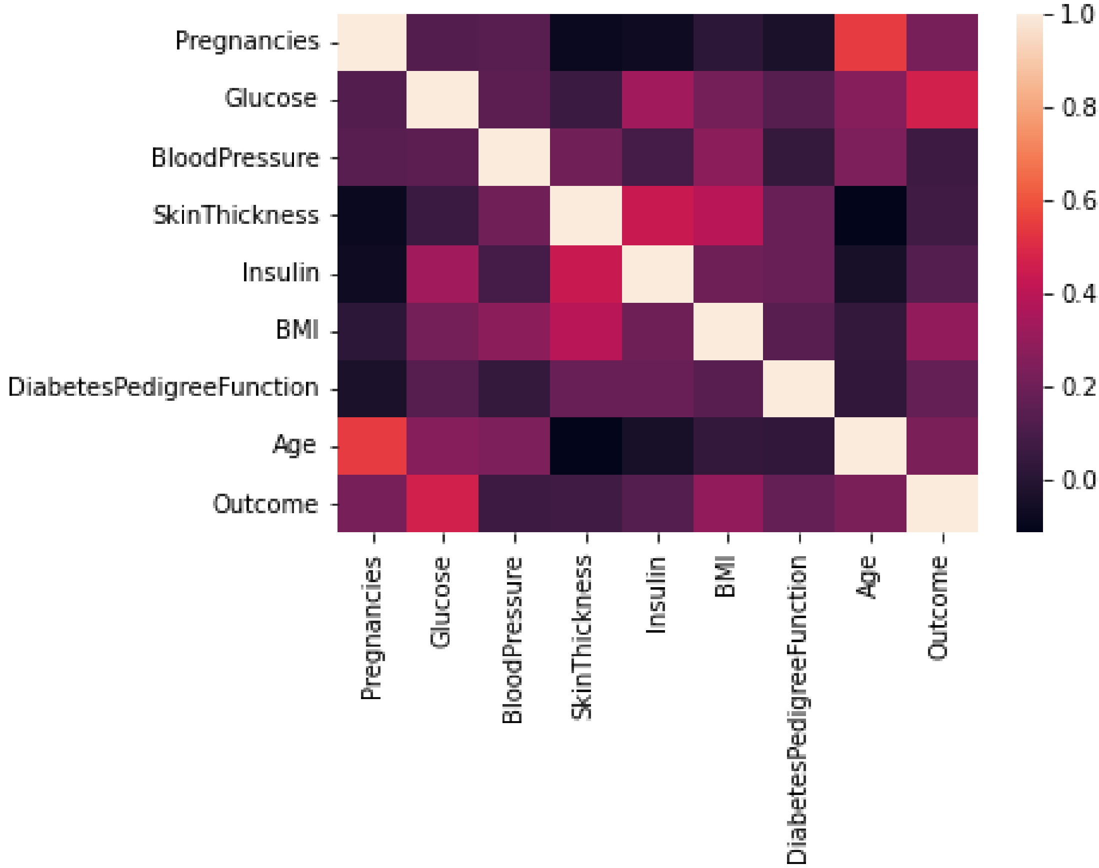
```
SkinThickness             -0.081672  0.057328       0.207371       1.000000
Insulin                   -0.073535  0.331357       0.088933       0.436783
BMI                        0.017683  0.221071       0.281805       0.392573
DiabetesPedigreeFunction  -0.033523  0.137337       0.041265       0.183928
Age                        0.544341  0.263514       0.239528      -0.113970
Outcome                    0.221898  0.466581       0.065068       0.074752

                           Insulin       BMI  DiabetesPedigreeFunction  \
Pregnancies              -0.073535  0.017683                 -0.033523
Glucose                   0.331357  0.221071                  0.137337
BloodPressure             0.088933  0.281805                  0.041265
SkinThickness             0.436783  0.392573                  0.183928
Insulin                   1.000000  0.197859                  0.185071
BMI                       0.197859  1.000000                  0.140647
DiabetesPedigreeFunction  0.185071  0.140647                  1.000000
Age                      -0.042163  0.036242                  0.033561
Outcome                   0.130548  0.292695                  0.173844

                               Age   Outcome
Pregnancies               0.544341  0.221898
Glucose                   0.263514  0.466581
BloodPressure             0.239528  0.065068
SkinThickness            -0.113970  0.074752
Insulin                  -0.042163  0.130548
BMI                       0.036242  0.292695
DiabetesPedigreeFunction  0.033561  0.173844
Age                       1.000000  0.238356
Outcome                   0.238356  1.000000
```

[7]:

```
sns.heatmap(df.corr())
```

[7]:

```
<AxesSubplot:>
```



1 Data split

[8]:
```
X = df.iloc[:,0:-1] # All features
Y = df.iloc[:,-1]  # Target
```

[9]:
```
X.head()
```

[9]:
```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age
0                     0.627   50
1                     0.351   31
2                     0.672   32
```

```
3                       0.167   21
4                       2.288   33
```

[10]:
```
Y.head()
```

[10]:
```
0    1
1    0
2    1
3    0
4    1
Name: Outcome, dtype: int64
```

[11]:
```
print("X.shape : ", X.shape)
print("Y.shape : ", Y.shape)
```

```
X.shape :  (768, 8)
Y.shape :  (768,)
```

[12]:
```
rus = RandomOverSampler(random_state=42)
X_res, Y_res = rus.fit_resample(X, Y)
```

[13]:
```
print("X_res.shape : ", X_res.shape)
print("Y_res.shape : ", Y_res.shape)
```

```
X_res.shape :  (1000, 8)
Y_res.shape :  (1000,)
```

[14]:
```
# Data split
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,␣
 ↪random_state=1)
# x_dev, x_test, y_dev, y_test = train_test_split(x_test, y_test, test_size= 0.
 ↪5)
```

[15]:
```
print("Original data size : ", X.shape, Y.shape)
print("Train data size : ", x_train.shape, y_train.shape)
# print("Dev data size : ", x_dev.shape, y_dev.shape)
print("Test data size : ", x_test.shape, y_test.shape)
```

```
Original data size :  (768, 8) (768,)
Train data size :  (614, 8) (614,)
Test data size :  (154, 8) (154,)
```

2 Preprocessing

[16]:
```
# replace zero bmi value with it's mean
print("Before BMI mean : ",round(x_train.loc[: ,'BMI'].mean(),1))
x_test.loc[: ,'BMI'] = x_test.loc[: ,'BMI'].replace(0, x_train.loc[: ,'BMI'].
 ↪mean())
```

```
x_train.loc[: ,'BMI'] = x_train.loc[: ,'BMI'].replace(0, x_train.loc[: ,'BMI'].
 ↪mean())
print("After BMI mean : ",round(x_train.loc[: ,'BMI'].mean(),1))
```

```
Before BMI mean :  31.8
After BMI mean :  32.2

/Users/kamal/opt/anaconda3/lib/python3.8/site-
packages/pandas/core/indexing.py:1773: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_column(ilocs[0], value, pi)
```

[17]:
```
# replace zero skinthickness value with it's mean
print("Before SkinThickness mean : ",round(x_train.loc[: ,'SkinThickness'].
 ↪mean(),1))
x_test.loc[: ,'SkinThickness'] = x_test.loc[: ,'SkinThickness'].replace(0,␣
 ↪x_train.loc[: ,'SkinThickness'].mean())
x_train.loc[: ,'SkinThickness'] = x_train.loc[: ,'SkinThickness'].replace(0,␣
 ↪x_train.loc[: ,'SkinThickness'].mean())
print("After SkinThickness mean : ",round(x_train.loc[: ,'SkinThickness'].
 ↪mean(),1))
```

```
Before SkinThickness mean :  19.8
After SkinThickness mean :  26.0
```

[18]:
```
# replace zero bloodpressure value with it's mean
print("Before BloodPressure mean : ",round(x_train.loc[: ,'BloodPressure'].
 ↪mean(),1))
x_test.loc[: ,'BloodPressure'] = x_test.loc[: ,'BloodPressure'].replace(0,␣
 ↪x_train.loc[: ,'BloodPressure'].mean())
x_train.loc[: ,'BloodPressure'] = x_train.loc[: ,'BloodPressure'].replace(0,␣
 ↪x_train.loc[: ,'BloodPressure'].mean())
print("After BloodPressure mean : ",round(x_train.loc[: ,'BloodPressure'].
 ↪mean(),1))
```

```
Before BloodPressure mean :  68.9
After BloodPressure mean :  72.1
```

[19]:
```
# replace zero Glucose value with it's mean
print("Before Glucose mean : ",round(x_train.loc[: ,'Glucose'].mean(),1))
x_test.loc[: ,'Glucose'] = x_test.loc[: ,'Glucose'].replace(0, x_train.loc[:␣
 ↪,'Glucose'].mean())
x_train.loc[: ,'Glucose'] = x_train.loc[: ,'Glucose'].replace(0, x_train.loc[:␣
 ↪,'Glucose'].mean())
```

```
print("After Glucose mean : ",round(x_train.loc[: ,'Glucose'].mean(),1))
```

```
Before Glucose mean :  121.3
After Glucose mean :  121.8
```

[20]:
```
# replace zero Insulin value with it's mean
print("Before Insulin mean : ",round(x_train.loc[: ,'Insulin'].mean(),1))
x_test.loc[: ,'Insulin'] = x_test.loc[: ,'Insulin'].replace(0, x_train.loc[:␣
↪,'Insulin'].mean())
x_train.loc[: ,'Insulin'] = x_train.loc[: ,'Insulin'].replace(0, x_train.loc[:␣
↪,'Insulin'].mean())
print("After Insulin mean : ",round(x_train.loc[: ,'Insulin'].mean(),1))
```

```
Before Insulin mean :  79.0
After Insulin mean :  118.4
```

3 Decision Tree

[21]:
```
accuracy = {}
```

3.0.1 criterion="gini", splitter="best"

[22]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="gini", splitter="best",␣
↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[23]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 1 1 0
 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 1 0 0 0 1 0 1 1 1 1 1 0
 0 1 1 0 0 1 1 1 0 1 0 0 0 1 1 0 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0
 0 0 0 1 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 1 0 1 1 0 0 0 1 0 0 1 1 0 1 0 1 0 0
 0 0 0 0 1 1]
```

[24]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[25]:
```
accuracy["dt_gini_best"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.6753246753246753
```

[26]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[73 26]
 [24 31]]
```

[27]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.75      0.74      0.74        99
           1       0.54      0.56      0.55        55

    accuracy                           0.68       154
   macro avg       0.65      0.65      0.65       154
weighted avg       0.68      0.68      0.68       154
```

3.0.2 criterion="gini", splitter="best", max_depth=8

[28]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="gini", splitter="best", max_depth=8,␣
↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[29]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 1 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 0 1 1 0 0 1 1 0 0 1 0 0 1 1 1 1 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0
 0 0 0 1 0 1 0 0 1 1 1 1 1 0 0 0 1 0 0 1 0 1 1 0 0 0 1 0 0 1 0 0 1 0 1 0 0
 1 0 0 1 1 1]
```

[30]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[31]:
```
accuracy["dt_gini_best_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7467532467532467
```

[32]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[74 25]
 [14 41]]
```

[33]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.75      0.79        99
           1       0.62      0.75      0.68        55

    accuracy                           0.75       154
   macro avg       0.73      0.75      0.73       154
weighted avg       0.76      0.75      0.75       154
```

3.0.3 criterion="entropy", splitter="best"

[34]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="best",␣
↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[35]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 1 1 0 0 1 0 1 0 1 0 1 0 1 1 0 0
 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 1 1 0 1 0 0
 0 0 1 1 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
 0 0 0 1 0 1 0 1 1 0 1 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 1 0 0 1 0 0 0 1 1 0 0
 1 0 0 0 1 1]
```

[36]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[37]:
```
accuracy["dt_entropy_best"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7012987012987013
```

[38]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[77 22]
 [24 31]]
```

[39]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support
```

```
              0       0.76      0.78      0.77        99
              1       0.58      0.56      0.57        55

       accuracy                           0.70       154
      macro avg       0.67      0.67      0.67       154
   weighted avg       0.70      0.70      0.70       154
```

3.0.4 criterion="entropy", splitter="best", max_depth=8

[40]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="best", max_depth=8,␣
 ↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[41]:
```
print(y_pred)
```

```
[1 0 0 0 1 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 1 0 1 1 1 0
 1 0 0 0 0 0 1 0 0 1 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 1 0 1 0 0 0 1 1 1 1 0 0
 0 1 1 1 0 1 1 1 0 1 0 1 1 1 0 1 0 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 0 1 0 0 1 0 1 1 0 1 1 1 0 0 0 0 1 1 1 0 1 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0
 1 0 0 0 1 1]
```

[42]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[43]:
```
accuracy["dt_entropy_best_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.6818181818181818
```

[44]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[68 31]
 [18 37]]
```

[45]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.79      0.69      0.74        99
           1       0.54      0.67      0.60        55

    accuracy                           0.68       154
```

```
   macro avg       0.67      0.68      0.67       154
weighted avg       0.70      0.68      0.69       154
```

3.0.5 criterion="entropy", splitter="random"

[46]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="random",␣
↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[47]:
```
print(y_pred)
```

```
[0 1 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 0 0
 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 0 0 0 1 0 0 0 1 1 1 0 0 0
 0 0 1 0 0 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1
 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0
 0 0 0 1 1 0]
```

[48]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[49]:
```
accuracy["dt_entropy_random"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.6493506493506493
```

[50]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[78 21]
 [33 22]]
```

[51]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.70      0.79      0.74        99
           1       0.51      0.40      0.45        55

    accuracy                           0.65       154
   macro avg       0.61      0.59      0.60       154
weighted avg       0.63      0.65      0.64       154
```

3.0.6 criterion="entropy", splitter="random", max_depth=8

[52]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="random",␣
 ↪max_depth=8, class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[53]:
```
print(y_pred)
```

```
[0 0 0 0 1 1 0 0 0 0 1 0 1 0 1 1 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 1 1 0
 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1 1 1 1 0 1 1 1 0 0
 1 1 1 0 0 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1 1 1 0 0 1 0 0 0 1 1 0 0 1 0 1 1 0
 0 1 1 1 0 1 0 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 1 1 0 1 0 0]
```

[54]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[55]:
```
accuracy["dt_entropy_random_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7207792207792207
```

[56]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[74 25]
 [18 37]]
```

[57]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.80      0.75      0.77        99
           1       0.60      0.67      0.63        55

    accuracy                           0.72       154
   macro avg       0.70      0.71      0.70       154
weighted avg       0.73      0.72      0.72       154
```

3.0.7 criterion="entropy", splitter="best", max_depth=3

[58]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="best", max_depth=3,␣
↪class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[59]:
```
print(y_pred)
```

```
[1 0 0 1 0 0 1 0 1 0 1 0 1 1 0 1 1 1 0 0 1 1 1 1 0 1 1 1 1 1 0 1 0 1 1 1 0
 1 1 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 0 1 0 1 0 1 1 1 1 1 1 0 0
 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 0
 0 1 0 1 0 1 0 0 1 1 1 1 1 0 1 1 1 1 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 1 0 0
 1 1 1 1 1 1]
```

[60]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[61]:
```
accuracy["dt_entropy_best_3"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.6753246753246753
```

[62]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[54 45]
 [ 5 50]]
```

[63]:
```
print(metrics.classification_report(y_test, y_pred))
```

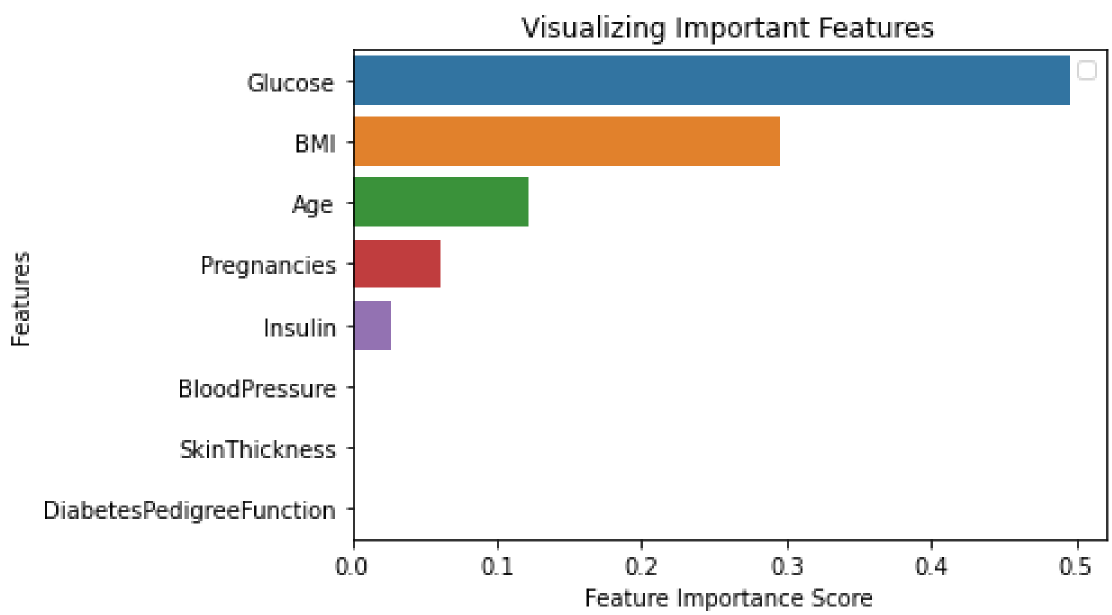
```
              precision    recall  f1-score   support

           0       0.92      0.55      0.68        99
           1       0.53      0.91      0.67        55

    accuracy                           0.68       154
   macro avg       0.72      0.73      0.68       154
weighted avg       0.78      0.68      0.68       154
```

[64]:
```
feature_imp = pd.Series(clf.feature_importances_,index=X.columns).
↪sort_values(ascending=False)
print(feature_imp)
# Creating a bar plot
```

```
sns.barplot(x=feature_imp, y=feature_imp.index)
# Add labels to your graph
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title("Visualizing Important Features")
plt.legend()
plt.show()
```

```
Glucose                     0.495224
BMI                         0.296275
Age                         0.121487
Pregnancies                 0.060543
Insulin                     0.026471
BloodPressure               0.000000
SkinThickness               0.000000
DiabetesPedigreeFunction    0.000000
dtype: float64
```

```
No handles with labels found to put in legend.
```



3.0.8 criterion="entropy", splitter="random", max_depth=3

[65]:
```
# Define and build model
clf = DecisionTreeClassifier(criterion="entropy", splitter="random",␣
↪max_depth=3, class_weight='balanced')
clf = clf.fit(x_train,y_train)
y_pred = clf.predict(x_test)
```

[66]:
```python
print(y_pred)
```

```
[1 1 0 0 1 1 1 0 1 1 1 0 1 1 0 1 1 0 0 0 0 0 1 0 0 1 1 1 0 1 1 0 0 1 0 1 0
 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 1 0 1 0 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0
 1 0 1 1 0 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 0 0 1 1 0 1 1
 0 1 0 1 0 0 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1 1 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 0 0 1 1 1 0]
```

[67]:
```python
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[68]:
```python
accuracy["dt_entropy_random_3"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.6558441558441559
```

[69]:
```python
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[56 43]
 [10 45]]
```

[70]:
```python
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.57      0.68        99
           1       0.51      0.82      0.63        55

    accuracy                           0.66       154
   macro avg       0.68      0.69      0.65       154
weighted avg       0.73      0.66      0.66       154
```

4 Accuracy visulization of Decision Tree

[71]:
```python
accuracy_df_dt = pd.DataFrame(list(zip(accuracy.keys(), accuracy.values())),␣
↪columns =['Arguments', 'Accuracy'])
accuracy_df_dt
```

[71]:

	Arguments	Accuracy
0	dt_gini_best	0.675325
1	dt_gini_best_8	0.746753
2	dt_entropy_best	0.701299
3	dt_entropy_best_8	0.681818

```
4     dt_entropy_random  0.649351
5   dt_entropy_random_8  0.720779
6     dt_entropy_best_3  0.675325
7   dt_entropy_random_3  0.655844
```

[72]:
```
fig = px.bar(accuracy_df_dt, x='Arguments', y='Accuracy')
fig.show()
```

5 Random Forest

[73]:
```
accuracy_rf = {}
```

5.0.1 n_estimators = 1000, criterion='entropy'

[74]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, criterion='entropy',␣
 ↪class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[75]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 1 0]
```

[76]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[77]:
```
accuracy_rf["rf_entropy_1000"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7987012987012987
```

[78]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[87 12]
 [19 36]]
```

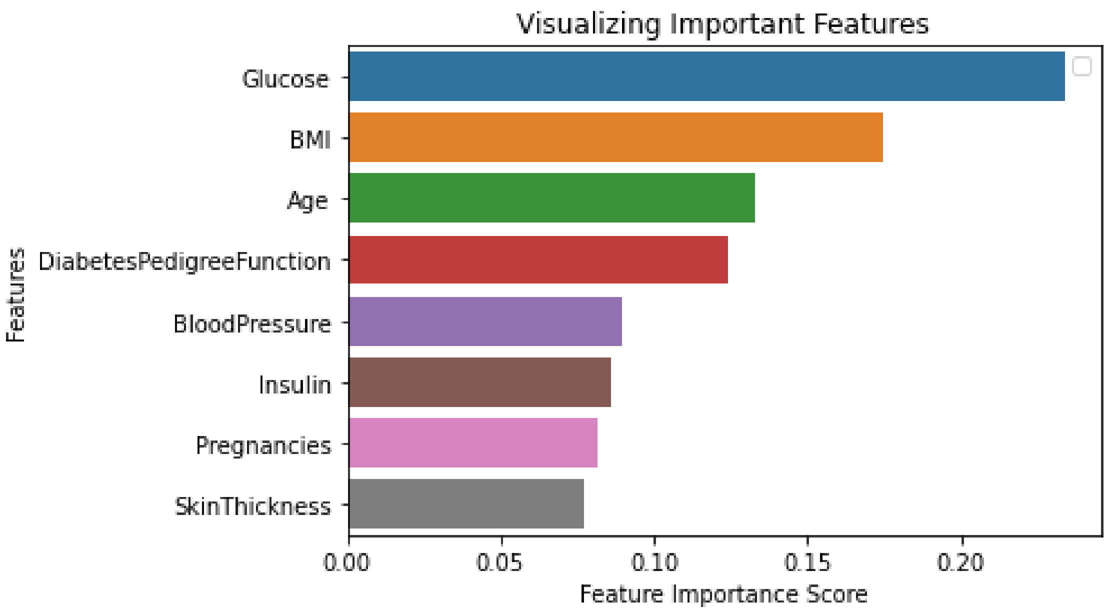
[79]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.82      0.88      0.85        99
           1       0.75      0.65      0.70        55

    accuracy                           0.80       154
   macro avg       0.79      0.77      0.77       154
weighted avg       0.80      0.80      0.80       154
```

[80]:
```
feature_imp = pd.Series(rf.feature_importances_,index=X.columns).
 ↪sort_values(ascending=False)
print(feature_imp)
# Creating a bar plot
sns.barplot(x=feature_imp, y=feature_imp.index)
# Add labels to your graph
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title("Visualizing Important Features")
plt.legend()
plt.show()
```

```
No handles with labels found to put in legend.
```

```
Glucose                     0.233693
BMI                         0.174989
Age                         0.133020
DiabetesPedigreeFunction    0.124056
BloodPressure               0.089496
Insulin                     0.086293
Pregnancies                 0.081532
SkinThickness               0.076921
dtype: float64
```



5.0.2 n_estimators = 100, criterion='entropy'

[81]:
```
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, criterion='entropy',␣
↪class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[82]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0
 1 0 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 1
 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 0 0 0 1 1 0]
```

[83]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[84]:
```
accuracy_rf["rf_entropy_100"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7987012987012987
```

[85]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[86 13]
 [18 37]]
```

[86]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.83      0.87      0.85        99
           1       0.74      0.67      0.70        55

    accuracy                           0.80       154
   macro avg       0.78      0.77      0.78       154
weighted avg       0.80      0.80      0.80       154
```

5.0.3 n_estimators = 1000, random_state = 42, criterion='entropy'

[87]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, random_state = 42,␣
 ↪criterion='entropy', class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[88]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 1 0]
```

[89]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[90]:
```
accuracy_rf["rf_entropy_1000_42"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7987012987012987
```

[91]:
```python
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[87 12]
 [19 36]]
```

[92]:
```python
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.82      0.88      0.85        99
           1       0.75      0.65      0.70        55

    accuracy                           0.80       154
   macro avg       0.79      0.77      0.77       154
weighted avg       0.80      0.80      0.80       154
```

5.0.4 n_estimators = 100, random_state = 42, criterion='entropy'

[93]:
```python
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, random_state = 42, max_depth =␣
↪8, criterion='entropy', class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[94]:
```python
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 1
 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 0 0 0 1 1 0]
```

[95]:
```python
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[96]:
```python
accuracy_rf["rf_entropy_100_42"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8181818181818182
```

[97]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[80 19]
 [ 9 46]]
```

[98]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.81      0.85        99
           1       0.71      0.84      0.77        55

    accuracy                           0.82       154
   macro avg       0.80      0.82      0.81       154
weighted avg       0.83      0.82      0.82       154
```

5.0.5 n_estimators = 1000, random_state = 42, max_depth = 8, criterion='entropy'

[99]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, random_state = 42, max_depth =␣
↪8, criterion='entropy', class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[100]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 0 0 1 1 0 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 1 0 1 0 0 1 0 1 0 1 1 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 1 0 0 1 1 0]
```

[101]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[102]:
```
accuracy_rf["rf_entropy_1000_42_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8246753246753247
```

[103]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[81 18]
 [ 9 46]]
```

[104]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.82      0.86        99
           1       0.72      0.84      0.77        55

    accuracy                           0.82       154
   macro avg       0.81      0.83      0.82       154
weighted avg       0.84      0.82      0.83       154
```

5.0.6 n_estimators = 100, random_state = 42, max_depth = 8, criterion='entropy'

[105]:
```
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, random_state = 42, max_depth =␣
 ↪8, criterion='entropy', class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[106]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 1
 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 0 0 0 1 1 0]
```

[107]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[108]:
```
accuracy_rf["rf_entropy_100_42_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8181818181818182
```

[109]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[80 19]
 [ 9 46]]
```

[110]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.81      0.85        99
           1       0.71      0.84      0.77        55

    accuracy                           0.82       154
   macro avg       0.80      0.82      0.81       154
weighted avg       0.83      0.82      0.82       154
```

5.0.7 n_estimators = 1000

[111]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[112]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 1 0]
```

[113]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[114]:
```
accuracy_rf["rf_gini_1000"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7922077922077922
```

[115]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[86 13]
 [19 36]]
```

[116]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.82      0.87      0.84        99
           1       0.73      0.65      0.69        55

    accuracy                           0.79       154
   macro avg       0.78      0.76      0.77       154
weighted avg       0.79      0.79      0.79       154
```

5.0.8 n_estimators = 100

[117]:
```
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[118]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0
 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 1 0]
```

[119]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[120]:
```
accuracy_rf["rf_gini_100"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8051948051948052
```

[121]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[86 13]
 [17 38]]
```

[122]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.83      0.87      0.85        99
```

```
           1       0.75      0.69      0.72        55

    accuracy                           0.81       154
   macro avg       0.79      0.78      0.78       154
weighted avg       0.80      0.81      0.80       154
```

5.0.9 n_estimators = 1000, random_state = 42

[123]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, random_state = 42,␣
↪class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[124]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 1 1 0 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 0 0 0 1 1 0]
```

[125]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[126]:
```
accuracy_rf["rf_gini_1000_42"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7857142857142857
```

[127]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[86 13]
 [20 35]]
```

[128]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.81      0.87      0.84        99
           1       0.73      0.64      0.68        55
```

```
    accuracy                           0.79       154
   macro avg       0.77      0.75      0.76       154
weighted avg       0.78      0.79      0.78       154
```

5.0.10 n_estimators = 100, random_state = 42

[129]:
```
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, random_state = 42, max_depth =␣
↪8, class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[130]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 1
 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 1 0 0 1 1 0]
```

[131]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[132]:
```
accuracy_rf["rf_gini_100_42"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8051948051948052
```

[133]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[82 17]
 [13 42]]
```

[134]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.83      0.85        99
           1       0.71      0.76      0.74        55

    accuracy                           0.81       154
   macro avg       0.79      0.80      0.79       154
```

```
weighted avg       0.81      0.81      0.81       154
```

5.0.11 n_estimators = 1000, random_state = 42, max_depth = 8

[135]:
```
# Instantiate model with 1000 decision trees
rf = RandomForestClassifier(n_estimators = 1000, random_state = 42, max_depth =
 ↪8, class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[136]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0
 0 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 1 0 0
 0 0 0 1 1 0]
```

[137]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[138]:
```
accuracy_rf["rf_gini_1000_42_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7987012987012987
```

[139]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[82 17]
 [14 41]]
```

[140]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.83      0.84        99
           1       0.71      0.75      0.73        55

    accuracy                           0.80       154
   macro avg       0.78      0.79      0.78       154
weighted avg       0.80      0.80      0.80       154
```

5.0.12 n_estimators = 100, random_state = 42, max_depth = 8

[141]:
```
# Instantiate model with 100 decision trees
rf = RandomForestClassifier(n_estimators = 100, random_state = 42, max_depth =
 ↪8, class_weight='balanced')
# Train the model on training data
rf.fit(x_train,y_train)
# Use the forest's predict method on the test data
y_pred = rf.predict(x_test)
```

[142]:
```
print(y_pred)
```

```
[1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0
 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 1 0
 1 1 1 0 0 1 1 0 0 0 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 1
 0 1 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0
 1 0 0 1 1 0]
```

[143]:
```
print(np.array(y_test))
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 0 1 1 0 0 1 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 0 0 0
 1 1 1 0 0 1 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1
 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 1 1 0 0
 1 0 0 1 0 0]
```

[144]:
```
accuracy_rf["rf_gini_100_42_8"] = metrics.accuracy_score(y_test, y_pred);
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.8051948051948052
```

[145]:
```
print(metrics.confusion_matrix(y_test, y_pred))
```

```
[[82 17]
 [13 42]]
```

[146]:
```
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.86      0.83      0.85        99
           1       0.71      0.76      0.74        55

    accuracy                           0.81       154
   macro avg       0.79      0.80      0.79       154
weighted avg       0.81      0.81      0.81       154
```

6 Accuracy visulization of Random Forest

[147]:
```
accuracy_df_rf = pd.DataFrame(list(zip(accuracy_rf.keys(), accuracy_rf.
 ↪values())), columns =['Arguments', 'Accuracy'])
accuracy_df_rf
```

[147]:

	Arguments	Accuracy
0	rf_entropy_1000	0.798701
1	rf_entropy_100	0.798701
2	rf_entropy_1000_42	0.798701
3	rf_entropy_100_42	0.818182
4	rf_entropy_1000_42_8	0.824675
5	rf_entropy_100_42_8	0.818182
6	rf_gini_1000	0.792208
7	rf_gini_100	0.805195
8	rf_gini_1000_42	0.785714
9	rf_gini_100_42	0.805195
10	rf_gini_1000_42_8	0.798701
11	rf_gini_100_42_8	0.805195

[148]:
```
fig = px.bar(accuracy_df_rf, x='Arguments', y='Accuracy')
fig.show()
```

[149]:
```
accuracy_df = pd.concat([accuracy_df_dt, accuracy_df_rf])
accuracy_df['Accuracy'] = round(accuracy_df['Accuracy'] * 100, 2)
fig = px.bar(accuracy_df, x='Arguments', y='Accuracy')
print(accuracy_df['Accuracy'].max())
fig.show()
```

82.47

[ ]:
```

```