SWINBURNE UNIVERSITY OF TECHNOLOGY

Title	Abalone Dataset Classification
Name	Sameer Tariq
Student ID	102880385
Contact Detail	10288085@student.swin.edu.au
Date	5/20/2022

Abstract

In this report, we are going to discuss how we find, acquire and load the dataset. After the appropriate dataset, the process comes for the data exploration, in this part we see the interdependency of different features and their impact on the target column. The next step comes with data modeling in which we created different chunks of the data for testing and training purposes at the end we applied two classification algorithms were used to predict the target feature and their outcomes and different aspect to measure their strength.

Introduction

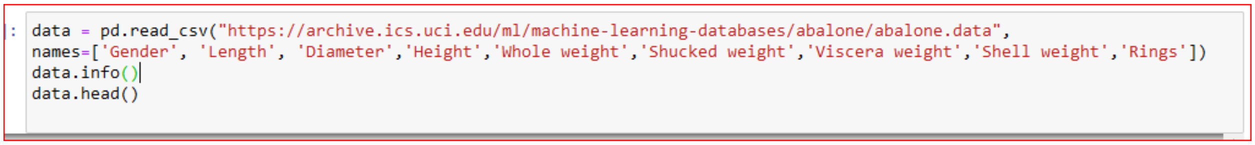
The dataset chosen for the assignment is Abalone. Abalone is a common name for any of a group of small to very large sea snails, marine gastropod mollusks in the family Haliotidae. The flesh of abalones is widely considered to be a desirable food and is consumed raw or cooked by a variety of cultures.

The abalone dataset has all the required needs as required in task 1 we are going to talk about it in detail further. After loading the dataset from UCI, the exploration process started did my best to see every aspect of the data and its independence of columns from one another. In the process of exploration used both visuals and statistics to cover the whole picture.

After the exploration, the toughest part is to determine the data modeling process because you have a vast number of options to choose from one of the main issues is you did not know the outcome of a particular step until the end. In this process, we use min-max scaler to transform the features because the data is pretty vast and then we transform the target column which is in categorical shape to number values to train the model. Split the data into training and testing chunks to train models. The models we have chosen for classification are Logistic regression and RandomForestClassifier in the end we use different measuring methods to evaluate the performance of the Classification models.

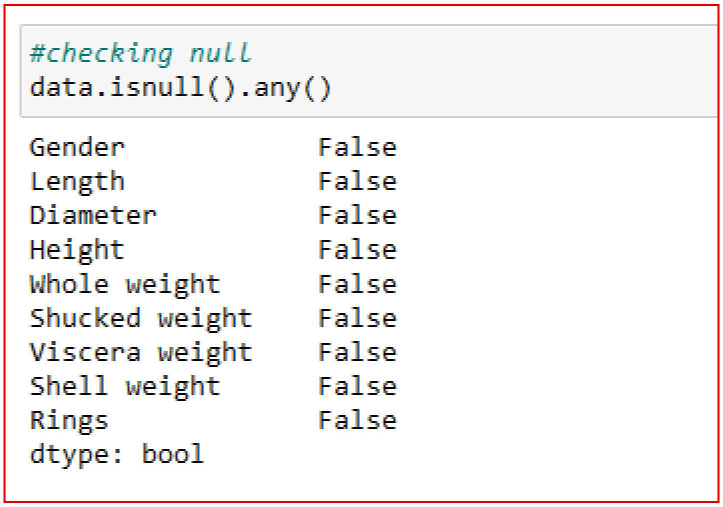
Task 1 – Data Acquisition and Preparation

After going through countless datasets even working on a few of them ended on the dataset Abalone, abalone is used to call the sea snails this dataset gather in 1995 by "The population Biology of Abalone" in Tasmania. This dataset consists of 8 columns and 4177 rows with no missing value and no multi-label. According to my knowledge, this dataset completely matches with dataset requirements. Data is loaded directly from the UCP data center archive and add the column's name respectively.



```
]: data = pd.read_csv("https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data",
   names=['Gender', 'Length', 'Diameter','Height','Whole weight','Shucked weight','Viscera weight','Shell weight','Rings'])
   data.info()
   data.head()
```

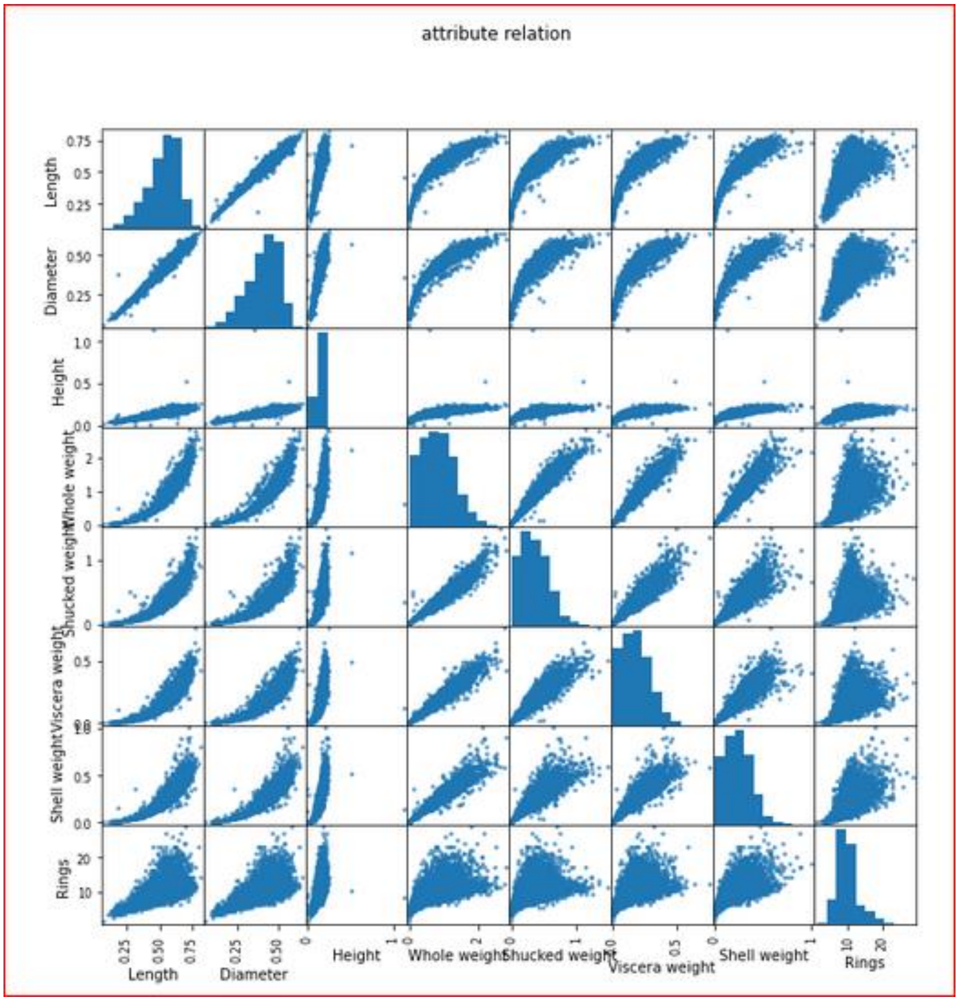
After loading the data get through the data by using method **info** and **head** to double-check the data is according to the given instructions at UCP after this checks the null values if there are any.



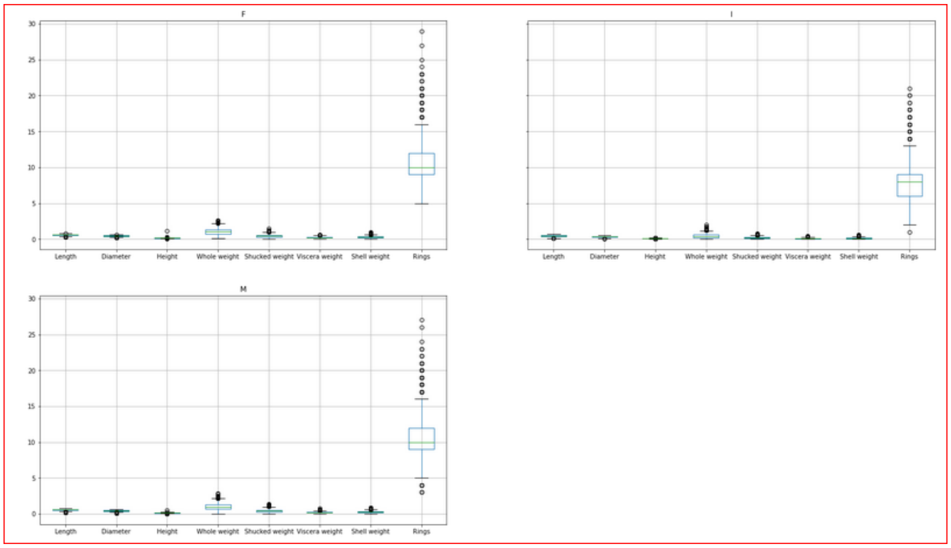
```
#checking null
data.isnull().any()

Gender            False
Length            False
Diameter          False
Height            False
Whole weight      False
Shucked weight    False
Viscera weight    False
Shell weight      False
Rings             False
dtype: bool
```

Task 2 – Data Exploration

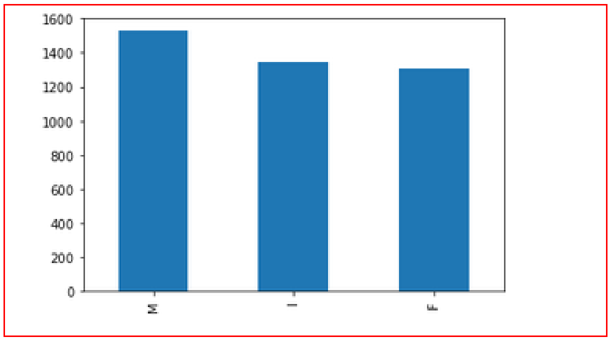
In the next process we need to go through our data in a detailed manner to be more precise about the actual meaning and its interdependency on one another for this we use both statically and visualization techniques to get into the depth of the picture. First with the help of **describe**d method get the mean, max, etc. of all the data. After that use the method **info** to make sure the data is already in the required data types. Most of the focus in this dataset is on column name **Gender** because that's the target for classification going to talk about it in further steps in the report. Create a group to make sure the data is balanced and among all fields. The second tool used to explore is the visualization first created scatter matrix to see the data relation among itself



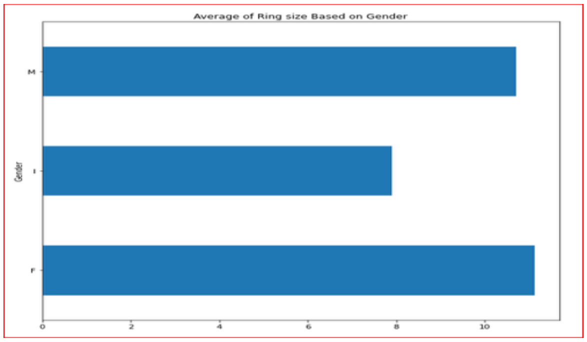
As our focus is on the gender we need to see the detail regarding it in the data created a boxplot to determine each column related to our each gender



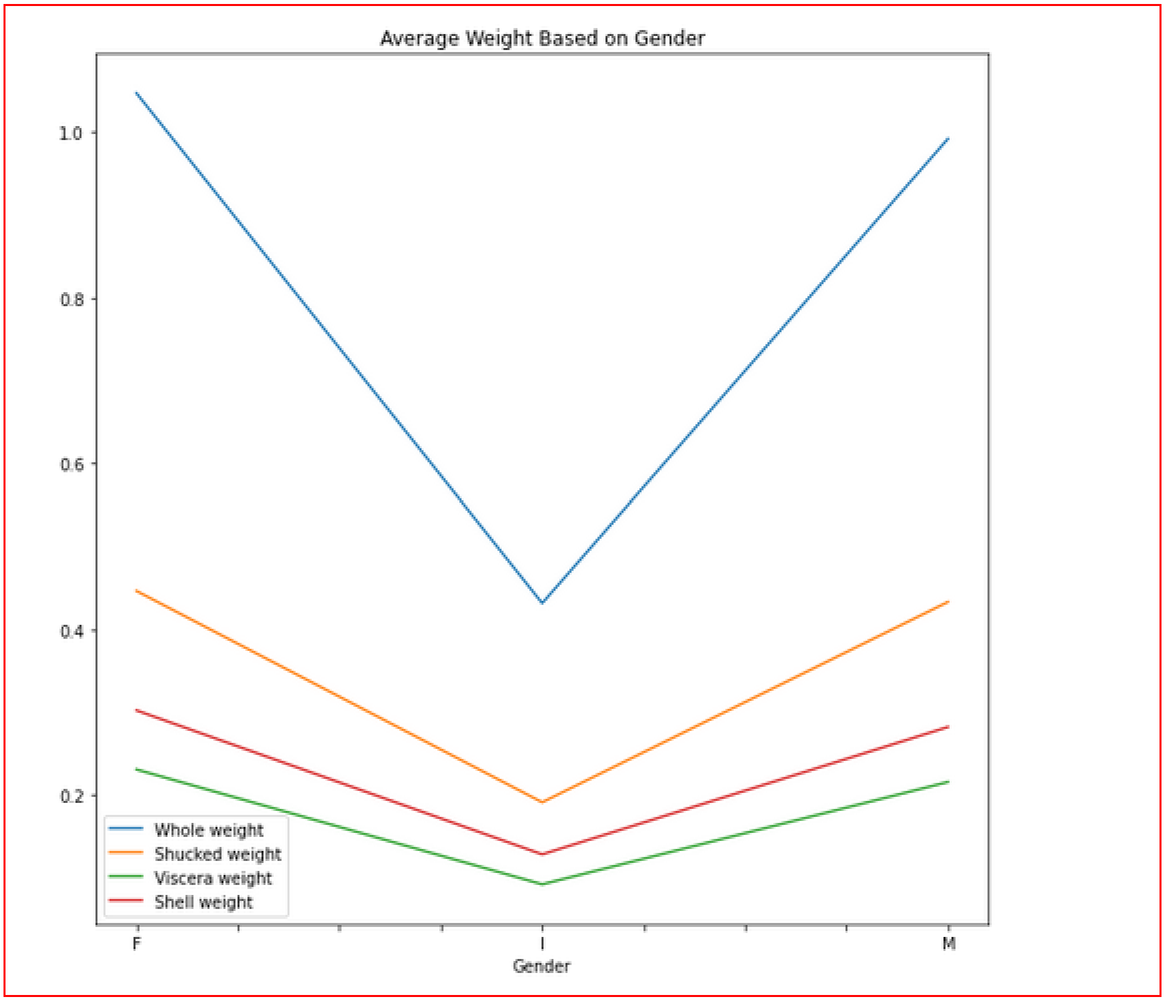
The above graph explains that Infant has all values smaller compared to the male and female. In the same way, created different questions to explain the data for the target like the total counts of gender for each gender. It shows equal distribution which is great for data classification.



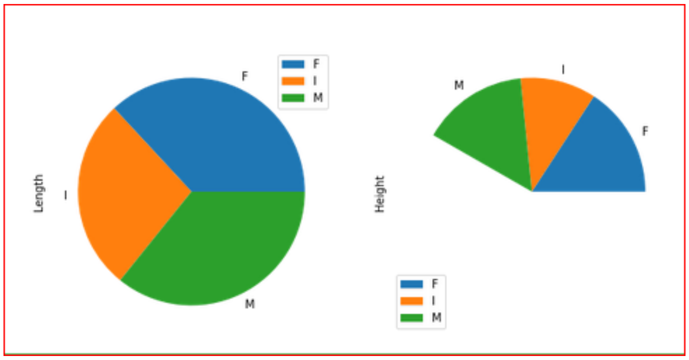
In the above boxplot, we find the ring size is the value that is greater than other values to further inspect it, creating a graph that explains the average size of the ring for each gender.



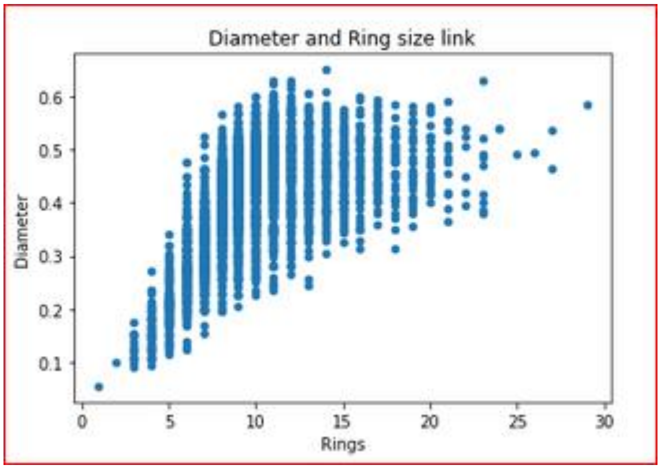
In the dataset, there are different weights as a whole weight is a whole abalone, and shucked weight is the weight of meat, etc. The below chart shows the mean weight for every gender



The next features which are related to mostly each other are height and length. The below graph shows the average height and length regarding gender



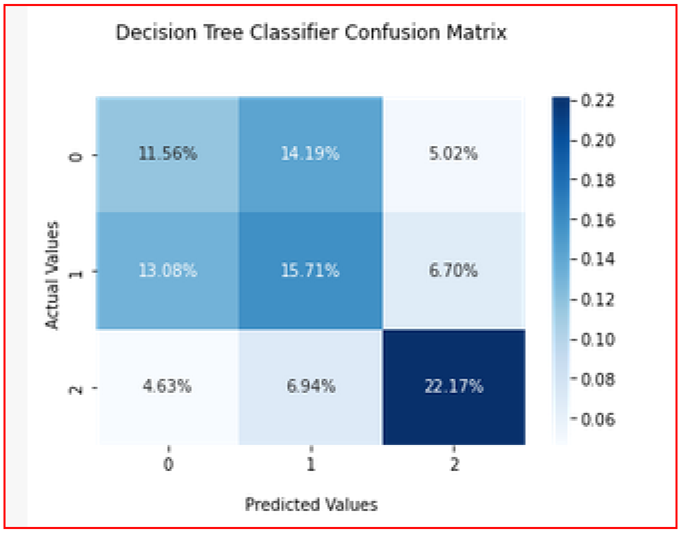
Rings and diameter are related to each other, creating a scatter plot to describe how they link between each other.



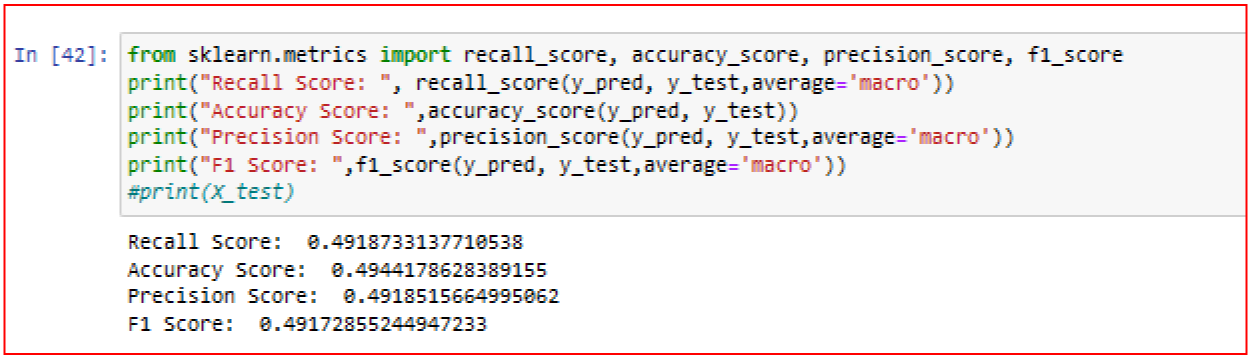
Task 3 – Data Modeling

Before data modeling, there are a couple of changes that have to be made to make the dataset the target column is in categorical variable to numeric by replacing it with 0, 1, 2 with F, M, I. After that dataset is converted into X, y. X consists of the data set and y consists of the target column. Data is split into seventy percent for the training and thirty percent for the testing because we have enough data to train and a good amount of 1254 instances to evaluate the performance of the model. The data is largely distributed to make it more relevant to each other applied a min-max scaler with a standard deviation of 1 and mean of 0. **Decision Tree**

Classifier is the first model chosen to train the data on it because it's easy to implement, work great on more than two outputs like in my case, and is easy to explain also but it might take some more time to train. After training the data model is tested on the testing chunk of the data result shows an accuracy of 0.49, which is quite low. Besides accuracy confusion matrix is the tool to see the output of your trained model which shows that data is quite good in picking Infant gender other than male and female. It means the data is quite similar for male and female which create confusion for the model.

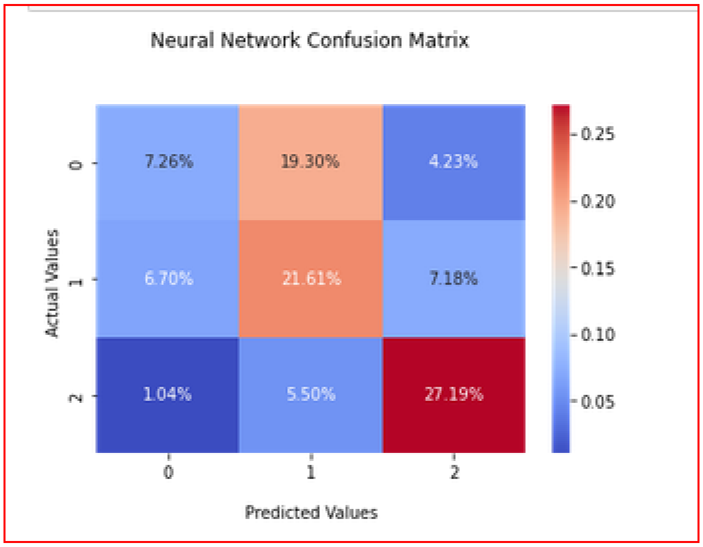


Other than that most of the performance measuring matrices have the same values near 0.49.

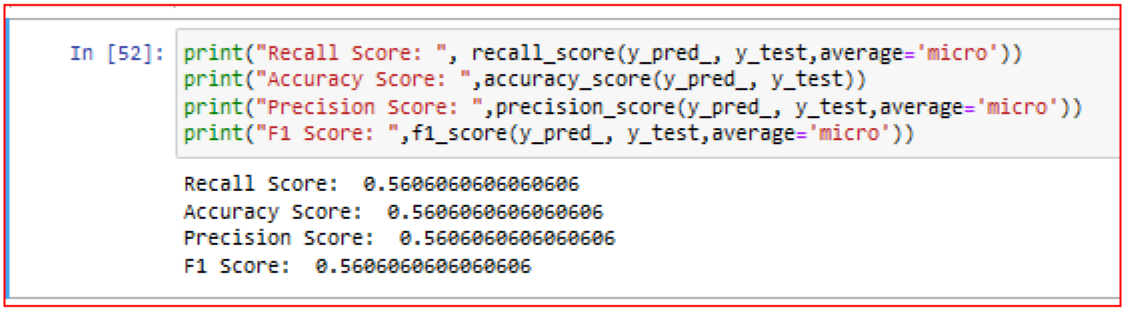


```
In [42]: from sklearn.metrics import recall_score, accuracy_score, precision_score, f1_score
         print("Recall Score: ", recall_score(y_pred, y_test,average='macro'))
         print("Accuracy Score: ",accuracy_score(y_pred, y_test))
         print("Precision Score: ",precision_score(y_pred, y_test,average='macro'))
         print("F1 Score: ",f1_score(y_pred, y_test,average='macro'))
         #print(X_test)

         Recall Score:  0.4918733137710538
         Accuracy Score:  0.4944178628389155
         Precision Score:  0.4918515664995062
         F1 Score:  0.49172855244947233
```

The second model chosen for the model is **Neural Network.** The reason for choosing this model is because it works great with a large number of features like we have around eight and it's completely different from the previous chosen model so it's interesting to see its result in the end. After training the data model is tested on the testing chunk of the data result shows an accuracy of 0.56, which good than the previous model. If we talk about the confusion matrix it pretty much shows the same behavior as the previous one but it perform well in catching Male(1) and Infants(2) but did worse on Female(0).

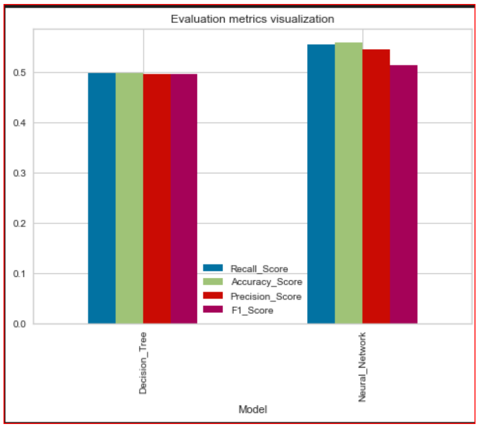


Other than that most of the performance measuring matrices have the same values near 0.56.

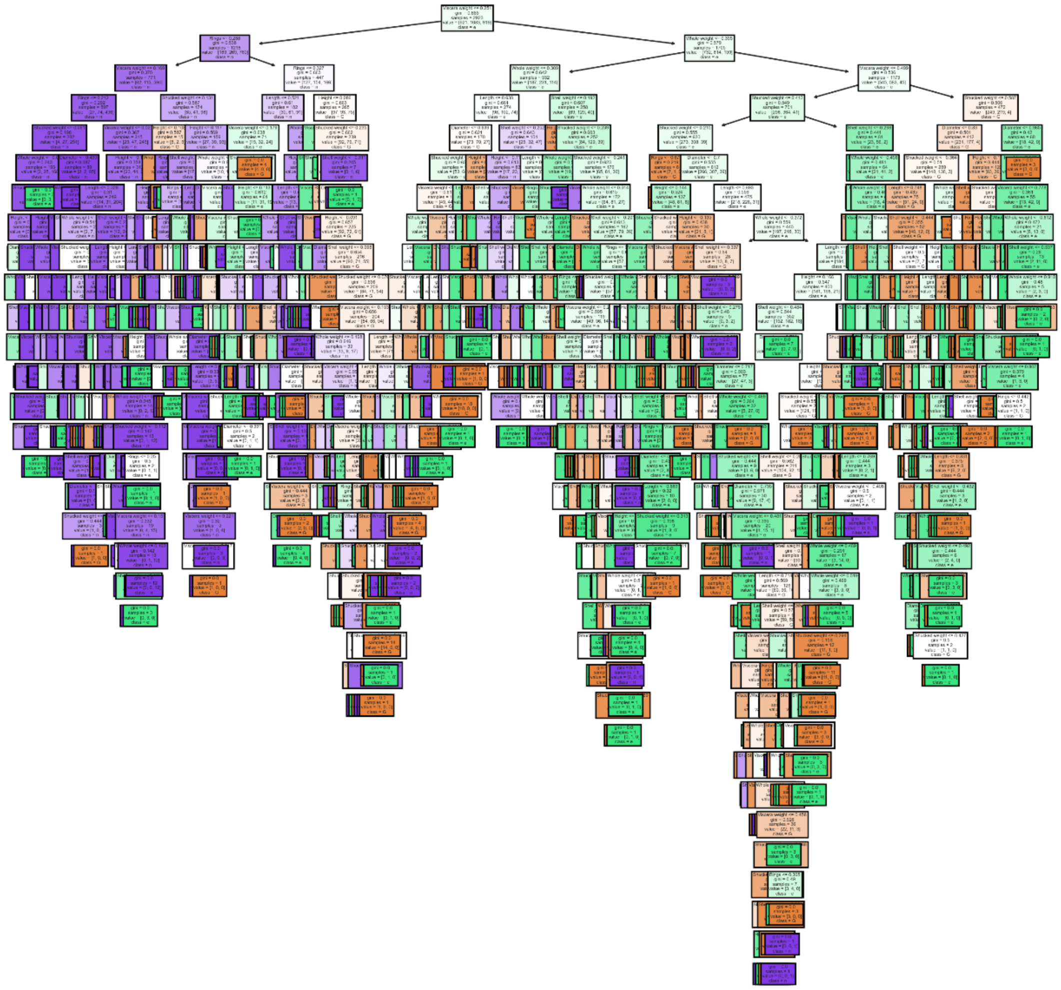


```
In [52]: print("Recall Score: ", recall_score(y_pred_, y_test,average='micro'))
         print("Accuracy Score: ",accuracy_score(y_pred_, y_test))
         print("Precision Score: ",precision_score(y_pred_, y_test,average='micro'))
         print("F1 Score: ",f1_score(y_pred_, y_test,average='micro'))

         Recall Score:  0.5606060606060606
         Accuracy Score:  0.5606060606060606
         Precision Score:  0.5606060606060606
         F1 Score:  0.5606060606060606
```

If we compare both models' visuals then there is more or less pretty much the same in performance but neural networks have some more good results.



In the research for the models come across a visualization of a decision tree which shows the whole tree structure which is made by the decision tree on our data it’s hard to grasp all the information in it but still, it's amazing visualization.



Conclusion:

In the conclusion, both models have a low score because of pretty much the same data for both gender Male and Females. If we compare the performance from both of the applied models, neural networks have performed well than the decision tree.