VIETNAM NATIONAL UNIVERSITY HO CHI MINH
CITY THE UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

Project Report

Artificial Intelligence

MEMBER :

Cao Anh Khoa - 22120161

Trần Anh Khoa - 22120164

Nguyễn Thanh Phong - 22120265

Điểu Kham - 19120441

Contents

I. Project evaluation

No.	Criteria	Score	Name	Rate
1	Analysis of the Wine Quality dataset	30%	Điểu Kham	100%
2	Analysis of the Breast Cancer dataset	30%	Trần Anh Khoa	100%
3	Analysis of an additional dataset.	30%	Cao Anh Khoa & Nguyễn Thanh Phong	100%
4	Comparative analysis of all three datasets.	5%	Cao Anh Khoa & Nguyễn Thanh Phong	100%
5	Well-structured and formatted notebooks.	5%	Cao Anh Khoa & Nguyễn Thanh Phong	100%
6	Report		Cao Anh Khoa & Nguyễn Thanh Phong	100%
	Total	100%		

II. Result

1. Binary Class

2.1. Load and analyze the data

```
...           id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0      842302         M        17.99         10.38          122.80     1001.0
1      842517         M        20.57         17.77          132.90     1326.0
2    84300903         M        19.69         21.25          130.00     1203.0
3    84348301         M        11.42         20.38           77.58      386.1
4    84358402         M        20.29         14.34          135.10     1297.0

     smoothness_mean  compactness_mean  concavity_mean  concave_points_mean  \
0            0.11840           0.27760          0.3001              0.14710
1            0.08474           0.07864          0.0869              0.07017
2            0.10960           0.15990          0.1974              0.12790
3            0.14250           0.28390          0.2414              0.10520
4            0.10030           0.13280          0.1980              0.10430

     ...  radius_worst  texture_worst  perimeter_worst  area_worst  \
0    ...         25.38          17.33           184.60      2019.0
1    ...         24.99          23.41           158.80      1956.0
2    ...         23.57          25.53           152.50      1709.0
3    ...         14.91          26.50            98.87       567.7
4    ...         22.54          16.67           152.20      1575.0

     smoothness_worst  compactness_worst  concavity_worst  concave_points_worst
0              0.1622             0.6656           0.7119                0.2654
1              0.1238             0.1866           0.2416                0.1860
2              0.1444             0.4245           0.4504                0.2430
...
 31  fractal_dimension_worst  569 non-null    float64
dtypes: float64(30), int64(1), object(1)
memory usage: 142.4+ KB
None
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Figure 1 - Output analysis the dataset

2.2 .Step to perform.

1.2.1 .Prepare the date

2.2.1.1 .Separate features and label

```
...   0      1
      1      1
      2      1
      3      1
      4      1
            ..
      564    1
      565    1
      566    1
      567    1
      568    0
      Name: diagnosis, Length: 569, dtype: int64
           radius_mean  texture_mean  perimeter_mean  area_mean  smoothness_mean  \
      0          17.99         10.38          122.80     1001.0          0.11840
      1          20.57         17.77          132.90     1326.0          0.08474
      2          19.69         21.25          130.00     1203.0          0.10960
      3          11.42         20.38           77.58      386.1          0.14250
      4          20.29         14.34          135.10     1297.0          0.10030
      ..           ...           ...             ...        ...              ...
      564        21.56         22.39          142.00     1479.0          0.11100
      565        20.13         28.25          131.20     1261.0          0.09780
      566        16.60         28.08          108.30      858.1          0.08455
      567        20.60         29.33          140.10     1265.0          0.11780
      568         7.76         24.54           47.92      181.0          0.05263
      ...
      567                 0.12400
      568                 0.07039

      [569 rows x 30 columns]
      Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Figure 2 - Output Split the dataset into features and labels

2.2.1.2 Analyze the original data.

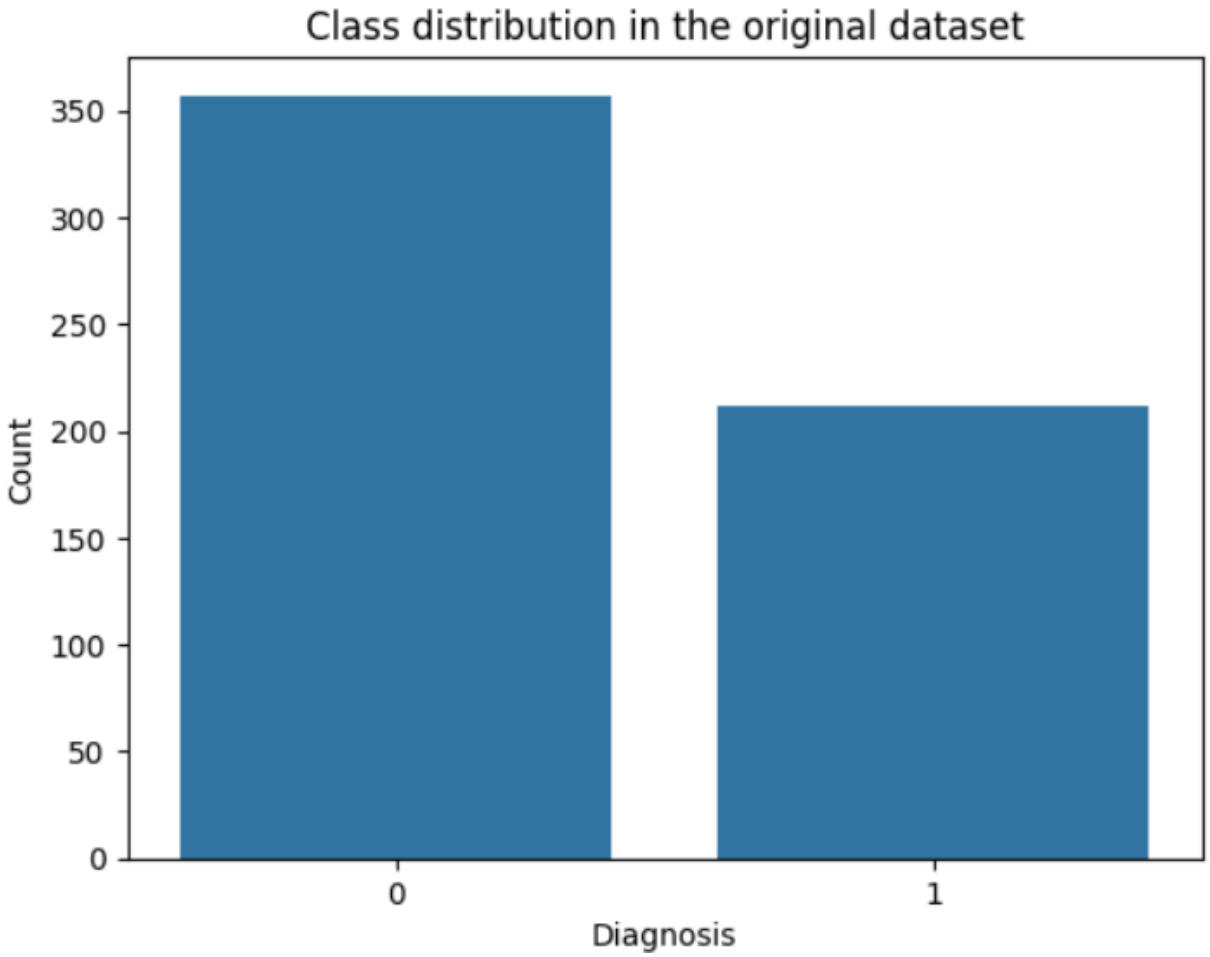


Figure 3-Analysis the original dataset

2.2.1.3 Split data into train/ test with different ratios.

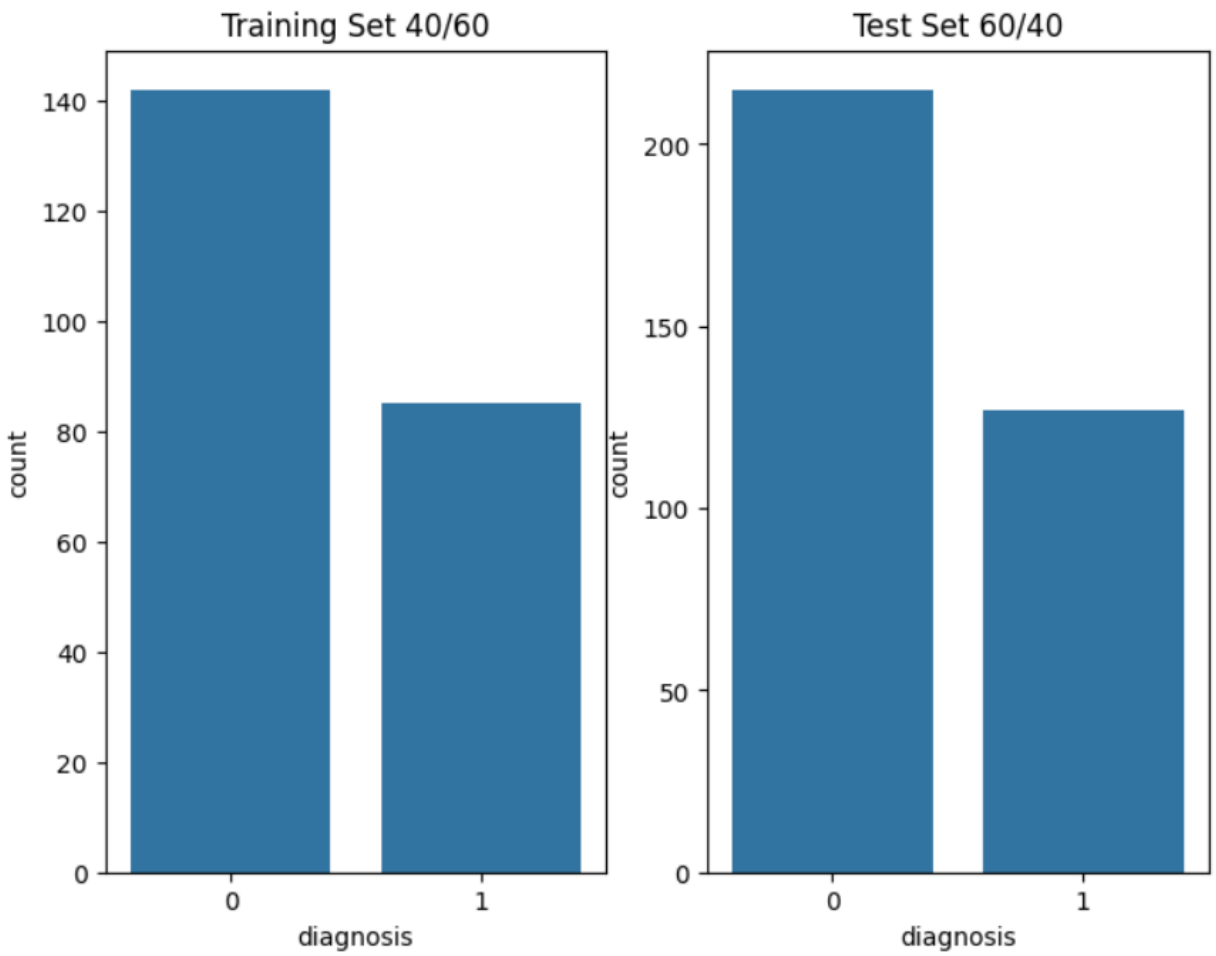


Figure 4 - Visualizing 40/60

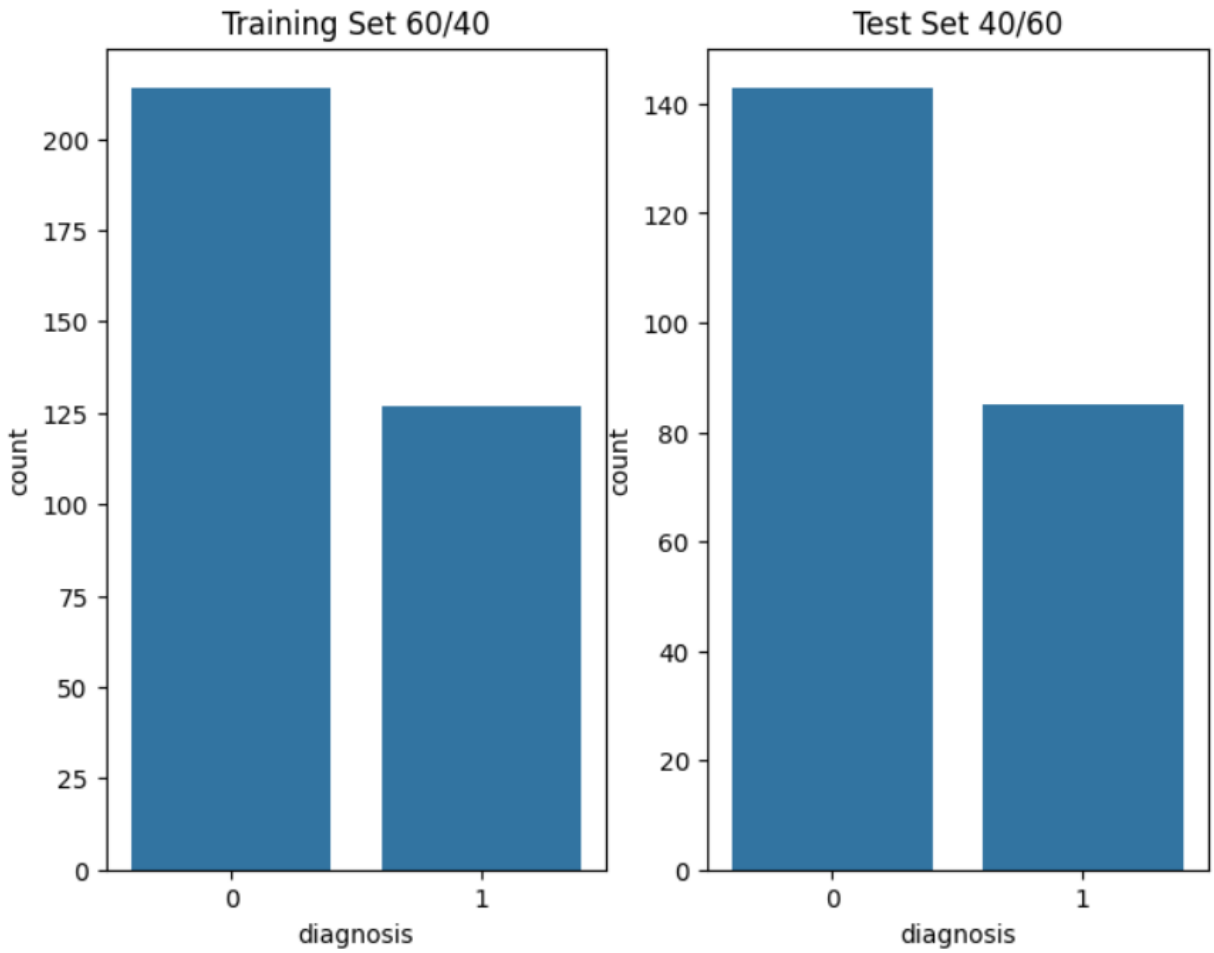


Figure 5 - Visualizing 60/40

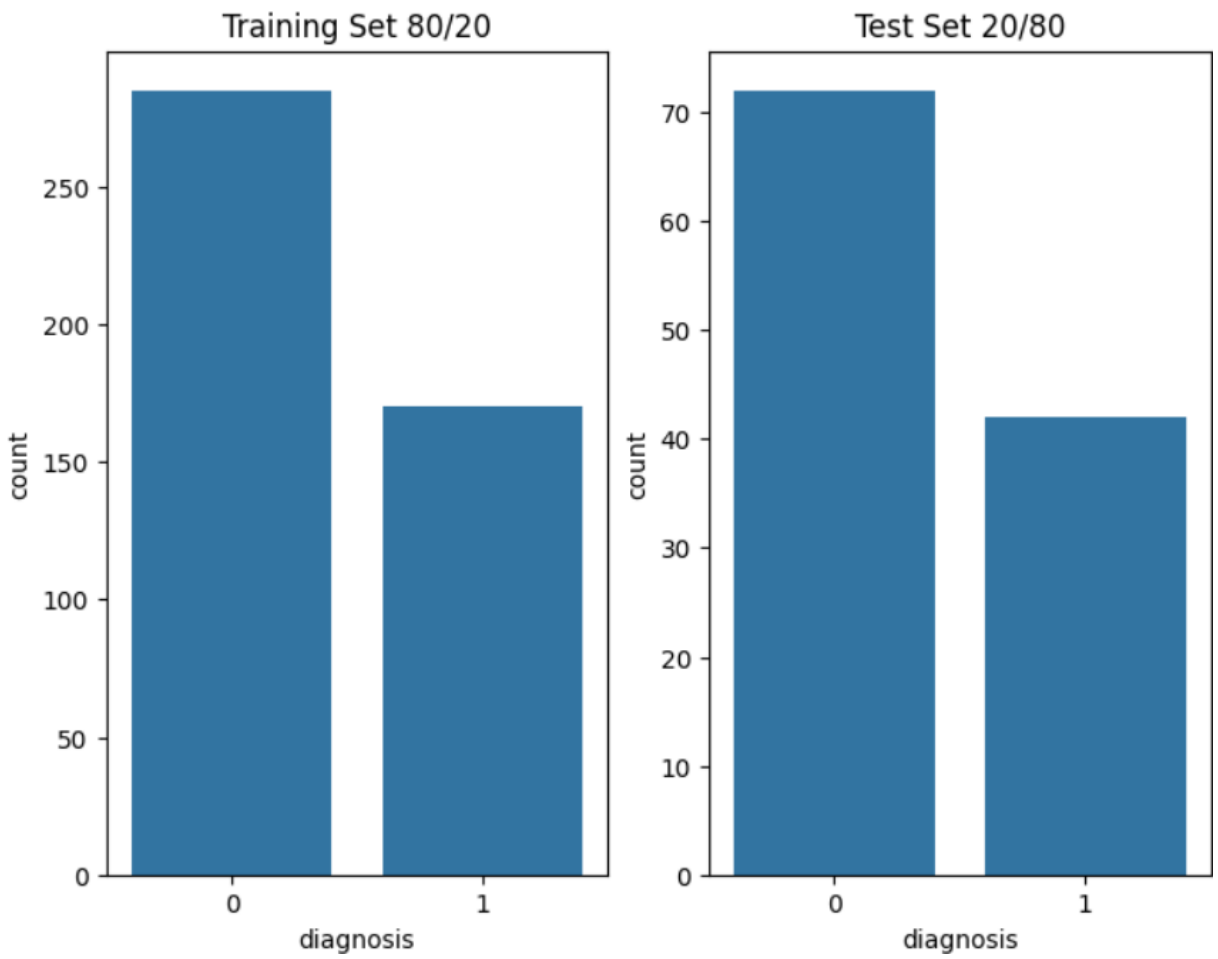


Figure 6 - Visualizing 80/20

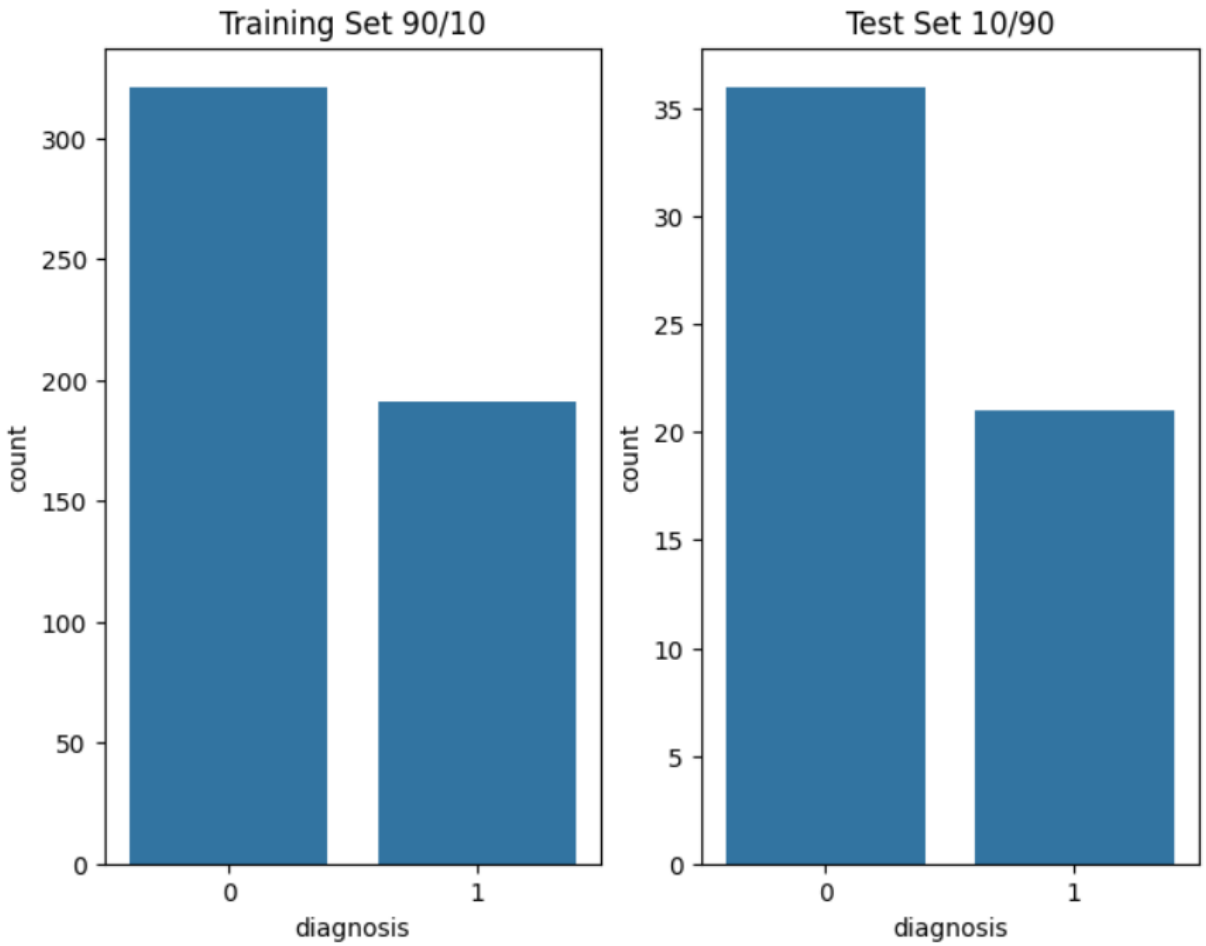


Figure 7 - Visualizing 90/10

2.2.1.4 Build a decision tree classifier and visualize it.

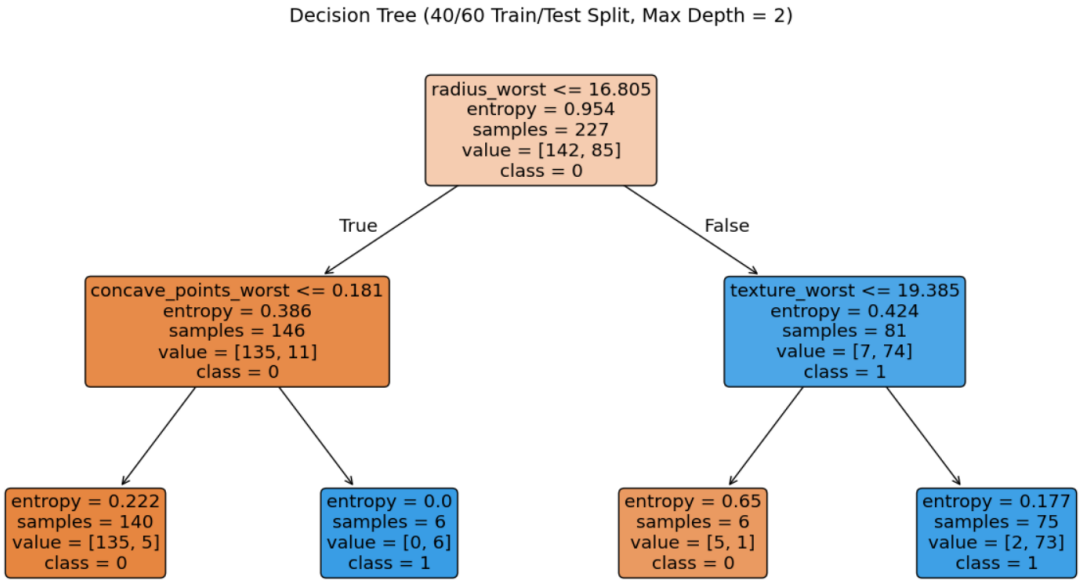


Figure 8 - Decision tree 40/60

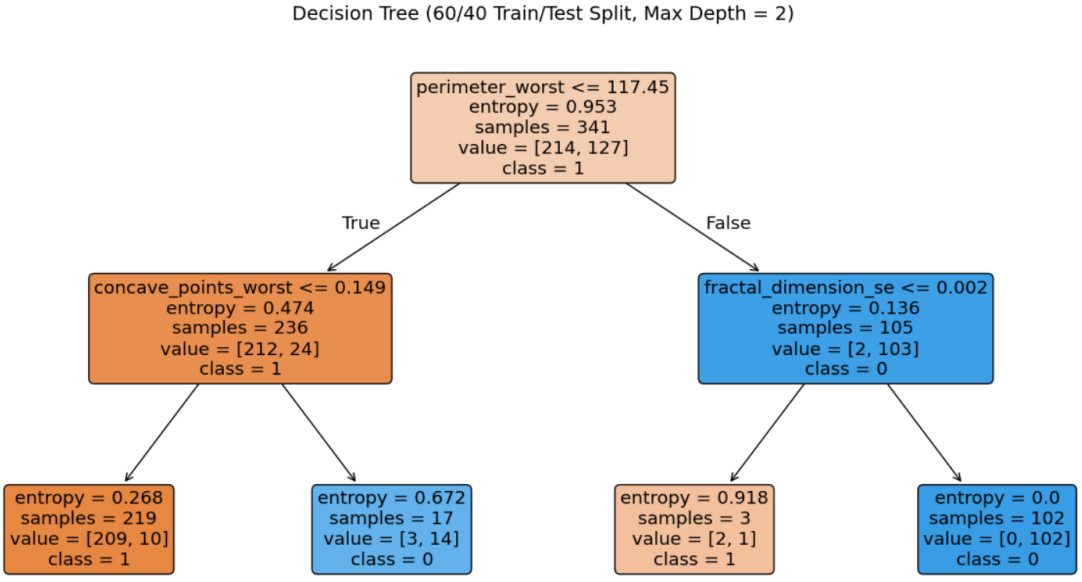


Figure 9 - Decision tree 60/40

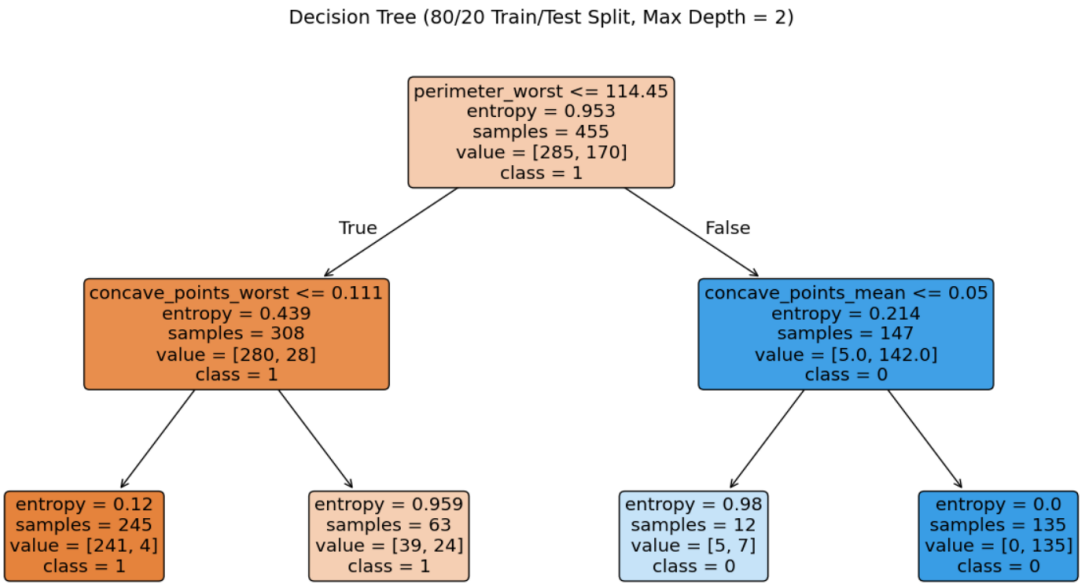


Figure 10 - Decision tree 80/20

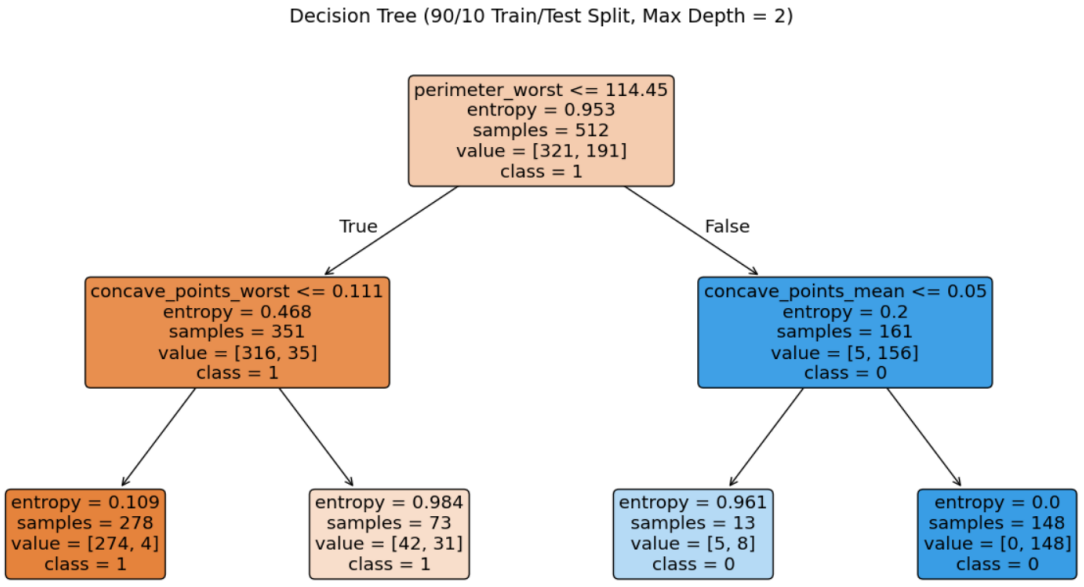


Figure 11 - Decision tree 90/10

2.2.1.5 Evaluate the decision tree classifier

```
Evaluating Decision Tree for split 40/60:
Classification Report:
              precision    recall  f1-score   support

           B       0.91      0.96      0.93       215
           M       0.92      0.83      0.88       127

    accuracy                           0.91       342
   macro avg       0.91      0.90      0.90       342
weighted avg       0.91      0.91      0.91       342

Confusion Matrix:
[[206   9]
 [ 21 106]]
```

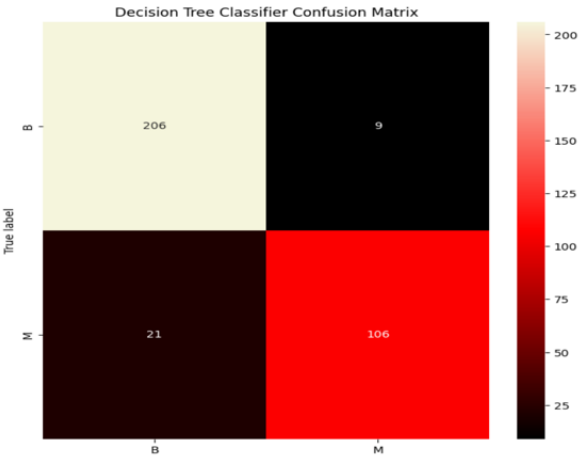


Figure 12 - Evaluating Decision tree classifiers 40/60

```
Evaluating Decision Tree for split 60/40:
Classification Report:
              precision    recall  f1-score   support

           B       0.94      0.96      0.95       143
           M       0.93      0.91      0.92        85

    accuracy                           0.94       228
   macro avg       0.94      0.93      0.93       228
weighted avg       0.94      0.94      0.94       228

Confusion Matrix:
[[137   6]
 [  8  77]]
```

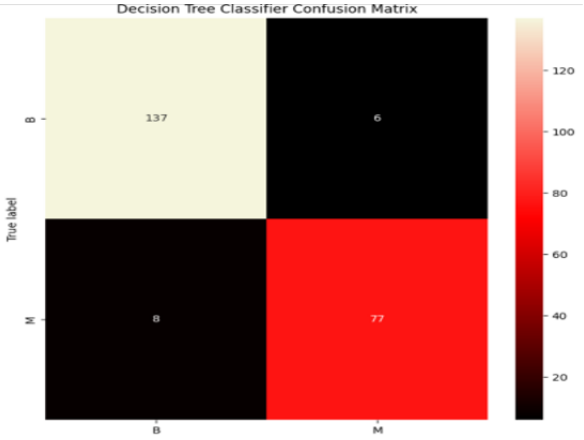


Figure 13 - Evaluating Decision tree classifiers 60/40

```
Evaluating Decision Tree for split 80/20:
Classification Report:
              precision    recall  f1-score   support

           B       0.95      0.99      0.97        72
           M       0.97      0.90      0.94        42

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114

Confusion Matrix:
[[71  1]
 [ 4 38]]
```

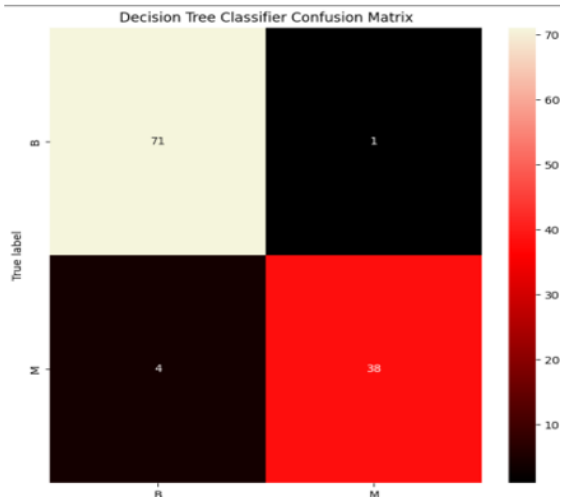


Figure 14 - Evaluating Decision tree classifiers 80/20

```
Evaluating Decision Tree for split 90/10:
Classification Report:
              precision    recall  f1-score   support

           B       0.95      0.97      0.96        36
           M       0.95      0.90      0.93        21

    accuracy                           0.95        57
   macro avg       0.95      0.94      0.94        57
weighted avg       0.95      0.95      0.95        57

Confusion Matrix:
[[35  1]
 [ 2 19]]
```

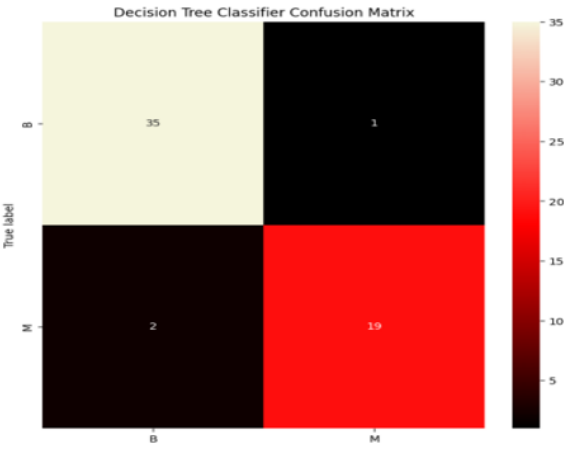


Figure 15 - Evaluating Decision tree classifiers 90/10

2.2.1.6 The depth and accuracy of a decision tree

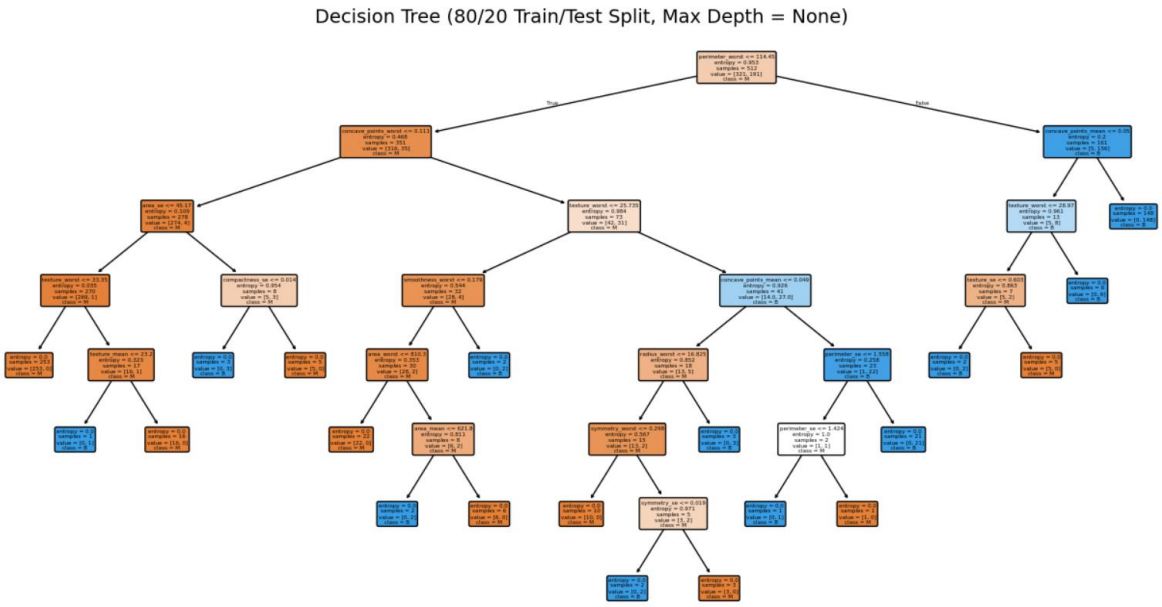


Figure 16 - The depth and accuracy of a decision tree (max depth = None)

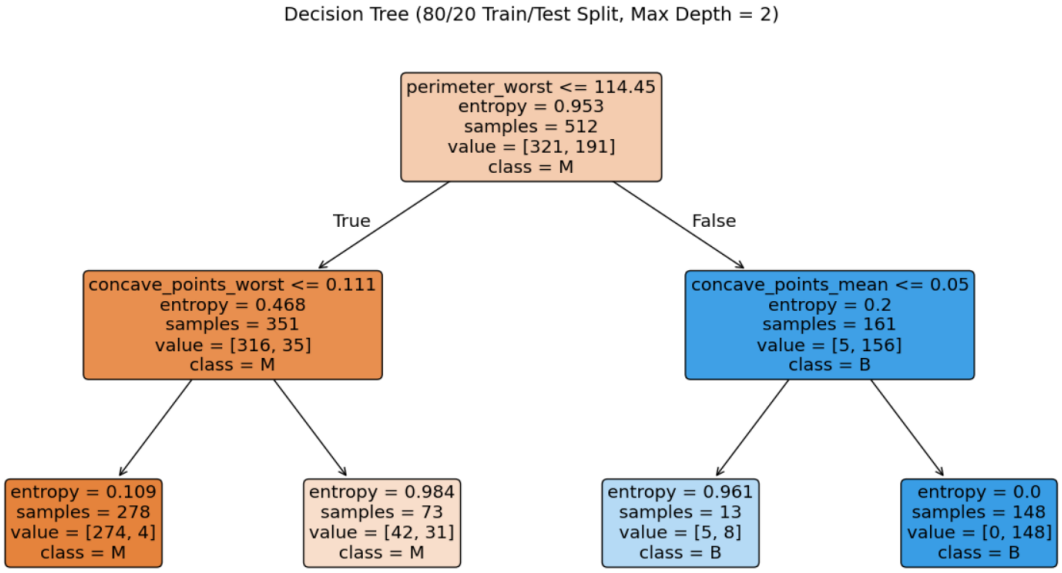


Figure 17 - The depth and accuracy of a decision tree (max depth = 2)

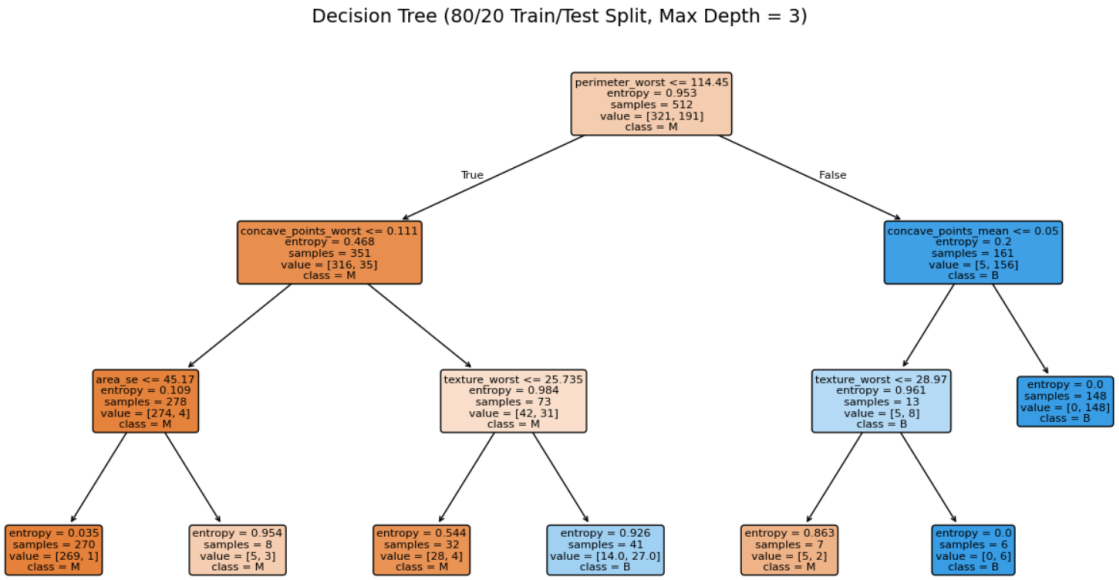


Figure 18 - The depth and accuracy of a decision tree (max depth = 3)

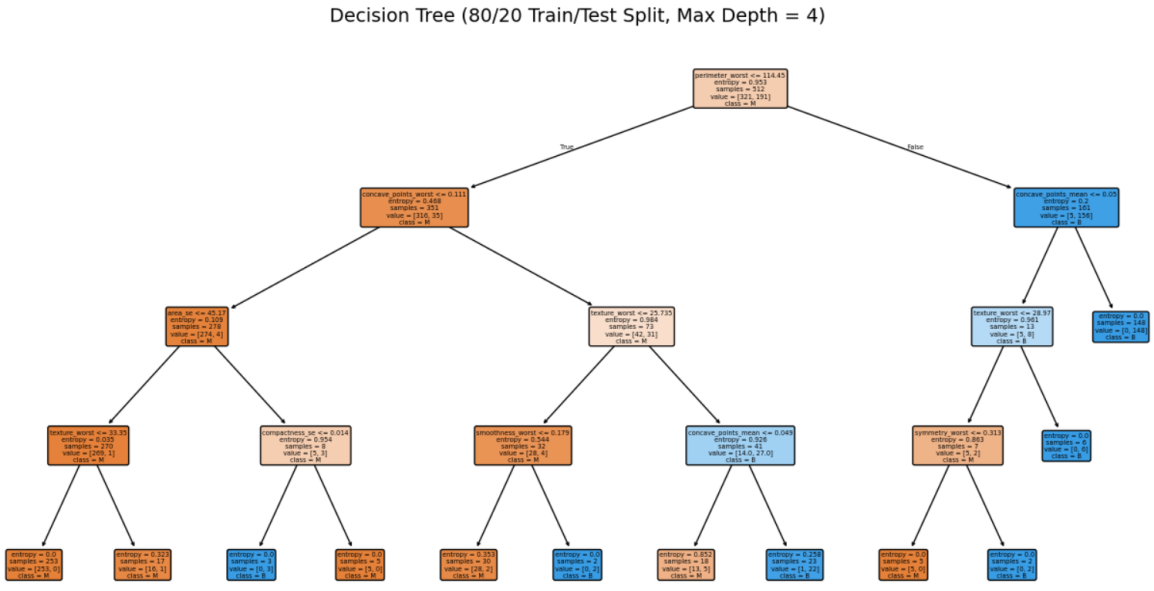


Figure 19 - The depth and accuracy of a decision tree (max depth = 4)

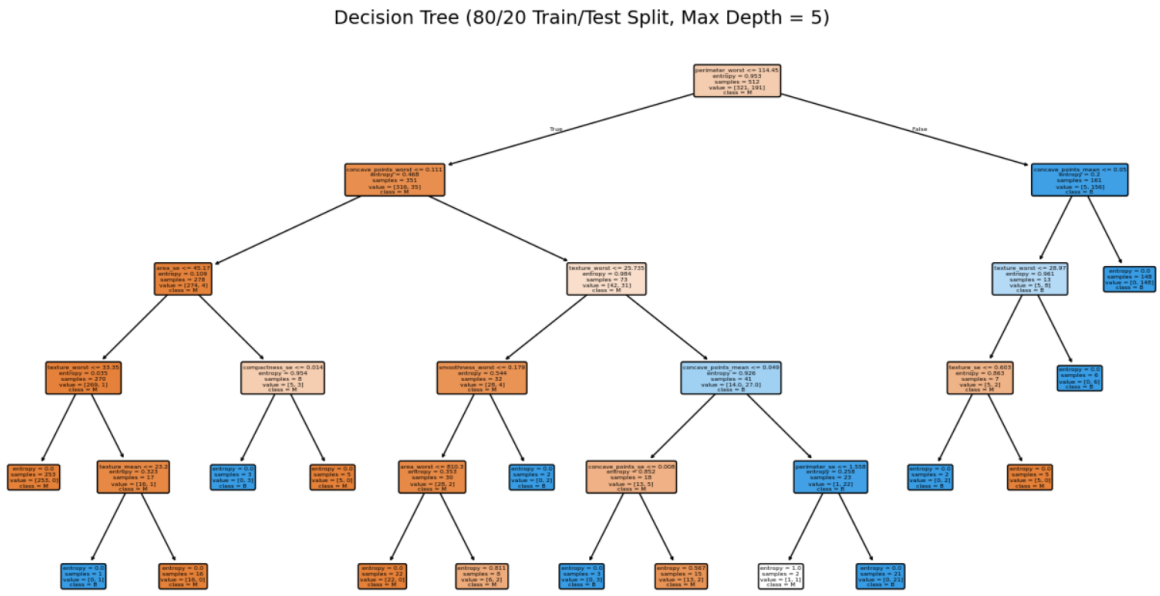


Figure 20 - The depth and accuracy of a decision tree (max depth = 5)

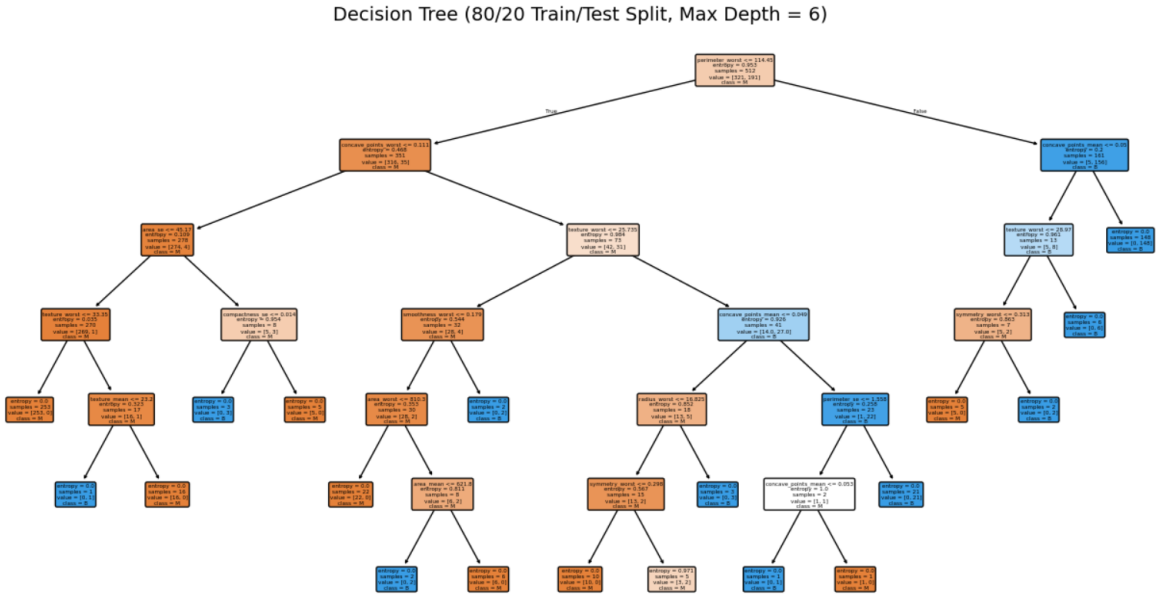


Figure 21 - The depth and accuracy of a decision tree (max depth = 6)

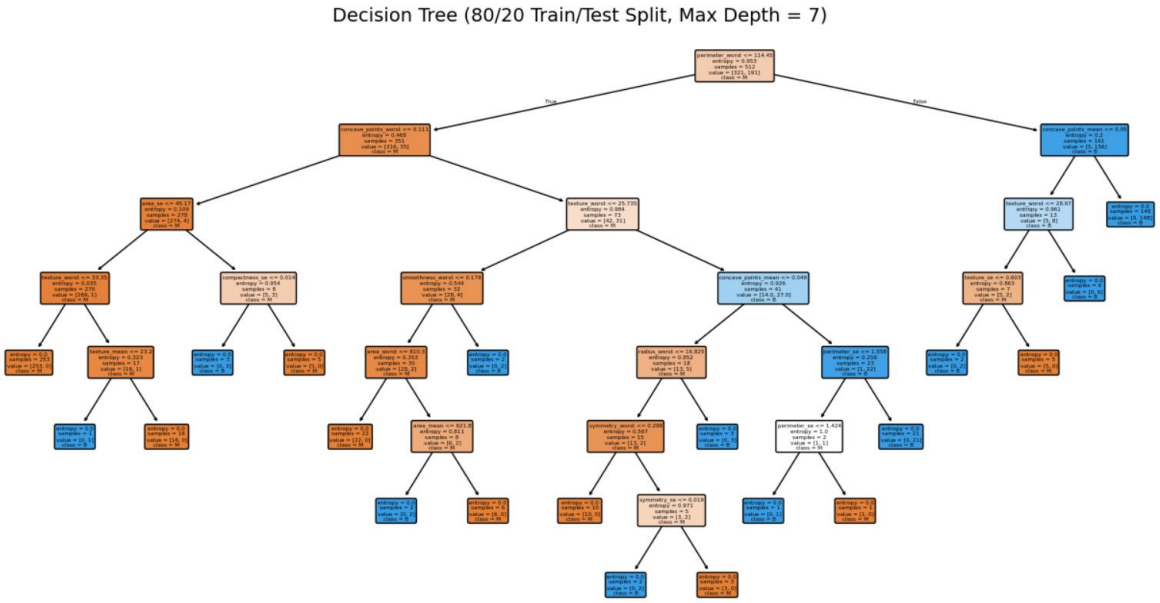
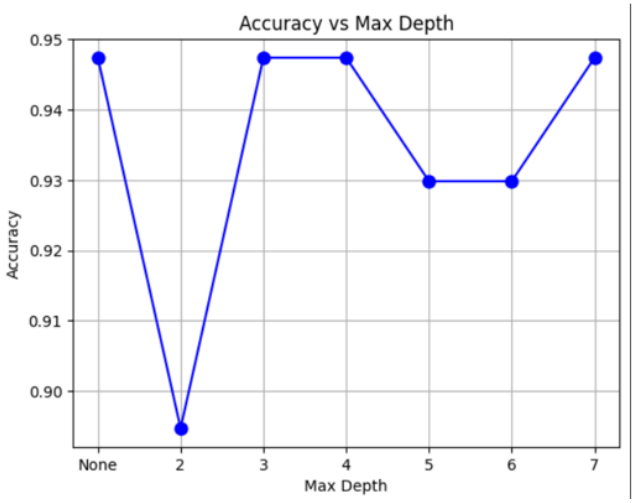


Figure 22 - The depth and accuracy of a decision tree (max depth = 7)




```
Accuracy for different max_depth values:

  Max Depth  Accuracy
0      None  0.947368
1         2  0.894737
2         3  0.947368
3         4  0.947368
4         5  0.929825
5         6  0.929825
6         7  0.947368
```

Figure 23 - Accuracy vs Max Depth

2. Wine quality

2.1 Load data

```
Dữ liệu White Wine:
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.0              0.27         0.36            20.7      0.045
1            6.3              0.30         0.34             1.6      0.049
2            8.1              0.28         0.40             6.9      0.050
3            7.2              0.23         0.32             8.5      0.058
4            7.2              0.23         0.32             8.5      0.058

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 45.0                 170.0   1.0010  3.00       0.45
1                 14.0                 132.0   0.9940  3.30       0.49
2                 30.0                  97.0   0.9951  3.26       0.44
3                 47.0                 186.0   0.9956  3.19       0.40
4                 47.0                 186.0   0.9956  3.19       0.40

   alcohol  quality
0      8.8        6
1      9.5        6
2     10.1        6
3      9.9        6
4      9.9        6

Thông tin dữ liệu:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
...
4     163
3      20
9       5
Name: count, dtype: int64
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

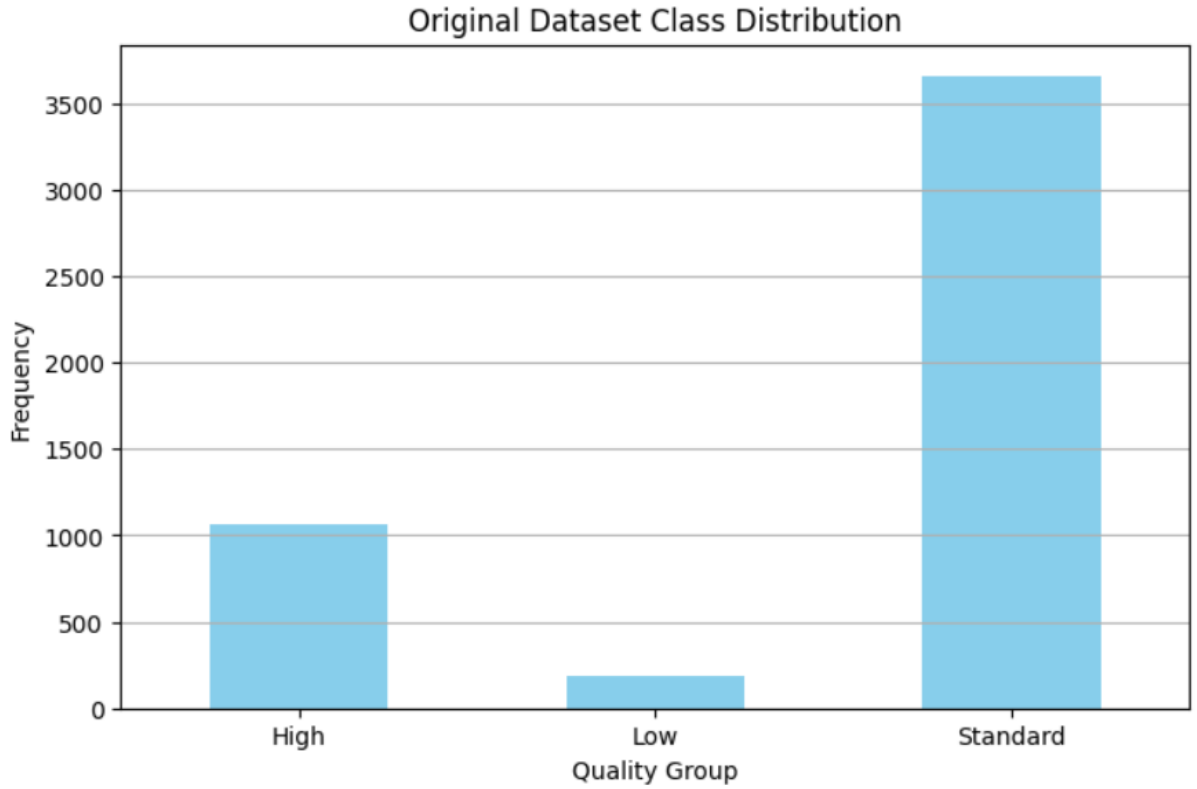
Group the quality

```
Phân phối nhóm chất lượng:
quality_grouped
Standard    3655
High        1060
Low          183
Name: count, dtype: int64
```

2.2 Step to perform

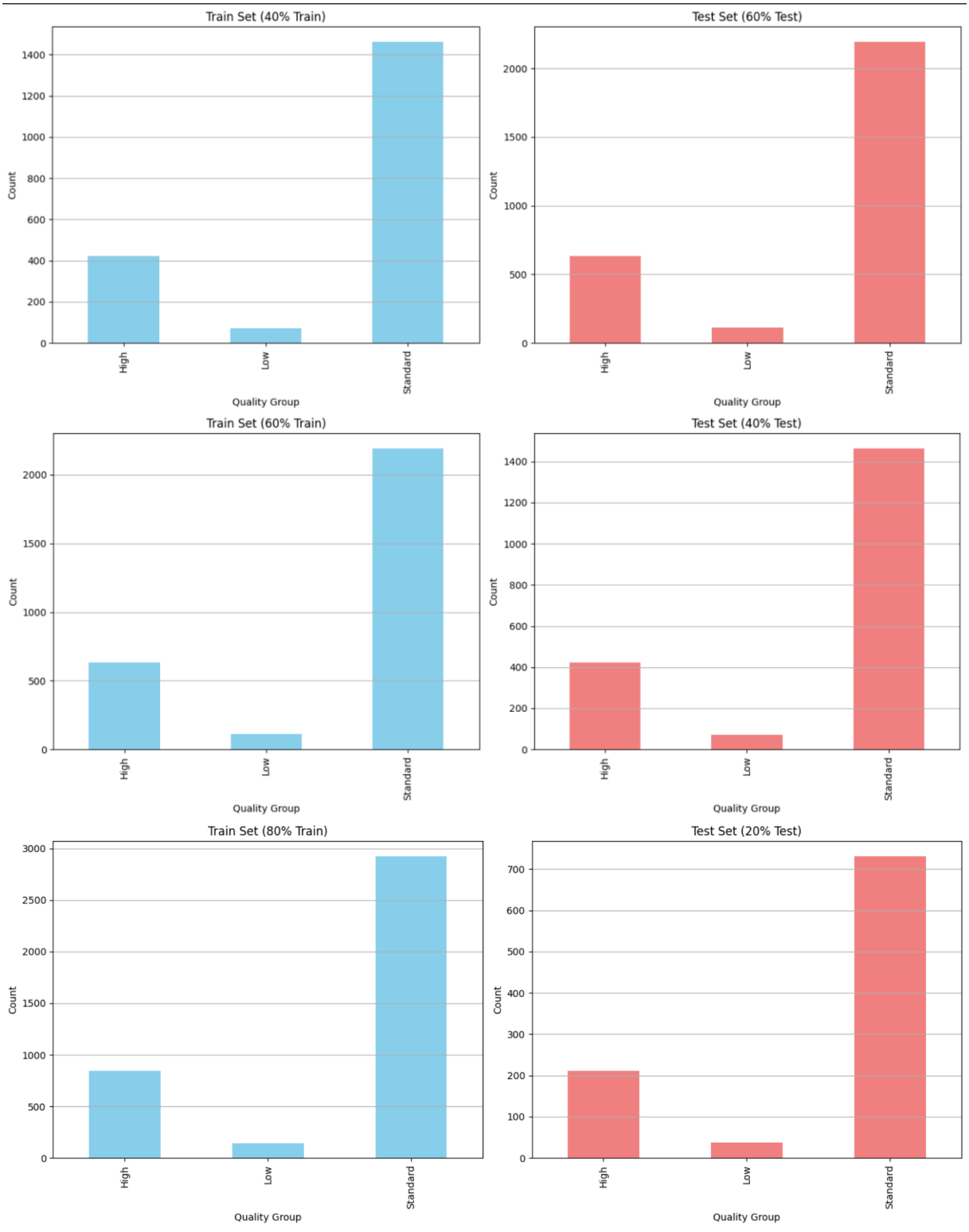
2.2.1 Preparing data

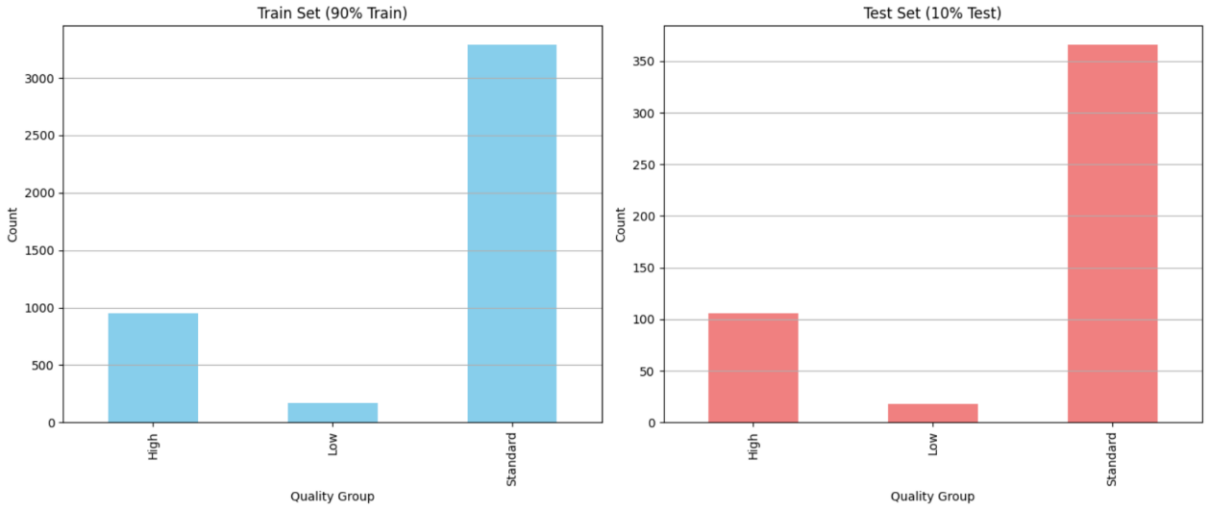
2.2.1.1 Analyze the original data



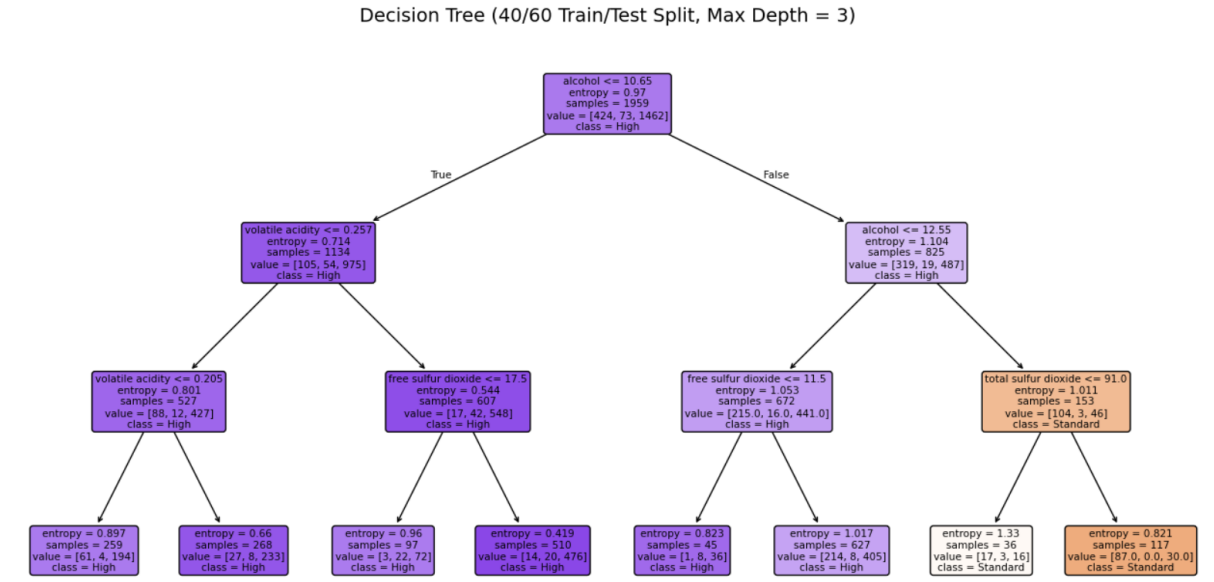
2.2.1.2 Split data into train/test with different ratios

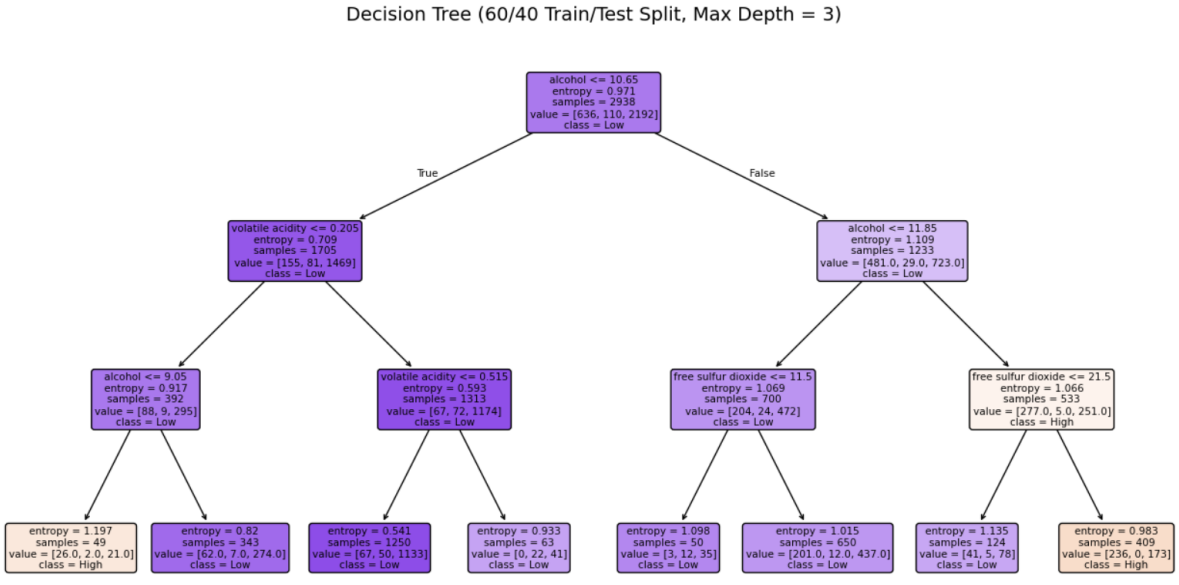
```
Tỉ lệ 40/60:
  Số lượng mẫu huấn luyện: 1959
  Số lượng mẫu kiểm tra: 2939
Tỉ lệ 60/40:
  Số lượng mẫu huấn luyện: 2938
  Số lượng mẫu kiểm tra: 1960
Tỉ lệ 80/20:
  Số lượng mẫu huấn luyện: 3918
  Số lượng mẫu kiểm tra: 980
Tỉ lệ 90/10:
  Số lượng mẫu huấn luyện: 4408
  Số lượng mẫu kiểm tra: 490
```

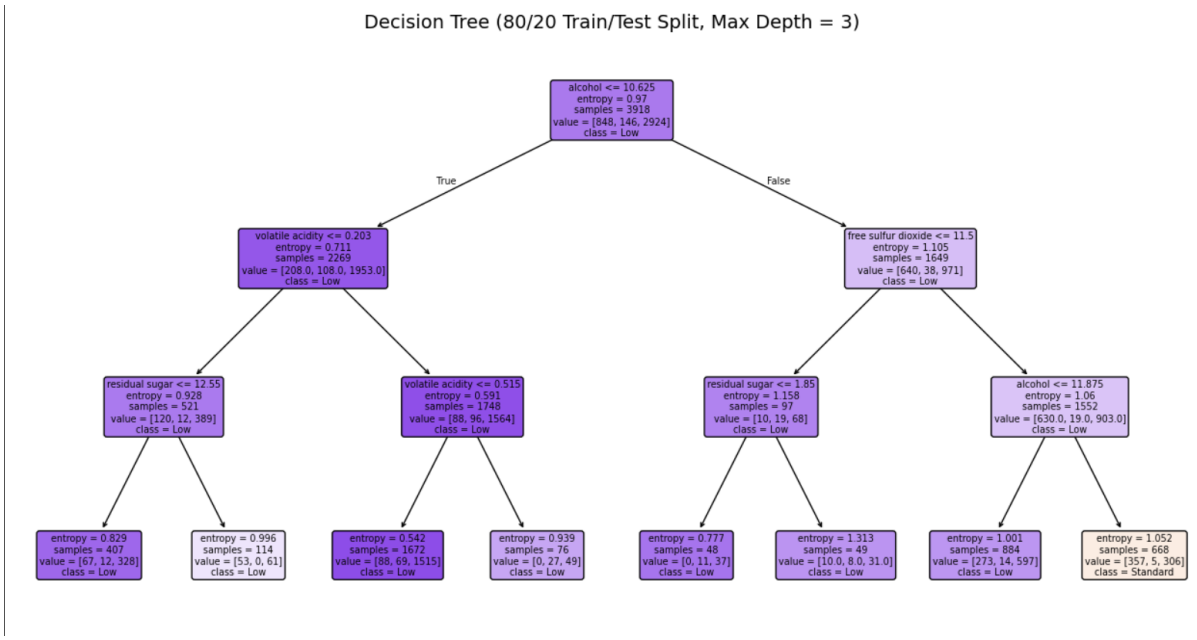

Train Set (40% Train)
Test Set (60% Test)
Train Set (60% Train)
Test Set (40% Test)
Train Set (80% Train)
Test Set (20% Test)
Count
Quality Group
High
Low
Standard

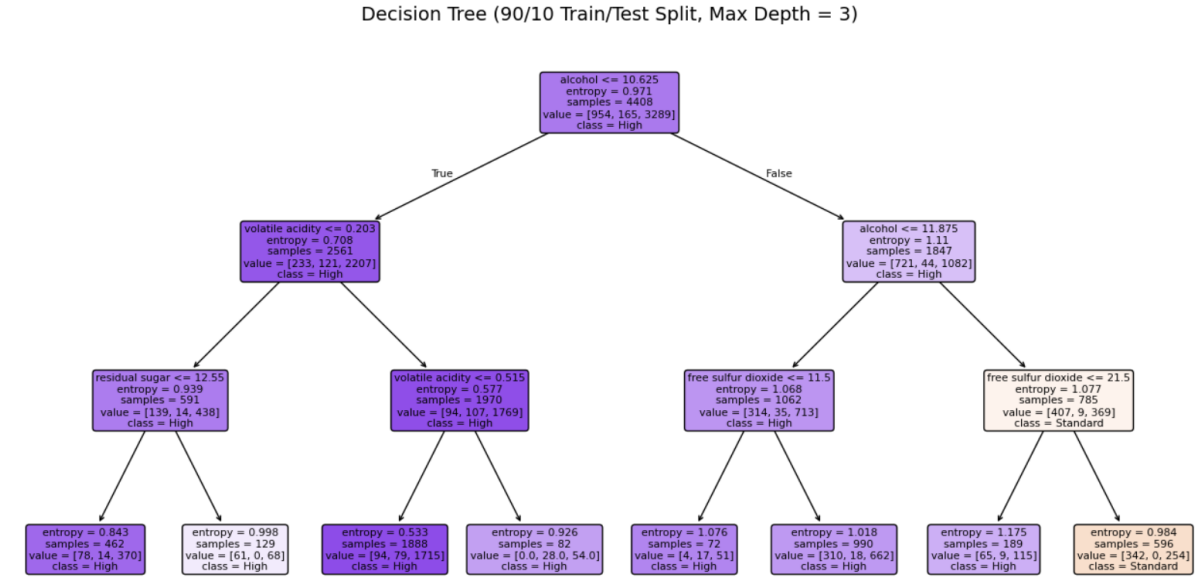


2.2.1.3 Build a decision classifier and visualize it.



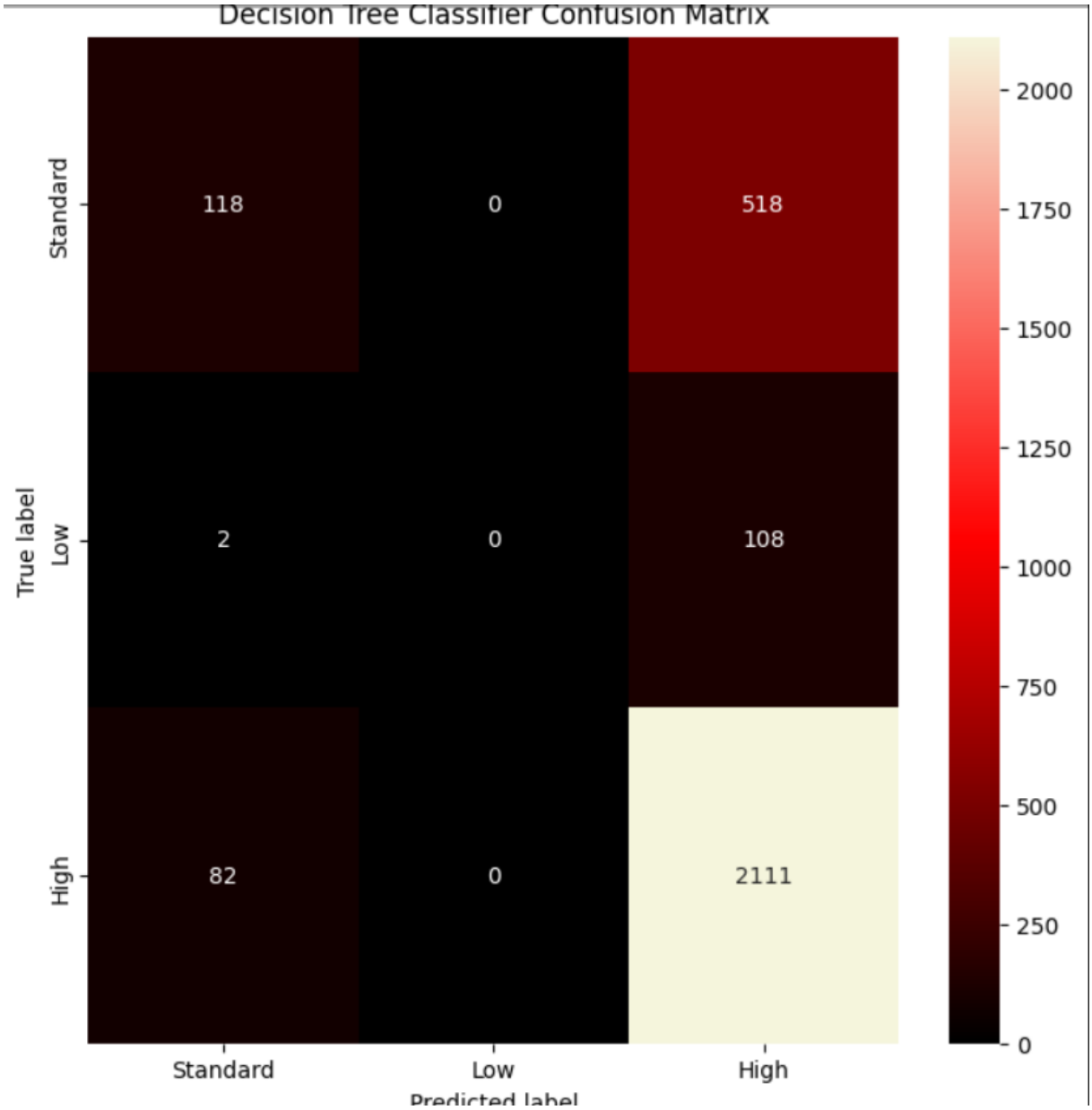



Decision Tree (80/20 Train/Test Split, Max Depth = 3)
alcohol <= 10.625
entropy = 0.97
samples = 3918
value = [848, 146, 2924]
class = Low
True
False
volatile acidity <= 0.203
entropy = 0.711
samples = 2269
value = [208.0, 108.0, 1953.0]
class = Low
free sulfur dioxide <= 11.5
entropy = 1.105
samples = 1649
value = [640, 38, 971]
class = Low
residual sugar <= 12.55
entropy = 0.928
samples = 521
value = [120, 12, 389]
class = Low
volatile acidity <= 0.515
entropy = 0.591
samples = 1748
value = [88, 96, 1564]
class = Low
residual sugar <= 1.85
entropy = 1.158
samples = 97
value = [10, 19, 68]
class = Low
alcohol <= 11.875
entropy = 1.06
samples = 1552
value = [630.0, 19.0, 903.0]
class = Low
entropy = 0.829
samples = 407
value = [67, 12, 328]
class = Low
entropy = 0.996
samples = 114
value = [53, 0, 61]
class = Low
entropy = 0.542
samples = 1672
value = [88, 69, 1515]
class = Low
entropy = 0.939
samples = 76
value = [0, 27, 49]
class = Low
entropy = 0.777
samples = 48
value = [0, 11, 37]
class = Low
entropy = 1.313
samples = 49
value = [10.0, 8.0, 31.0]
class = Low
entropy = 1.001
samples = 884
value = [273, 14, 597]
class = Low
entropy = 1.052
samples = 668
value = [357, 5, 306]
class = Standard

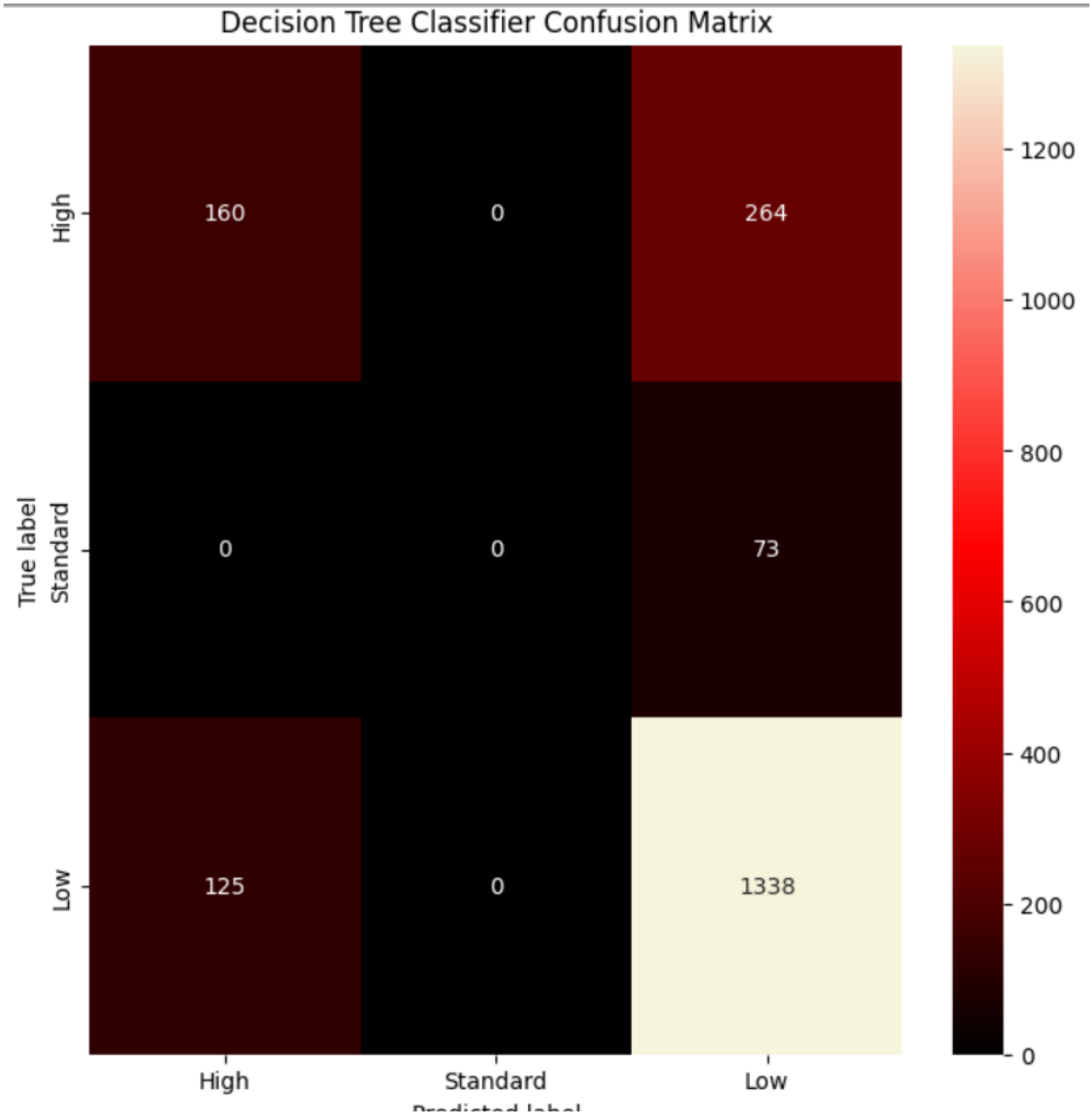

Decision Tree (90/10 Train/Test Split, Max Depth = 3)
alcohol <= 10.625
entropy = 0.971
samples = 4408
value = [954, 165, 3289]
class = High
True
False
volatile acidity <= 0.203
entropy = 0.708
samples = 2561
value = [233, 121, 2207]
class = High
alcohol <= 11.875
entropy = 1.11
samples = 1847
value = [721, 44, 1082]
class = High
residual sugar <= 12.55
entropy = 0.939
samples = 591
value = [139, 14, 438]
class = High
volatile acidity <= 0.515
entropy = 0.577
samples = 1970
value = [94, 107, 1769]
class = High
free sulfur dioxide <= 11.5
entropy = 1.068
samples = 1062
value = [314, 35, 713]
class = High
free sulfur dioxide <= 21.5
entropy = 1.077
samples = 785
value = [407, 9, 369]
class = Standard
entropy = 0.843
samples = 462
value = [78, 14, 370]
class = High
entropy = 0.998
samples = 129
value = [61, 0, 68]
class = High
entropy = 0.533
samples = 1888
value = [94, 79, 1715]
class = High
entropy = 0.926
samples = 82
value = [0.0, 28.0, 54.0]
class = High
entropy = 1.076
samples = 72
value = [4, 17, 51]
class = High
entropy = 1.018
samples = 990
value = [310, 18, 662]
class = High
entropy = 1.175
samples = 189
value = [65, 9, 115]
class = High
entropy = 0.984
samples = 596
value = [342, 0, 254]
class = Standard

2.2.1.4 Evaluate the decision tree classifier

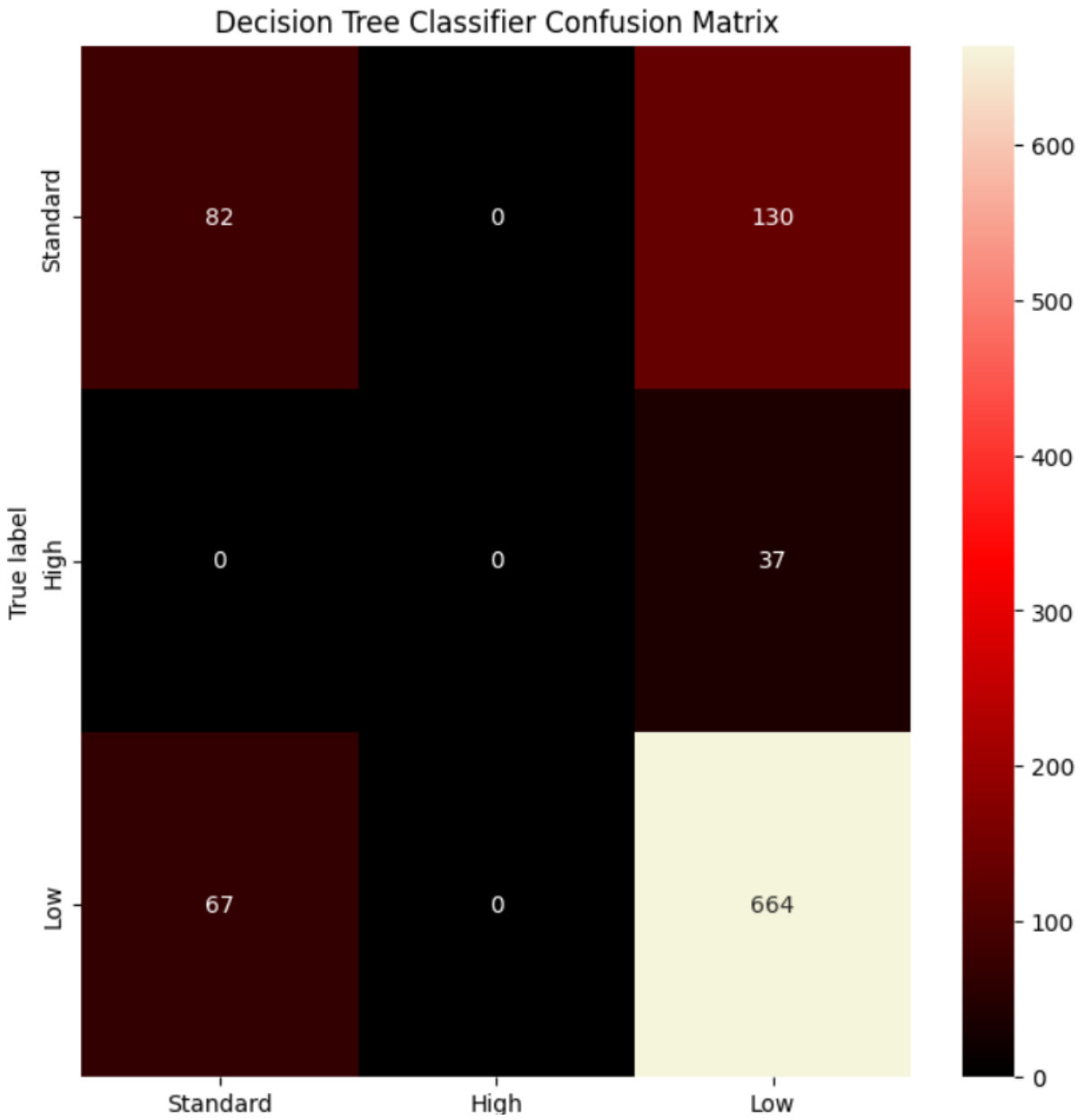
```
Evaluating Decision Tree for split 40/60:
Classification Report:
              precision    recall  f1-score   support

        High       0.58      0.19      0.28       636
         Low       0.00      0.00      0.00       110
    Standard       0.77      0.96      0.86      2193

    accuracy                           0.76      2939
   macro avg       0.45      0.38      0.38      2939
weighted avg       0.70      0.76      0.70      2939

Confusion Matrix:
[[ 118    0  518]
 [   2    0  108]
 [  82    0 2111]]
```


Decision Tree Classifier Confusion Matrix
Standard
Low
High
True label
118
0
518
2
0
108
82
0
2111
Standard
Low
High
Predicted label
2000
1750
1500
1250
1000
750
500
250
0

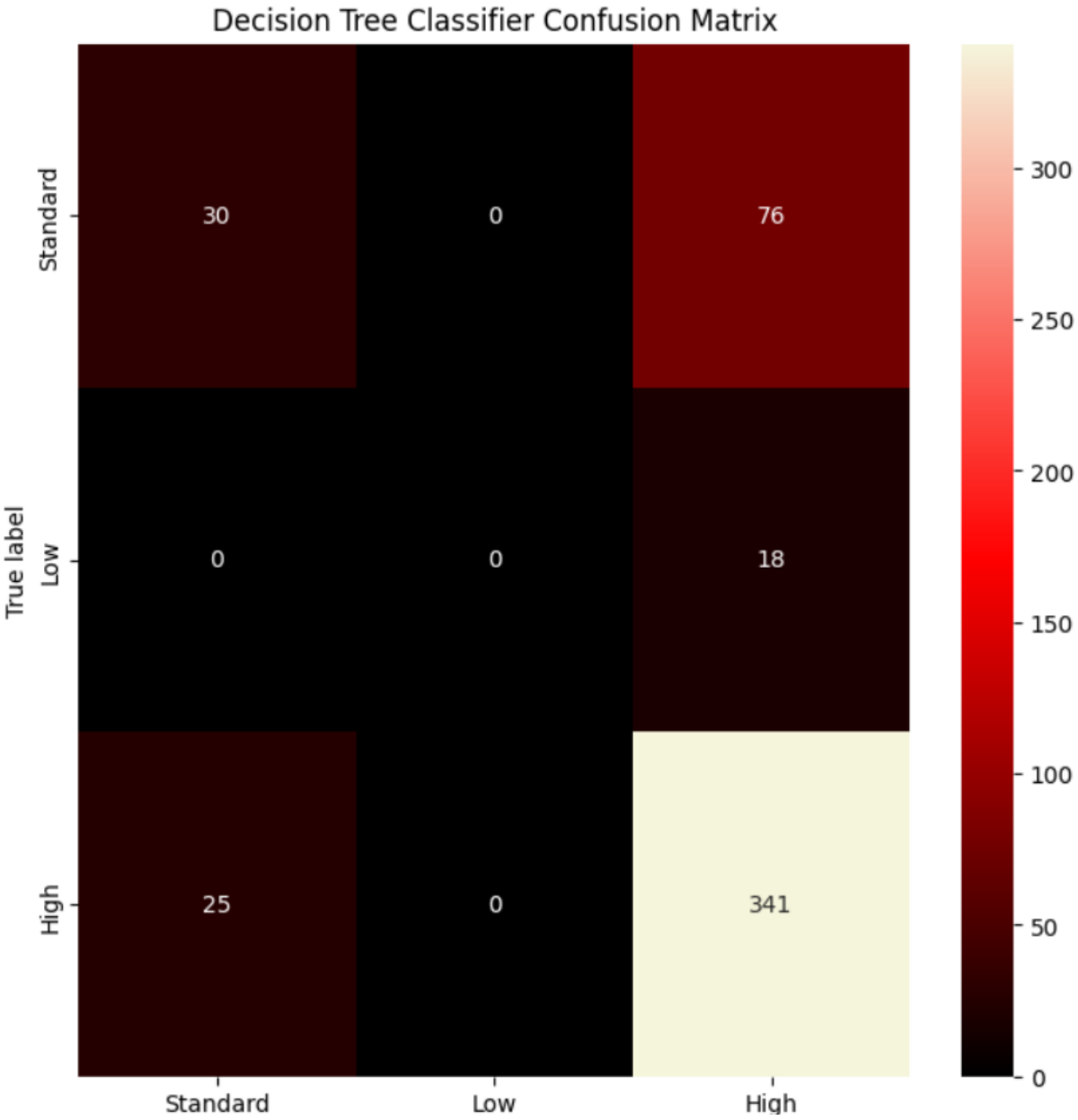
```
Evaluating Decision Tree for split 60/40:
Classification Report:
              precision    recall  f1-score   support

        High       0.56      0.38      0.45       424
         Low       0.00      0.00      0.00        73
    Standard       0.80      0.91      0.85      1463

    accuracy                           0.76      1960
   macro avg       0.45      0.43      0.43      1960
weighted avg       0.72      0.76      0.73      1960

Confusion Matrix:
[[ 160    0  264]
 [   0    0   73]
 [ 125    0 1338]]
```


Decision Tree Classifier Confusion Matrix
160
0
264
0
0
73
125
0
1338
High
Standard
Low
True label
High
Standard
Low
1200
1000
800
600
400
200
0

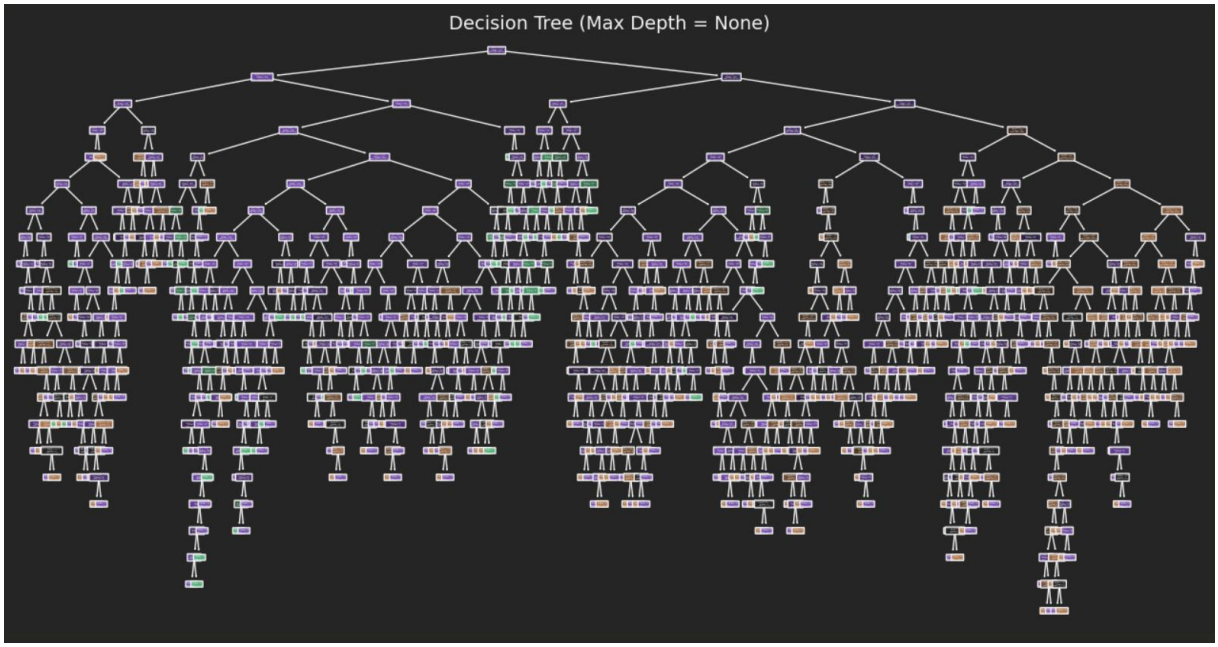
```
Evaluating Decision Tree for split 80/20:
Classification Report:
              precision    recall  f1-score   support

        High       0.55      0.39      0.45       212
         Low       0.00      0.00      0.00        37
    Standard       0.80      0.91      0.85       731

    accuracy                           0.76       980
   macro avg       0.45      0.43      0.43       980
weighted avg       0.72      0.76      0.73       980

Confusion Matrix:
[[ 82   0 130]
 [  0   0  37]
 [ 67   0 664]]
```

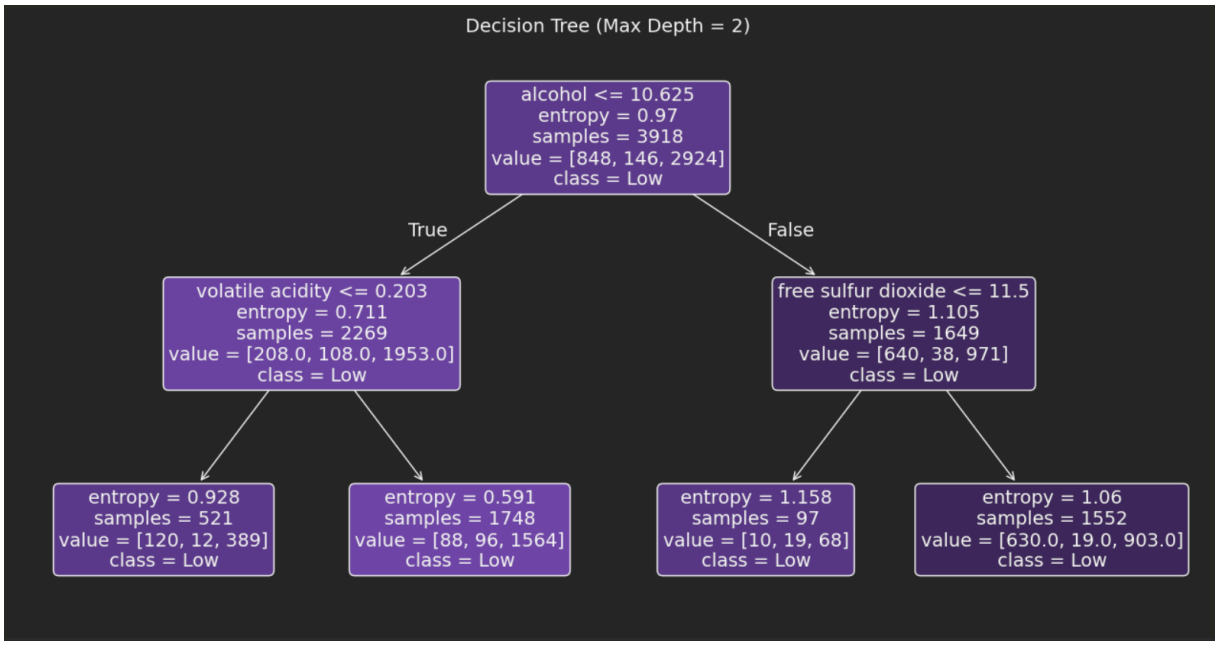

Decision Tree Classifier Confusion Matrix
True label
Standard
High
Low
82
0
130
0
0
37
67
0
664
Standard
High
Low
600
500
400
300
200
100
0

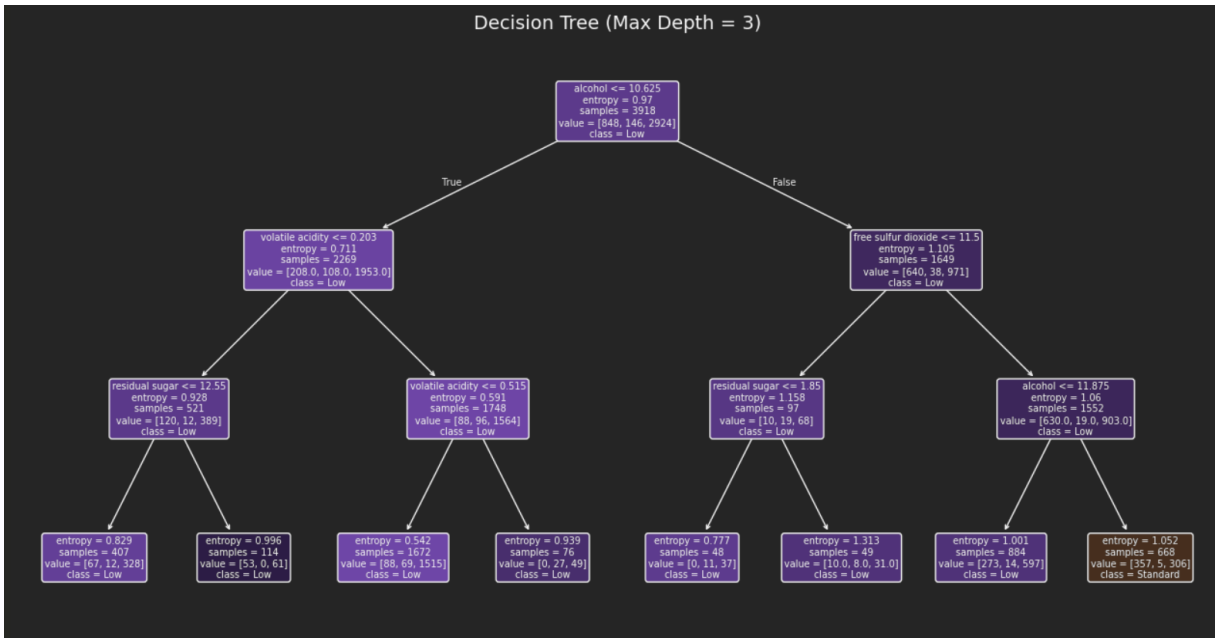
```
Evaluating Decision Tree for split 90/10:
Classification Report:
              precision    recall  f1-score   support

        High       0.55      0.28      0.37       106
         Low       0.00      0.00      0.00        18
    Standard       0.78      0.93      0.85       366

    accuracy                           0.76       490
   macro avg       0.44      0.40      0.41       490
weighted avg       0.70      0.76      0.72       490

Confusion Matrix:
[[ 30   0  76]
 [  0   0  18]
 [ 25   0 341]]
```

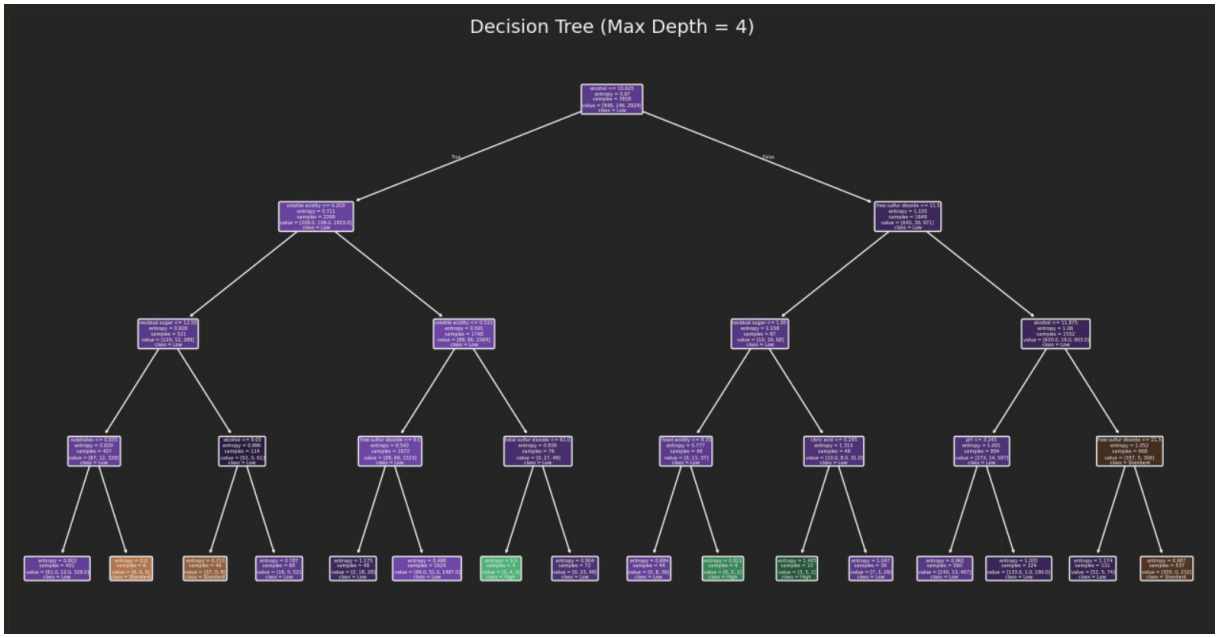

Decision Tree Classifier Confusion Matrix
True label
Standard
Low
High
30
0
76
0
0
18
25
0
341
Standard
Low
High
300
250
200
150
100
50
0

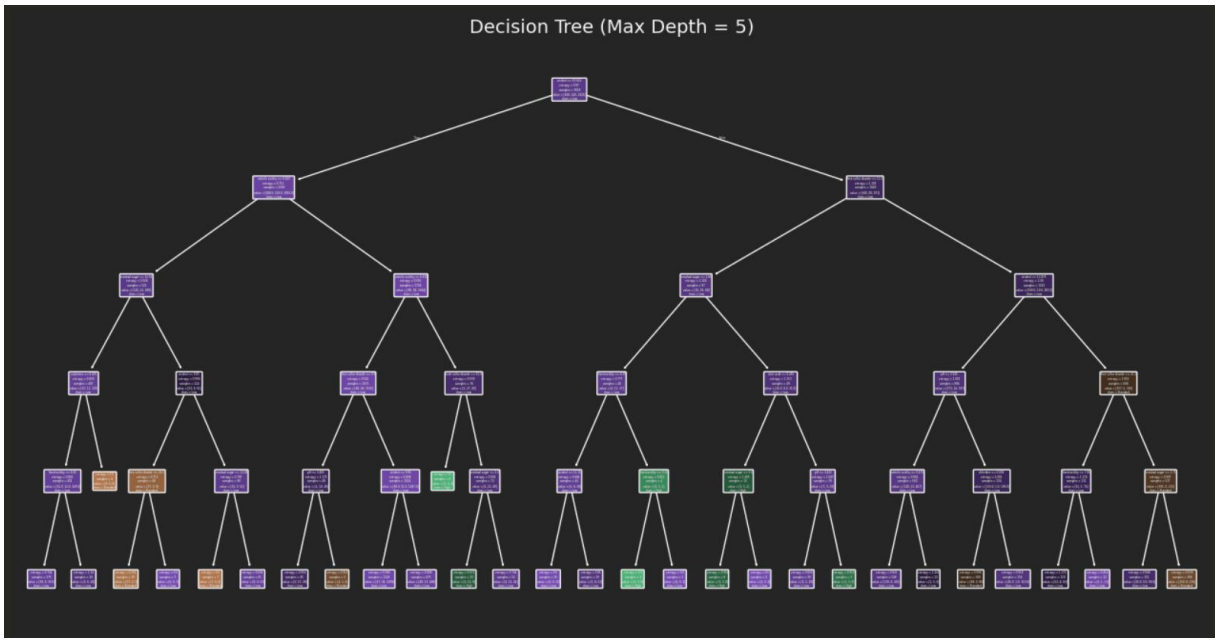
2.2.1.5 Depth and accuracy of the decision tree

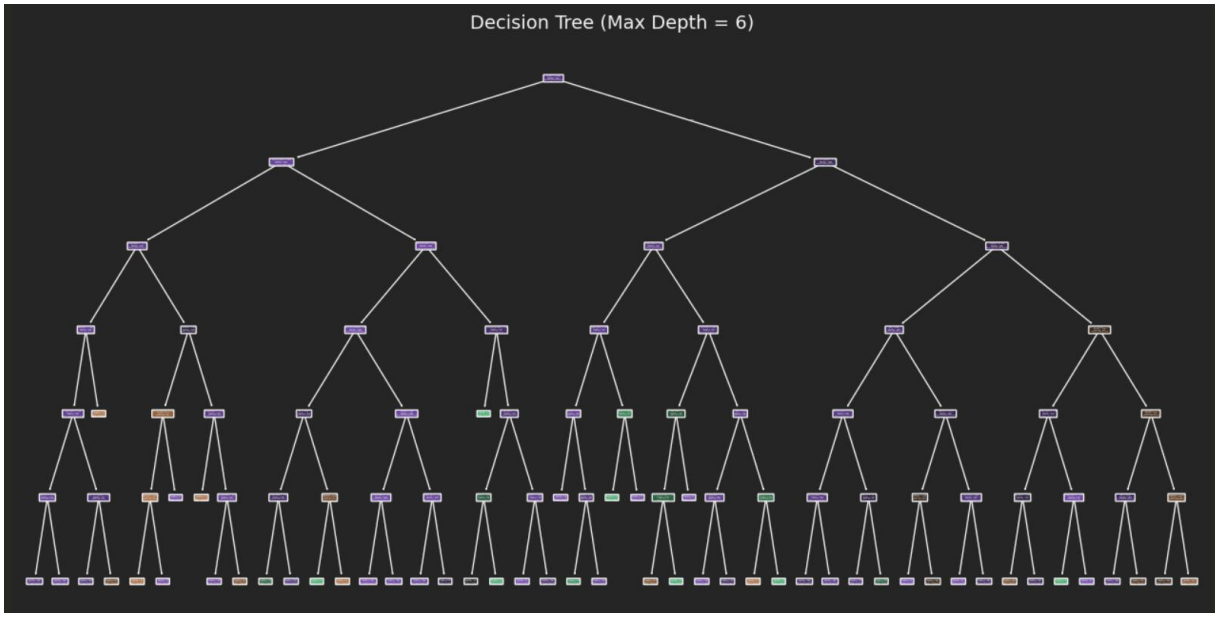


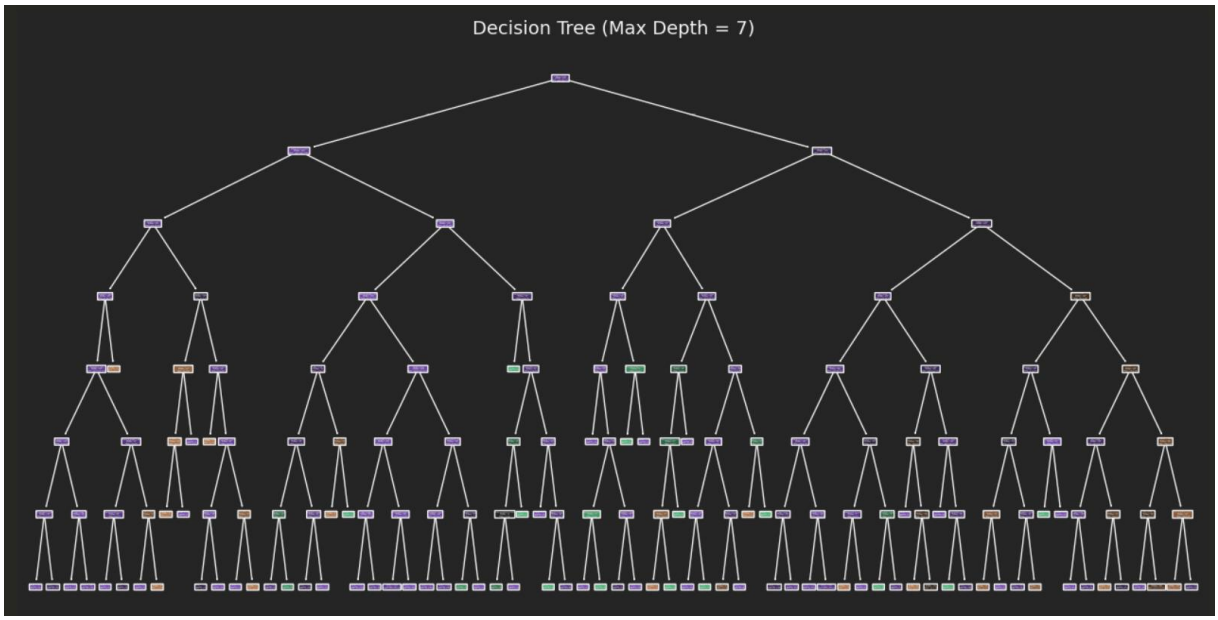


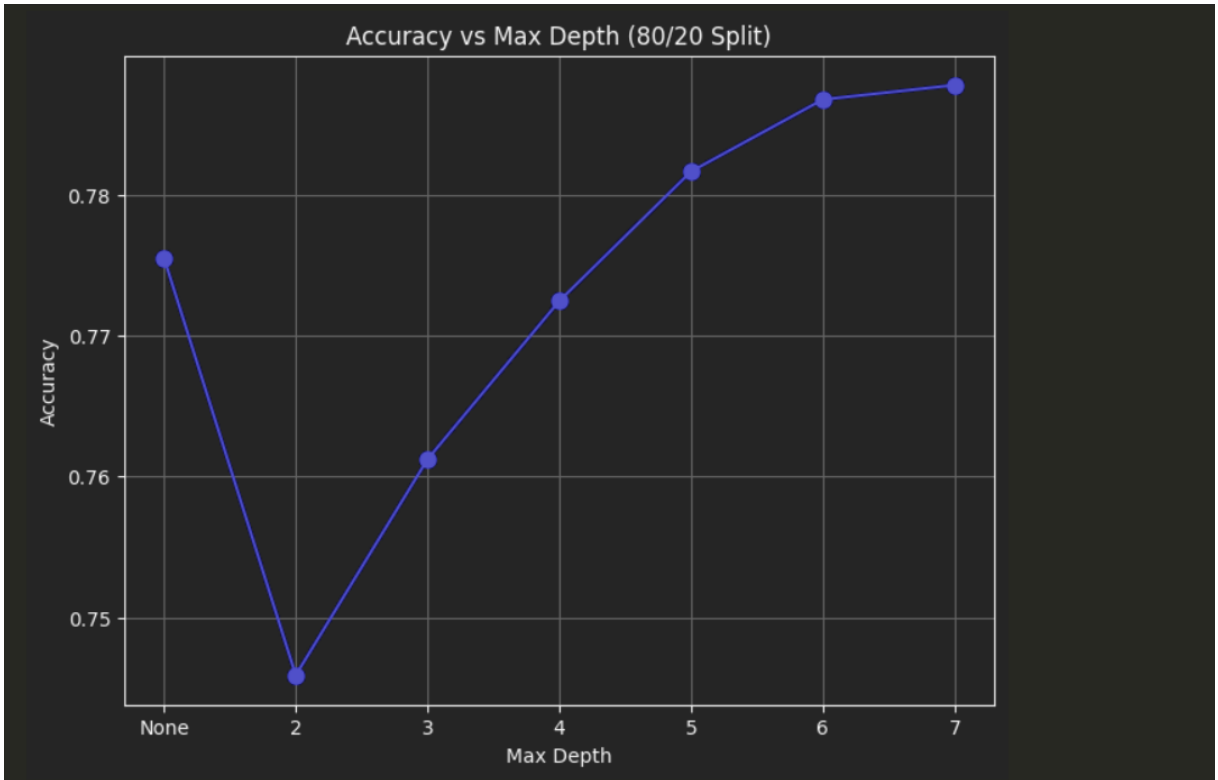

Decision Tree (Max Depth = 3)
alcohol <= 10.625
entropy = 0.97
samples = 3918
value = [848, 146, 2924]
class = Low
True
False
volatile acidity <= 0.203
entropy = 0.711
samples = 2269
value = [208.0, 108.0, 1953.0]
class = Low
free sulfur dioxide <= 11.5
entropy = 1.105
samples = 1649
value = [640, 38, 971]
class = Low
residual sugar <= 12.55
entropy = 0.928
samples = 521
value = [120, 12, 389]
class = Low
volatile acidity <= 0.515
entropy = 0.591
samples = 1748
value = [88, 96, 1564]
class = Low
residual sugar <= 1.85
entropy = 1.158
samples = 97
value = [10, 19, 68]
class = Low
alcohol <= 11.875
entropy = 1.06
samples = 1552
value = [630.0, 19.0, 903.0]
class = Low
entropy = 0.829
samples = 407
value = [67, 12, 328]
class = Low
entropy = 0.996
samples = 114
value = [53, 0, 61]
class = Low
entropy = 0.542
samples = 1672
value = [88, 69, 1515]
class = Low
entropy = 0.939
samples = 76
value = [0, 27, 49]
class = Low
entropy = 0.777
samples = 48
value = [0, 11, 37]
class = Low
entropy = 1.313
samples = 49
value = [10.0, 8.0, 31.0]
class = Low
entropy = 1.001
samples = 884
value = [273, 14, 597]
class = Low
entropy = 1.052
samples = 668
value = [357, 5, 306]
class = Standard


Decision Tree (Max Depth = 4)


Decision Tree (Max Depth = 5)


Decision Tree (Max Depth = 6)





Accuracy for different max_depth values (Horizontal Format):

	None	2	3	4	5	6	7
Accuracy	0.77551	0.745918	0.761224	0.772449	0.781633	0.786735	0.787755

```
Accuracy for different max_depth values (Horizontal Format):
             None         2         3         4         5         6         7
Accuracy  0.77551  0.745918  0.761224  0.772449  0.781633  0.786735  0.787755
```

3. Additional dataset.

3.1. Load data.

3. 1.1. Import libraries.

```
1  import pandas as pd
2  import matplotlib.pyplot as plt
3  import seaborn as sns
4  from sklearn.model_selection import train_test_split
5  from sklearn.tree import DecisionTreeClassifier, plot_tree
6  from sklearn.metrics import classification_report, accuracy_score, confusion_matrix
7  import matplotlib.colors as mcolors
8
   [2]
```

3.1.2. Load and analyze the data.

```
Dữ liệu ban đầu:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

```
Thông tin dữ liệu:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

Phân phối nhãn:
Survived
0    549
1    342
Name: count, dtype: int64
```

3.2 Steps to perform.

3.2.1. Prepare the data.

3.2.1.1. Separate features and labels.

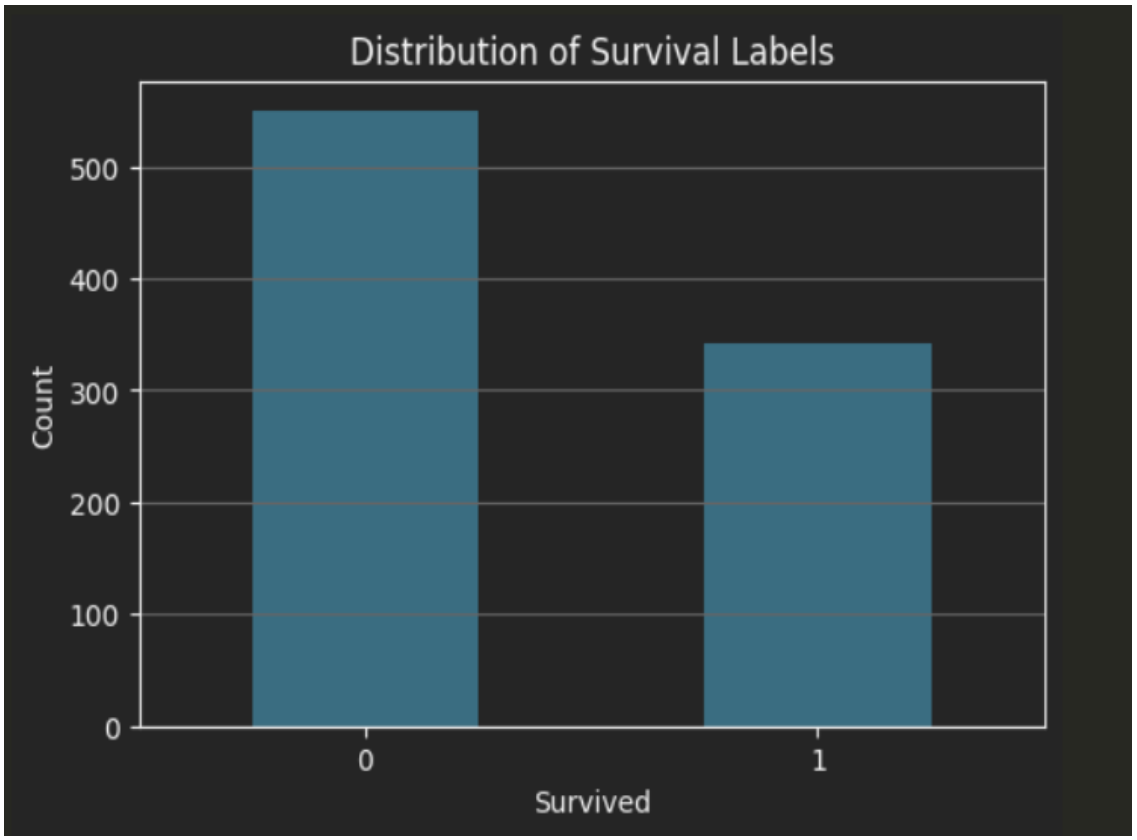
```
Dữ liệu đặc trưng (X):
   Pclass   Age  SibSp  Parch     Fare  Name_Abbott, Mr. Rossmore Edward  \
0       3  22.0      1      0   7.2500                              False
1       1  38.0      1      0  71.2833                              False
2       3  26.0      0      0   7.9250                              False
3       1  35.0      1      0  53.1000                              False
4       3  35.0      0      0   8.0500                              False

   Name_Abbott, Mrs. Stanton (Rosa Hunt)  Name_Abelson, Mr. Samuel  \
0                                  False                     False
1                                  False                     False
2                                  False                     False
3                                  False                     False
4                                  False                     False

   Name_Abelson, Mrs. Samuel (Hannah Wizosky)  \
0                                       False
1                                       False
2                                       False
3                                       False
4                                       False

   Name_Adahl, Mr. Mauritz Nils Martin  ...  Cabin_F G63  Cabin_F G73  \
0                                False  ...        False        False
1                                False  ...        False        False
2                                False  ...        False        False
3                                False  ...        False        False
4                                False  ...        False        False
```

```
   Cabin_F2  Cabin_F33  Cabin_F38  Cabin_F4  Cabin_G6  Cabin_T  Embarked_Q  \
0     False      False      False     False     False    False       False
1     False      False      False     False     False    False       False
2     False      False      False     False     False    False       False
3     False      False      False     False     False    False       False
4     False      False      False     False     False    False       False

   Embarked_S
0        True
1       False
2        True
3        True
4        True

[5 rows x 1724 columns]

Nhãn (y):
0    0
1    1
2    1
3    1
4    0
Name: Survived, dtype: int64
```

3.2.1.2. Analyze the original data.

```
Thống kê mô tả dữ liệu:
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200

Số lượng giá trị bị thiếu:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
```

```
Phân phối các giá trị nhãn (Survived):
Survived
0    549
1    342
Name: count, dtype: int64
```

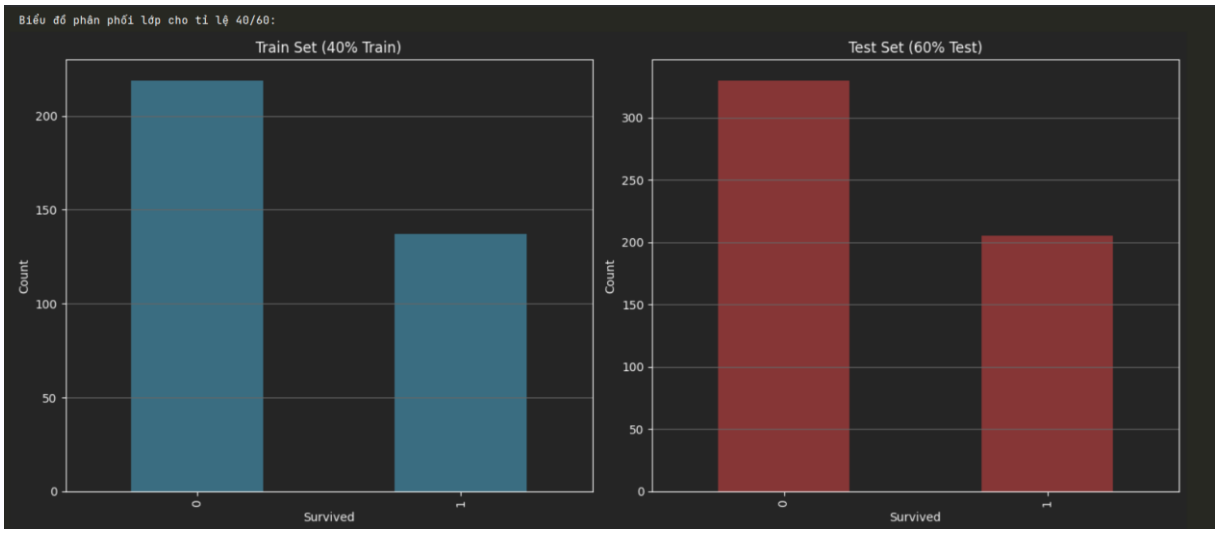


2.1.3. Split data into train/test with different ratios.

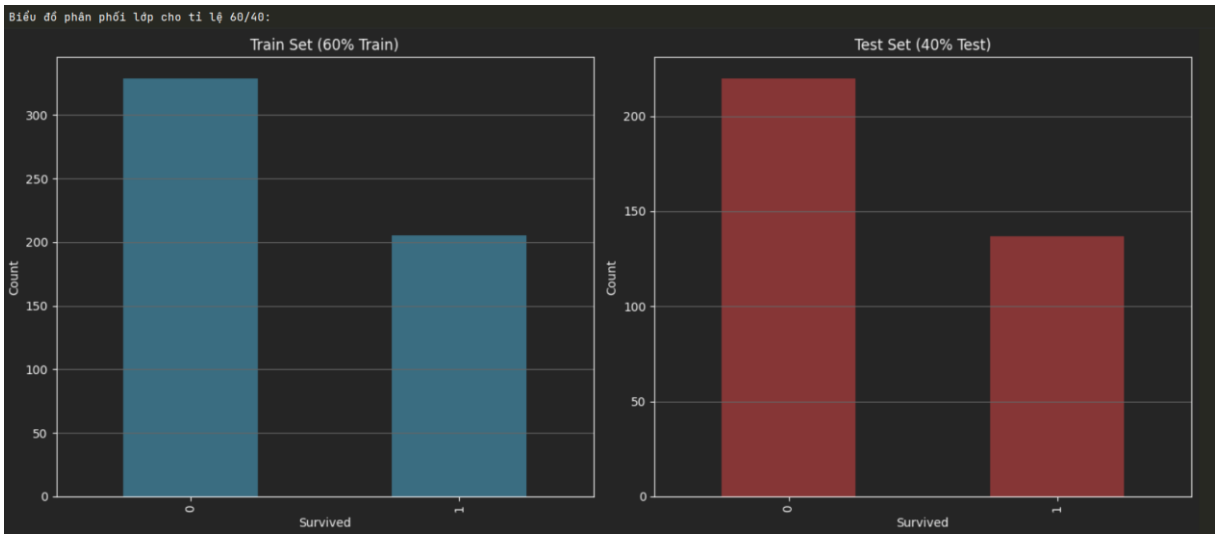
```
Tỉ lệ 40/60:
  Số lượng mẫu huấn luyện: 356
  Số lượng mẫu kiểm tra: 535
Tỉ lệ 60/40:
  Số lượng mẫu huấn luyện: 534
  Số lượng mẫu kiểm tra: 357
Tỉ lệ 80/20:
  Số lượng mẫu huấn luyện: 712
  Số lượng mẫu kiểm tra: 179
Tỉ lệ 90/10:
  Số lượng mẫu huấn luyện: 801
  Số lượng mẫu kiểm tra: 90
```

2.1.4. Chart showing train/test ratios.

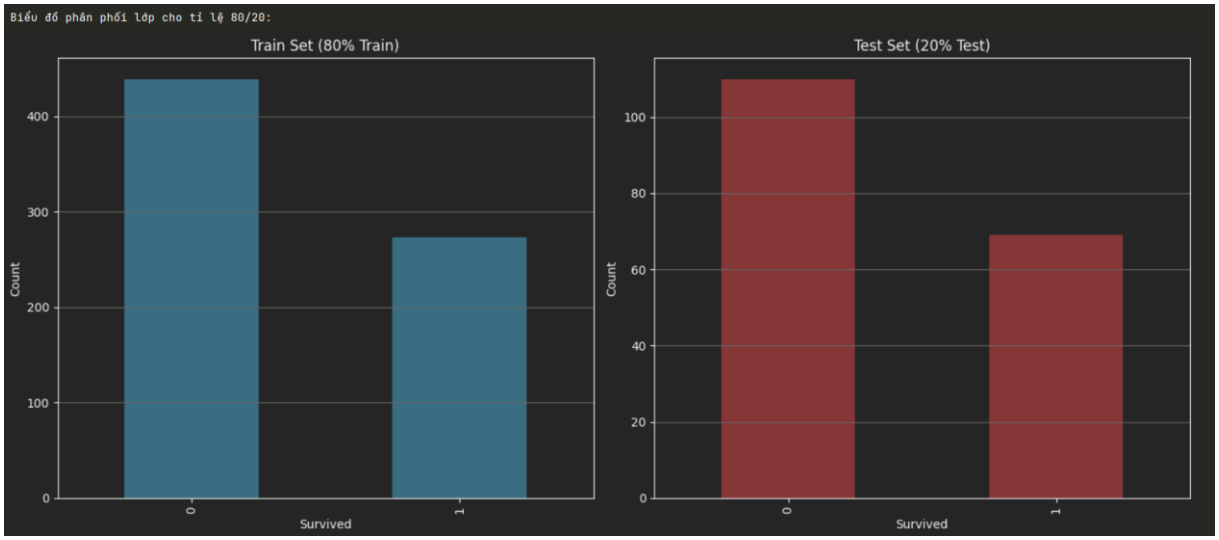
a. Chart for class distribution ratio of 40/60



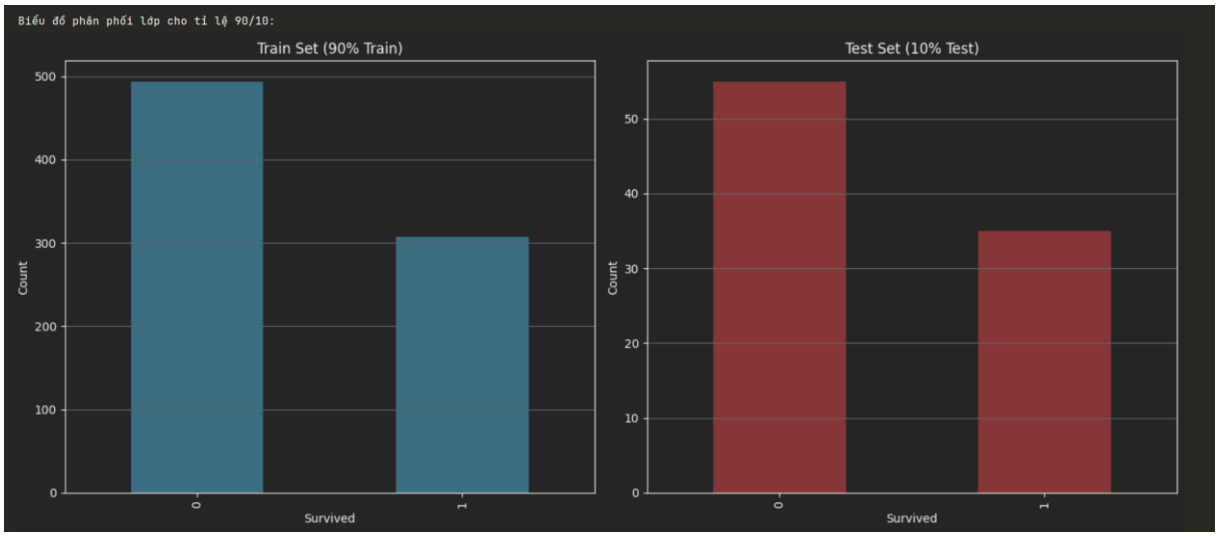
b. Chart for class distribution ratio of 60/40



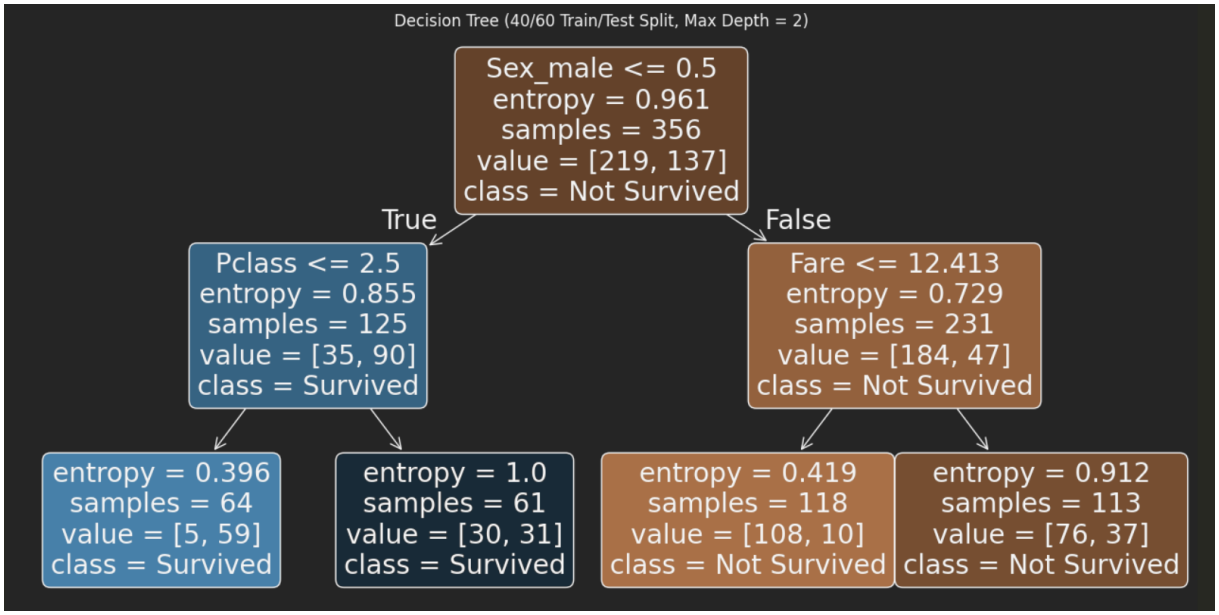
c. Chart for class distribution ratio of 80/20

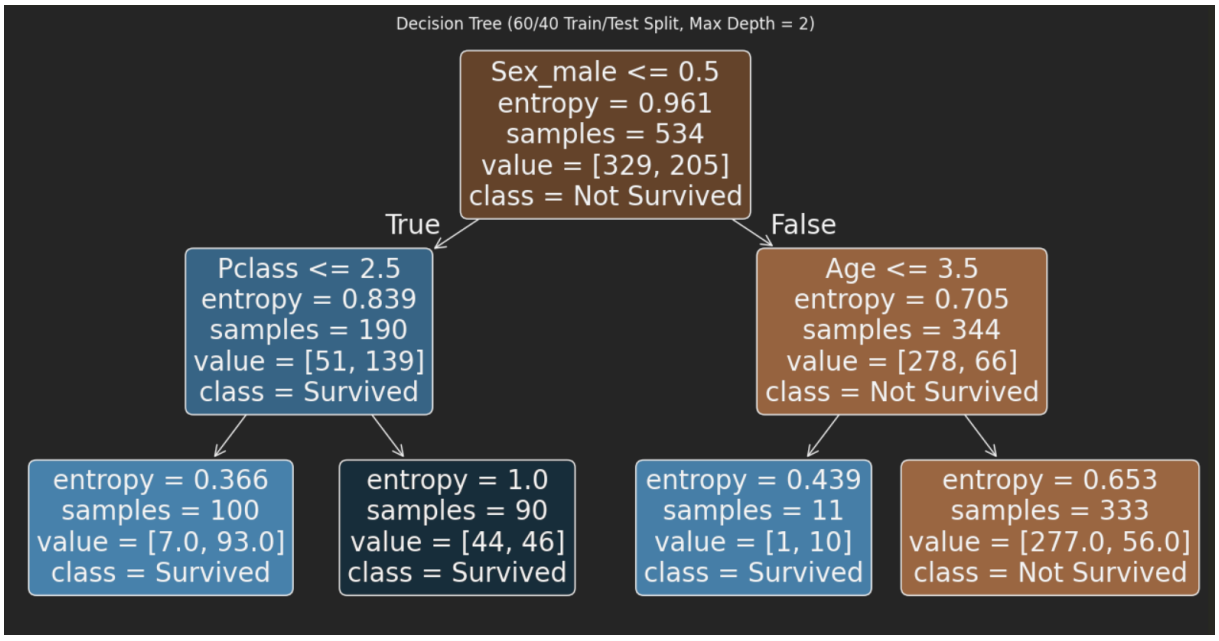


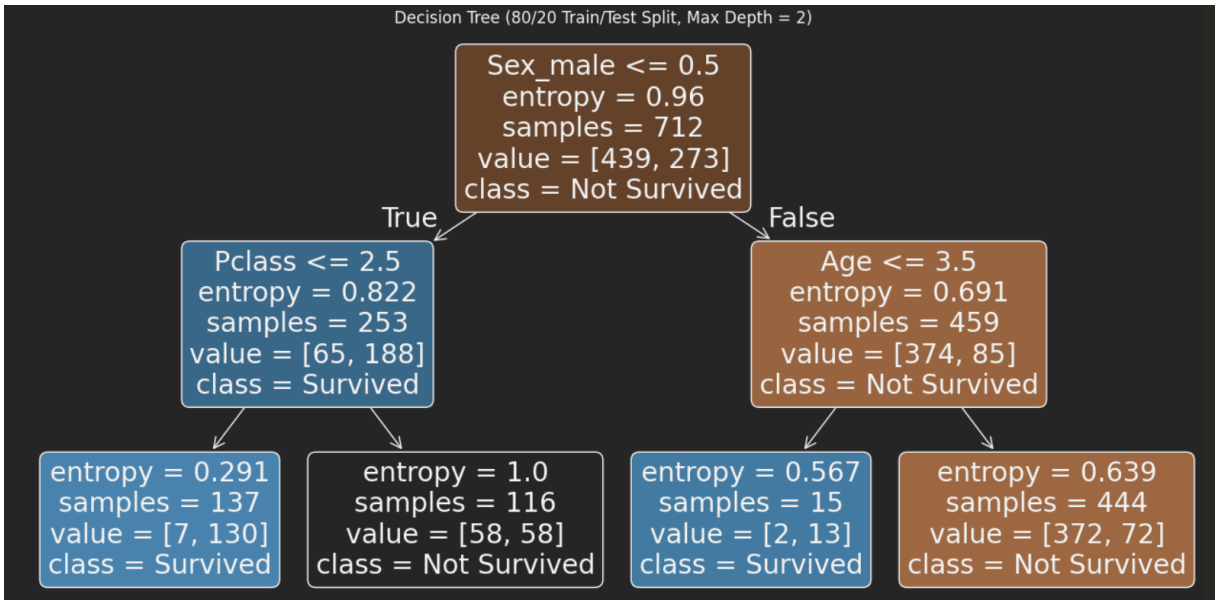
d. Chart for class distribution ratio of 90/10.

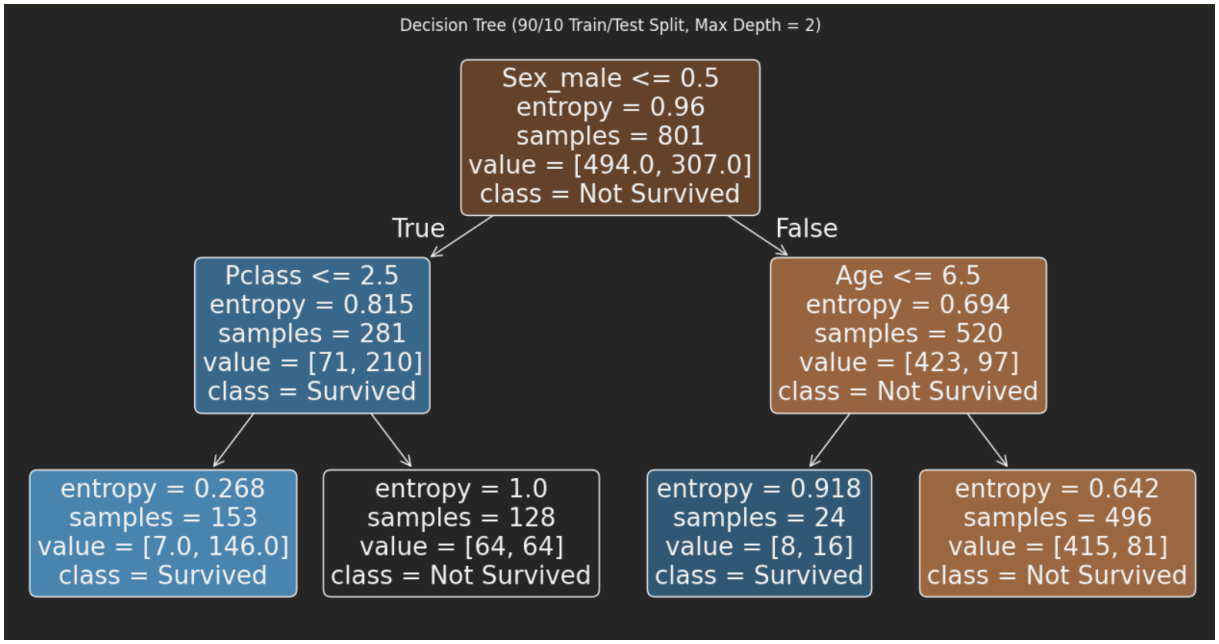


2.2. Build a decision tree classifier and visualize it.



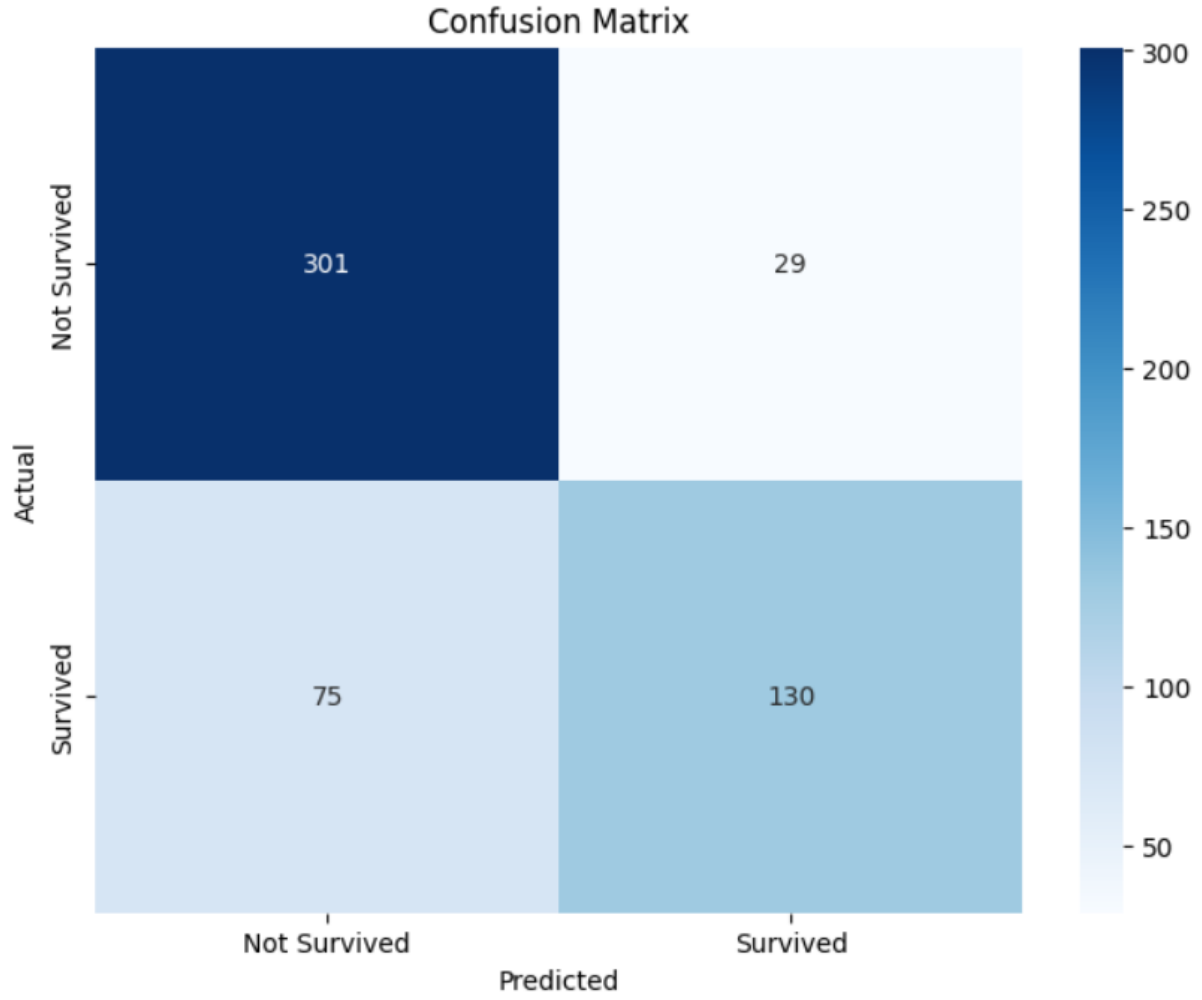



Decision Tree (80/20 Train/Test Split, Max Depth = 2)
Sex_male <= 0.5
entropy = 0.96
samples = 712
value = [439, 273]
class = Not Survived
True
False
Pclass <= 2.5
entropy = 0.822
samples = 253
value = [65, 188]
class = Survived
Age <= 3.5
entropy = 0.691
samples = 459
value = [374, 85]
class = Not Survived
entropy = 0.291
samples = 137
value = [7, 130]
class = Survived
entropy = 1.0
samples = 116
value = [58, 58]
class = Not Survived
entropy = 0.567
samples = 15
value = [2, 13]
class = Survived
entropy = 0.639
samples = 444
value = [372, 72]
class = Not Survived


Decision Tree (90/10 Train/Test Split, Max Depth = 2)
Sex_male <= 0.5
entropy = 0.96
samples = 801
value = [494.0, 307.0]
class = Not Survived
True
False
Pclass <= 2.5
entropy = 0.815
samples = 281
value = [71, 210]
class = Survived
Age <= 6.5
entropy = 0.694
samples = 520
value = [423, 97]
class = Not Survived
entropy = 0.268
samples = 153
value = [7.0, 146.0]
class = Survived
entropy = 1.0
samples = 128
value = [64, 64]
class = Not Survived
entropy = 0.918
samples = 24
value = [8, 16]
class = Survived
entropy = 0.642
samples = 496
value = [415, 81]
class = Not Survived

2.3. Evaluate the decision tree classifier.

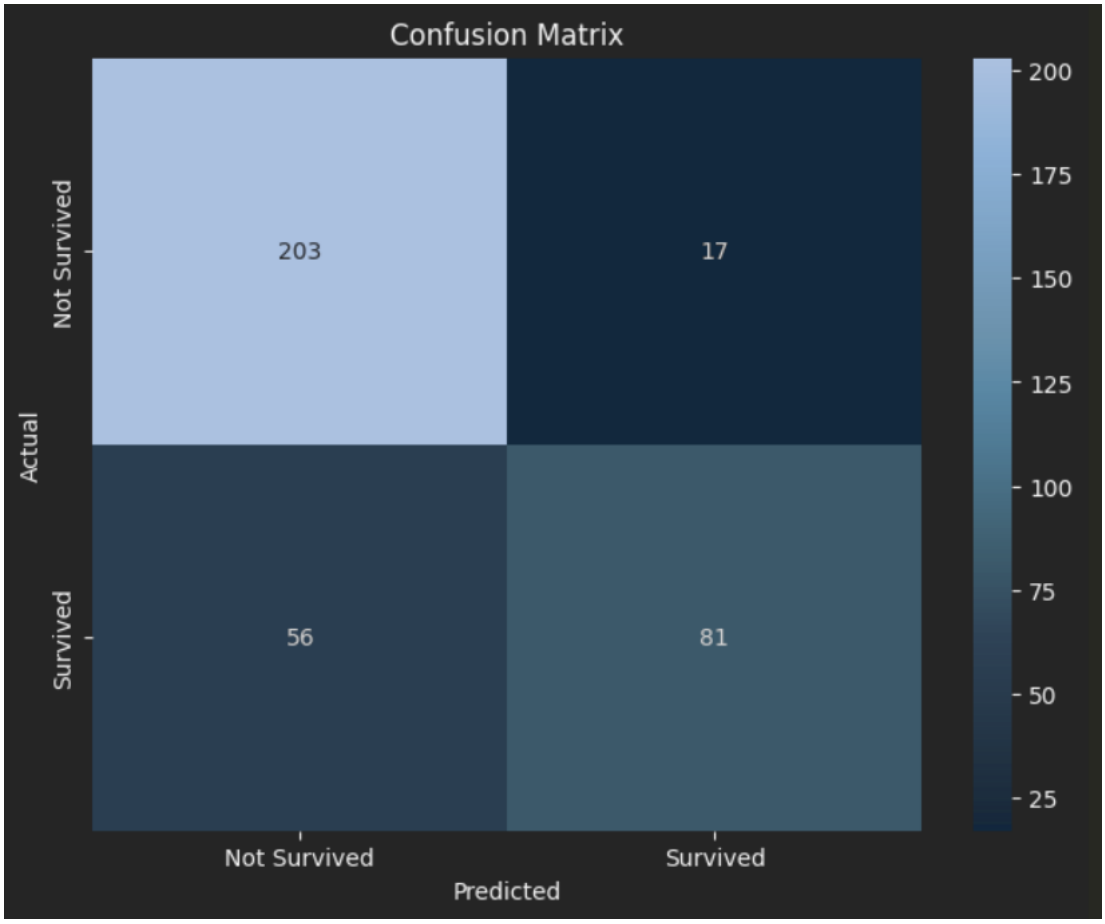
a. Ratio (40/60)



- Insights:

```
Evaluating Decision Tree for split 40/60:
Classification Report:
              precision    recall  f1-score   support

Not Survived       0.80      0.91      0.85       330
    Survived       0.82      0.63      0.71       205

    accuracy                           0.81       535
   macro avg       0.81      0.77      0.78       535
weighted avg       0.81      0.81      0.80       535


Confusion Matrix:
[[301  29]
 [ 75 130]]
```

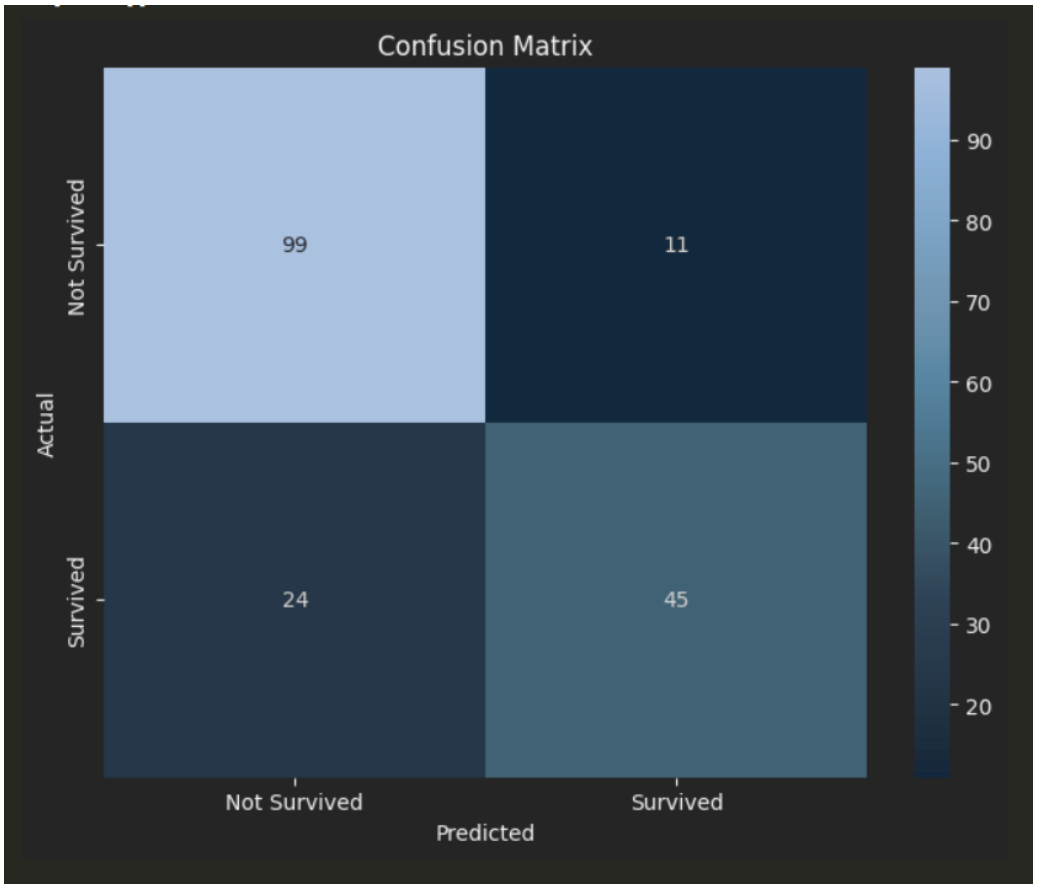
b. Ratio 60/40



- Insights:

```
Evaluating Decision Tree for split 60/40:
Classification Report:
              precision    recall  f1-score   support

Not Survived       0.78      0.92      0.85       220
    Survived       0.83      0.59      0.69       137

    accuracy                           0.80       357
   macro avg       0.81      0.76      0.77       357
weighted avg       0.80      0.80      0.79       357


Confusion Matrix:
[[203  17]
 [ 56  81]]
```

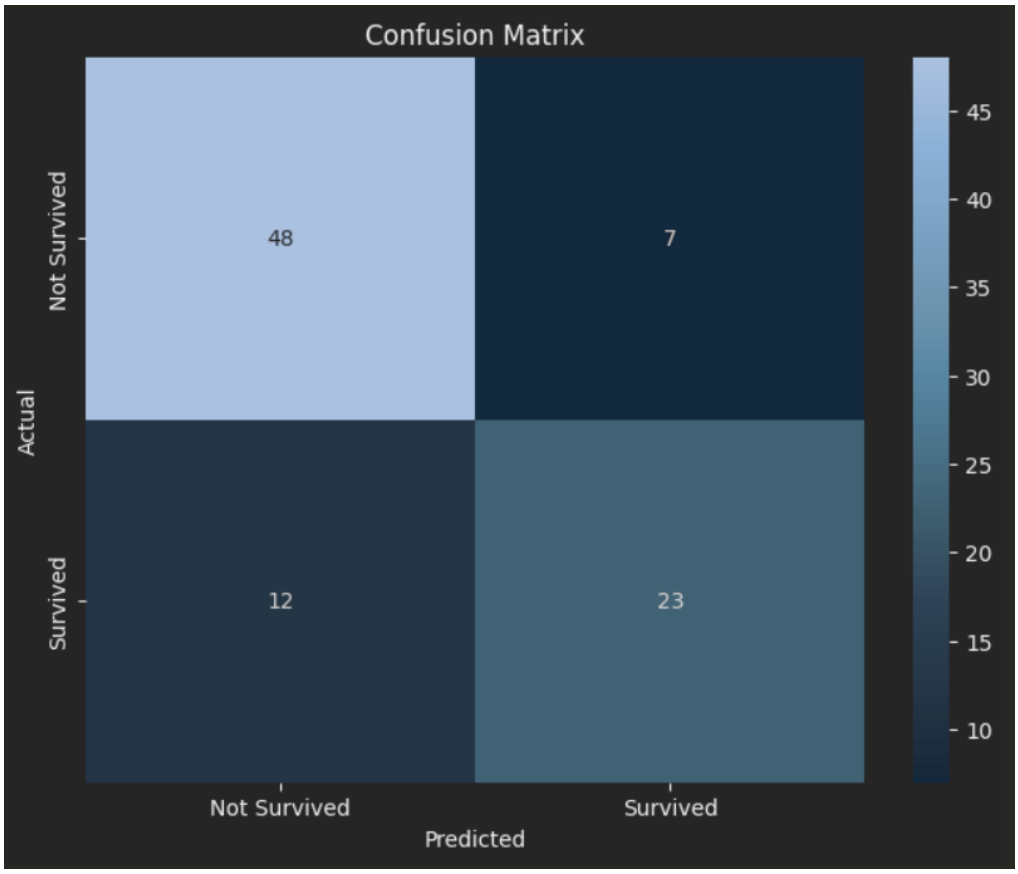
c. Ratio 80/20



- Insightst:

```
Evaluating Decision Tree for split 80/20:
Classification Report:
              precision    recall  f1-score   support

Not Survived       0.80      0.90      0.85       110
    Survived       0.80      0.65      0.72        69

    accuracy                           0.80       179
   macro avg       0.80      0.78      0.78       179
weighted avg       0.80      0.80      0.80       179


Confusion Matrix:
[[99 11]
 [24 45]]
```

c. Ratio 90/10

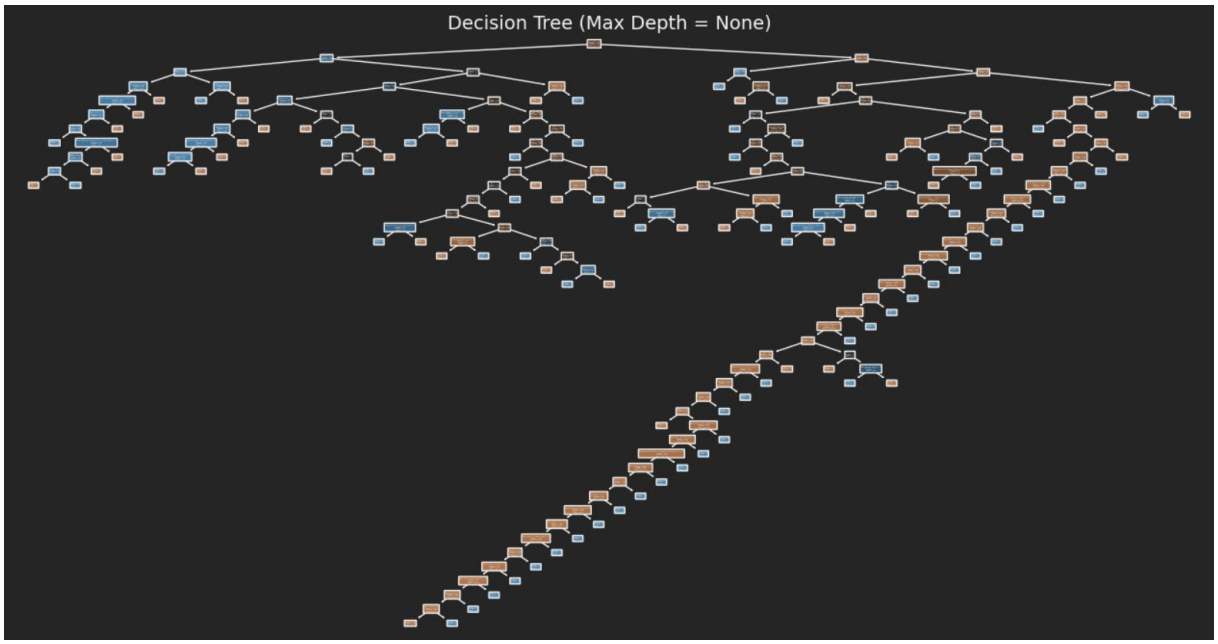


- Insights:

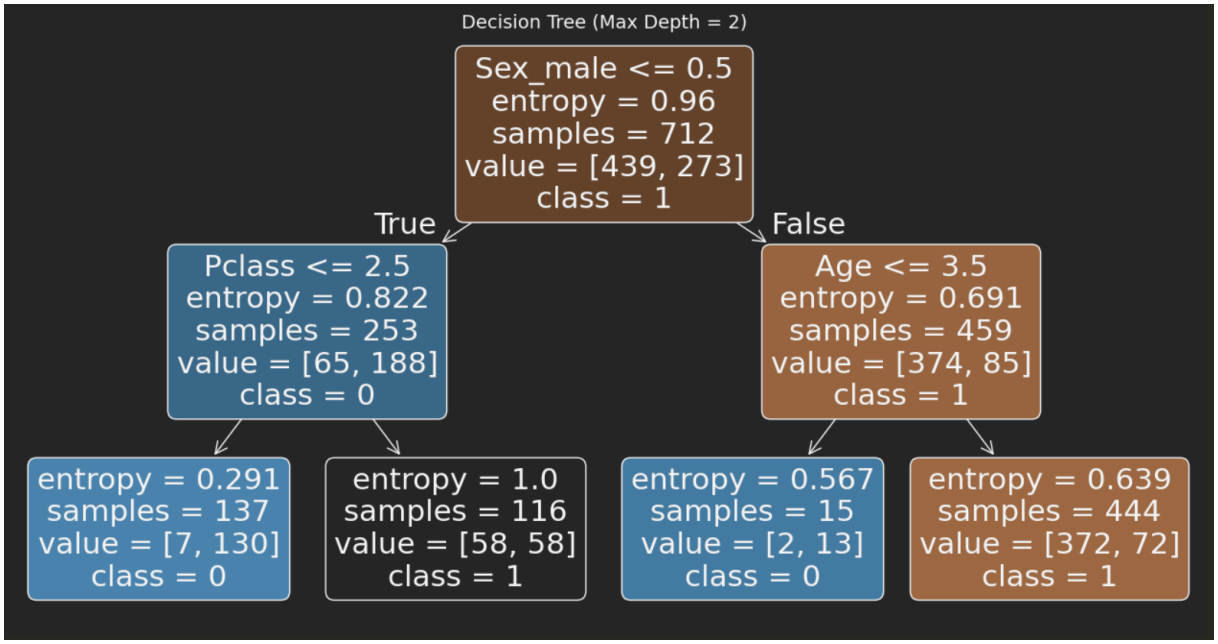
```
Evaluating Decision Tree for split 90/10:
Classification Report:
              precision    recall  f1-score   support

Not Survived       0.80      0.87      0.83        55
    Survived       0.77      0.66      0.71        35

    accuracy                           0.79        90
   macro avg       0.78      0.76      0.77        90
weighted avg       0.79      0.79      0.79        90


Confusion Matrix:
[[48  7]
 [12 23]]
```

2.4. Depth and accuracy of the decision tree.

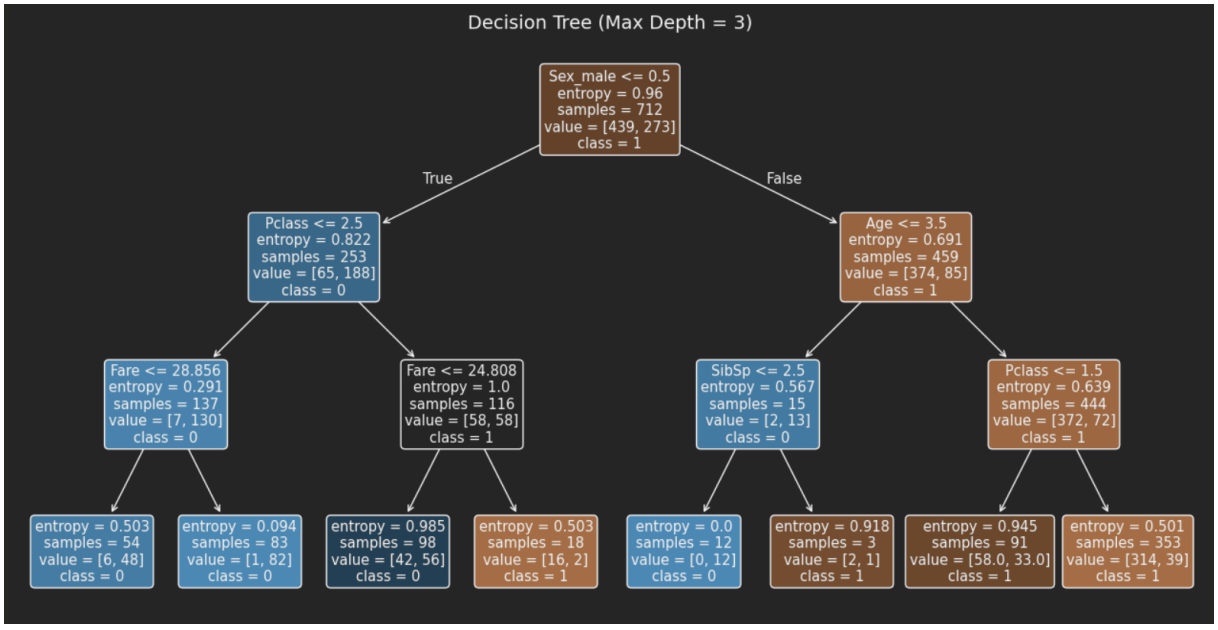
- Max deep = None


Decision Tree (Max Depth = None)

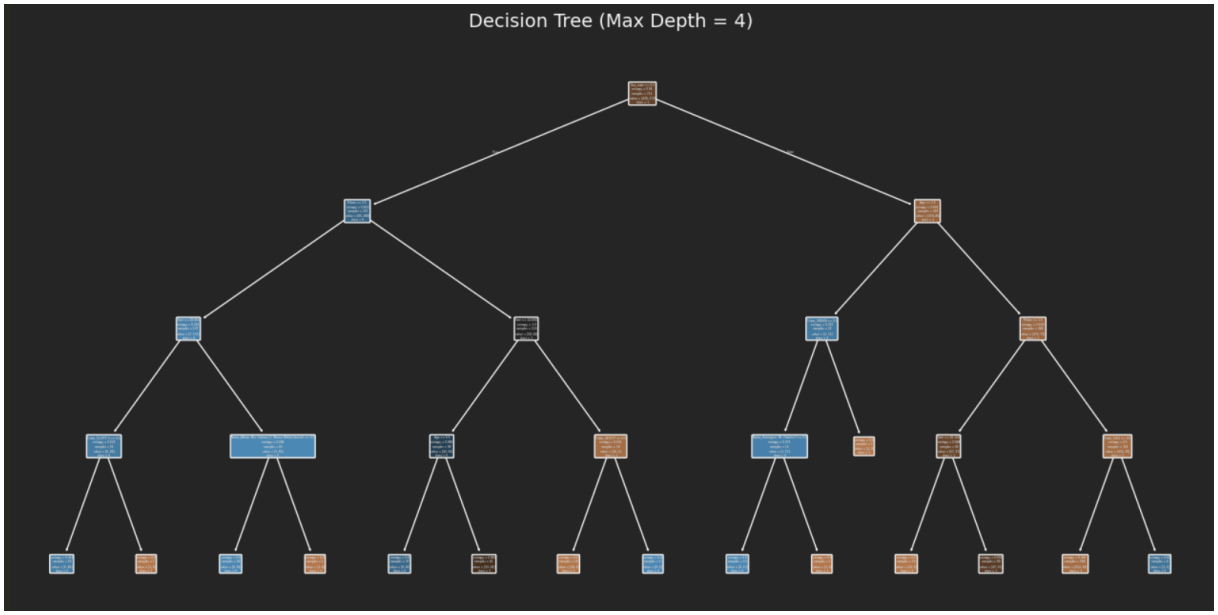
- Max deep = 2



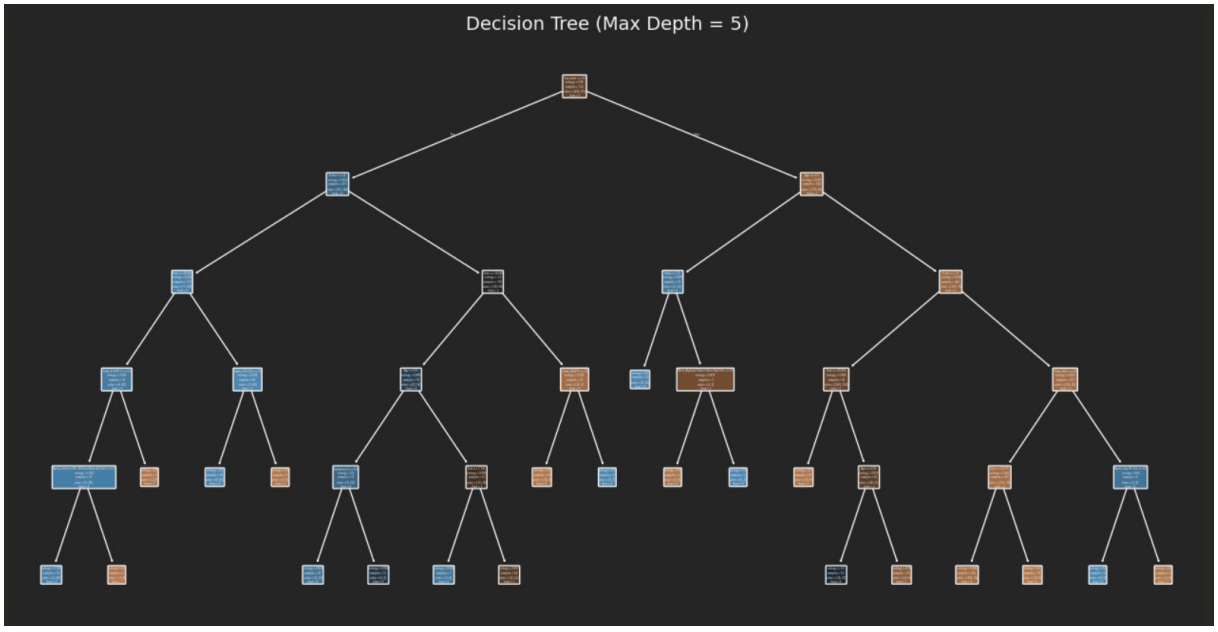
- Max deep = 3



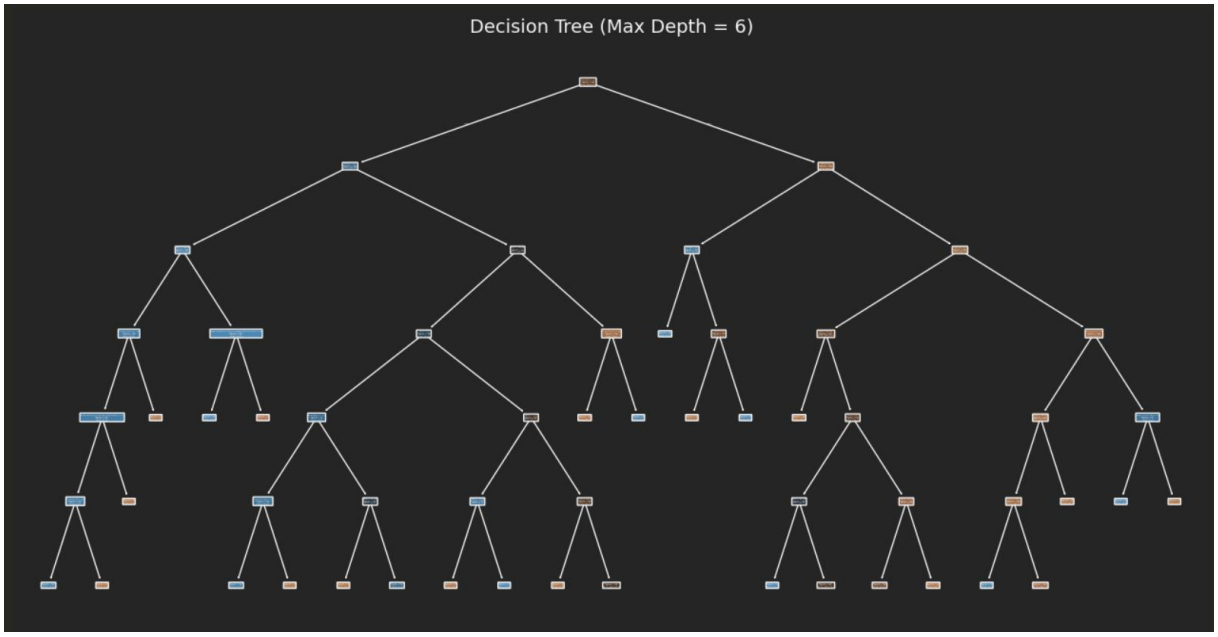
- Max deep = 4



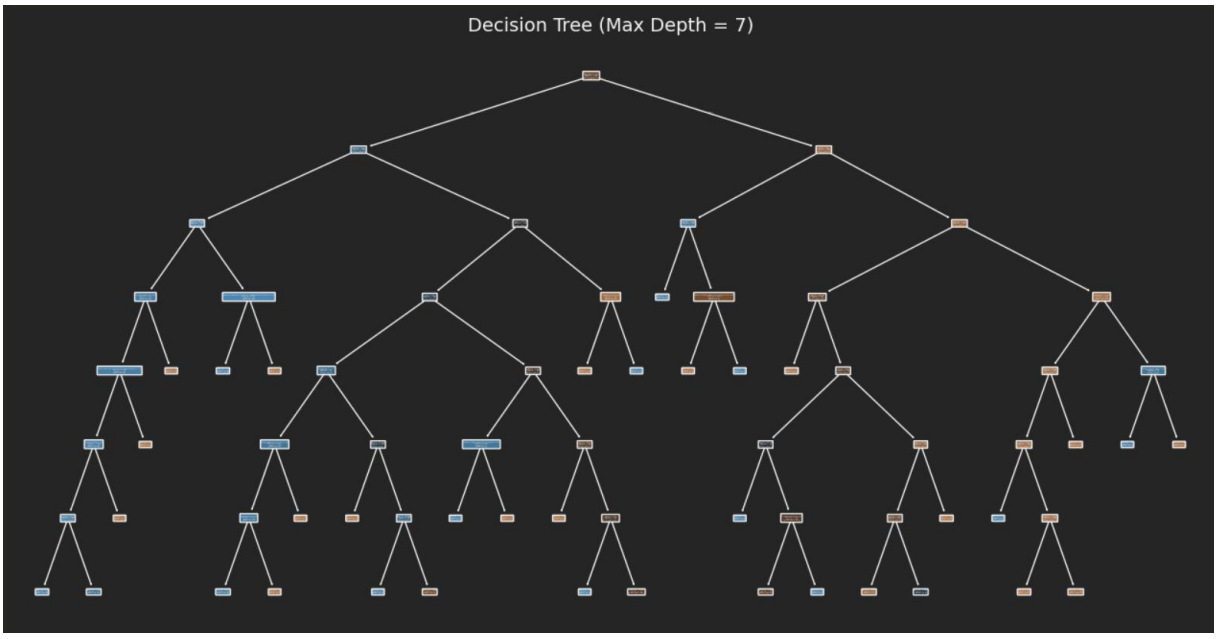
- Max deep = 5



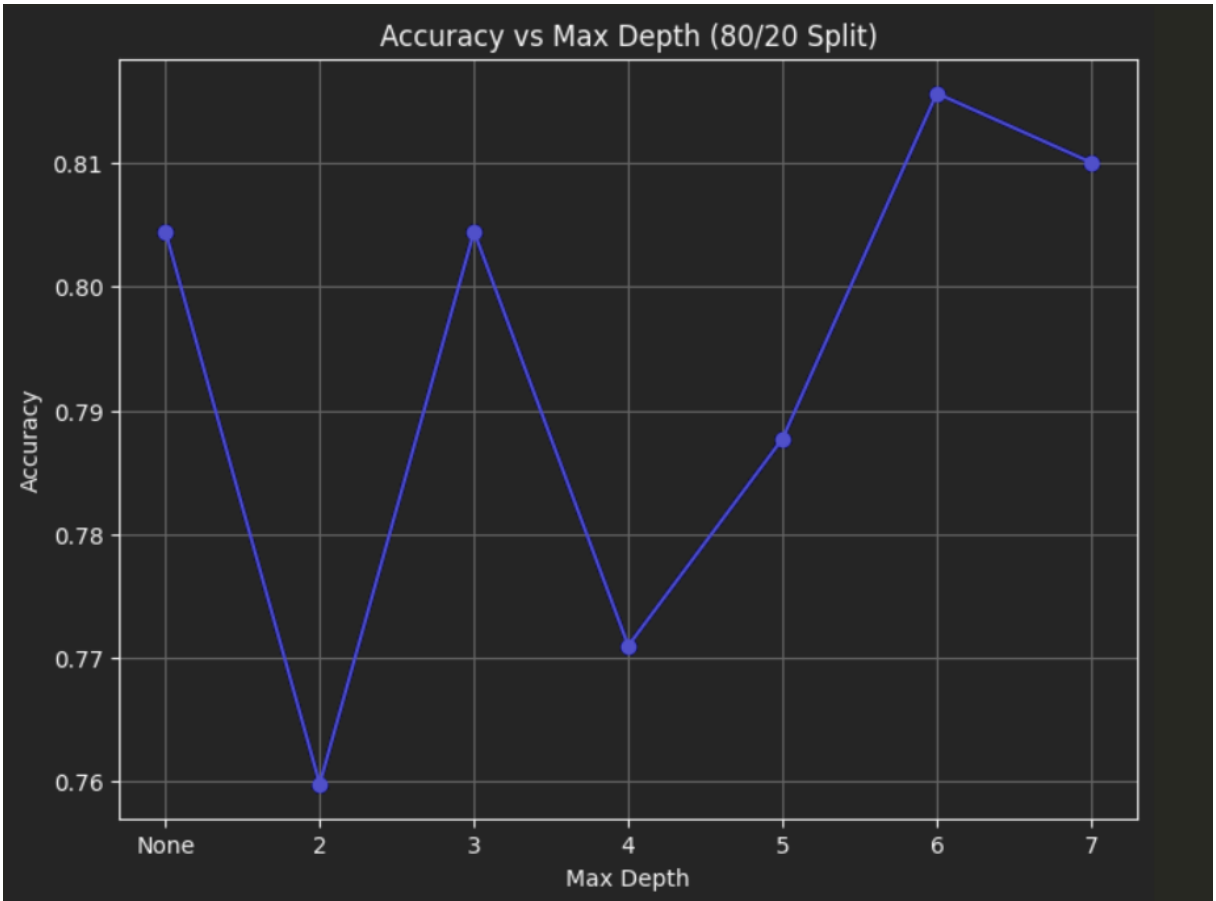
- Max deep = 6



- Max deep = 7



- Graph




```
### Accuracy for Different Max Depth Values (Horizontal Format):
              None         2         3        4         5         6         7
Accuracy  0.804469  0.759777  0.804469  0.77095  0.787709  0.815642  0.810056
```

3. Summarize the results.

```
Dataset Summary:
   Number of Classes  Number of Features  Sample Size  Accuracy
0                  2                1724          891  0.810056
```

IV. Compare dataset

1. Additional Dataset

- **Number of classes:** 2 (binary classification).
- **Number of features:** 1724 – Very high compared to other datasets.
- **Sample size:** 891 –Average.
- **Accuracy: 0.810056.**

- **Remarks:**
    - Although it has a large number of features, the accuracy is not as high as the Breast Cancer dataset. This may be because the model struggles to handle excessive redundant features (high dimensionality).
    - To improve the results, feature selection can be applied to reduce the number of features.

---

2. Breast Cancer

- **Number of classes:** 2 (binary classification).
- **Number of features:** 30 – Reasonable and moderate.
- **Sample size:** 569 – The smallest among the datasets.
- **Accuracy: 0.947368**.
- **Remarks:**
    - With a moderate number of features and sample size, the model performs efficiently and achieves the highest accuracy.
    - The data is clearer and less noisy compared to other datasets.
    - The Breast Cancer dataset demonstrates that high accuracy does not rely on large data size if the data is of high quality and has well-distinguishable features.

---

3. Wine Quality

- **Number of classes:** 3 (multi-class classification).
- **Number of features:** 11 – The lowest among the three datasets.
- **Sample size:** 4898 – The largest among the three datasets.
- **Accuracy: 0.787755**.
- **Remarks:**
    - A larger number of classes and a larger sample size increase the complexity of the problem.
    - Lower accuracy due to imbalanced data or a lack of clear distinction between classes.
    - Techniques such as class balancing or experimenting with more powerful models like Random Forest or Gradient Boosting may be needed.

Conclusion:

1. **Number of classes:** The more classes (multi-class classification), the lower the accuracy tends to be, as observed in the Wine Quality dataset.
2. **Number of features:** A large number of features (Additional Dataset) can lead to overfitting and reduced model performance if feature selection is not applied.
3. **Sample size:** A large sample size (Wine Quality) helps reduce overfitting, but the complexity of the problem increases with a higher number of classes.
4. **Model performance:**
    - The Breast Cancer dataset achieves the highest accuracy thanks to well-distinguishable features and a moderate data size.
    - The Additional Dataset can be improved by reducing redundant features.
    - The Wine Quality dataset may require additional preprocessing techniques to improve accuracy.

Recommendations for improvement:

- **Feature Selection**: Reduce the number of features for datasets with many columns, such as the Additional Dataset.
- **Class Balancing**: Balance the data for the Wine Quality dataset.
- **Alternative Models**: Experiment with models such as Random Forest, XGBoost, or Gradient Boosting to improve classification performance.