REPORT

-Princy Gautam(B20BB051)

Task 1: Decision Tree Classifier

Method

Import necessary libraries for performing mathematical calculations and reading csv file. Also, import encoder to encode categorical data to numerical.

Read the file using pandas and assign it a name df

The features for the given dataset are: Age, Sex, BP, Cholesterol, Na to K columns and the target is the Drug column.

Using isnull (). any () we check for the missing values in the data.

Age: numerical

Sex: nominal

BP: ordinal

Cholesterol: ordinal

Na to K: numerical

Drug: categorical

Assign the name data to the dataset and read it. Then using label encoder, fit and transform the data which are non-numerical. Drop the earlier columns from the data and print new numerical dataset.

Then for lab part (2), we encode all the columns except the output column. Split the dataset into train and test sets in the ratio of 80:20 with reproducibility = 55 throughout the task. Train and fit the model using decision tree classifier with criterion as entropy. Then measure accuracy by predicting y and print it.

Split the dataset into 70:30 and 90:10 ratio of train and test sets. And then, train and fit the data using decision tree classifier with criterion as gini. Again, check the accuracy of each model.

Print confusion matrix and classification report for all three models.

Print the graphical visualization of all three trees using graphviz.

Where, model1 – 80:20 ratio one

Model2 – 70:30 ratio one

Model3 – 90:10 ratio one

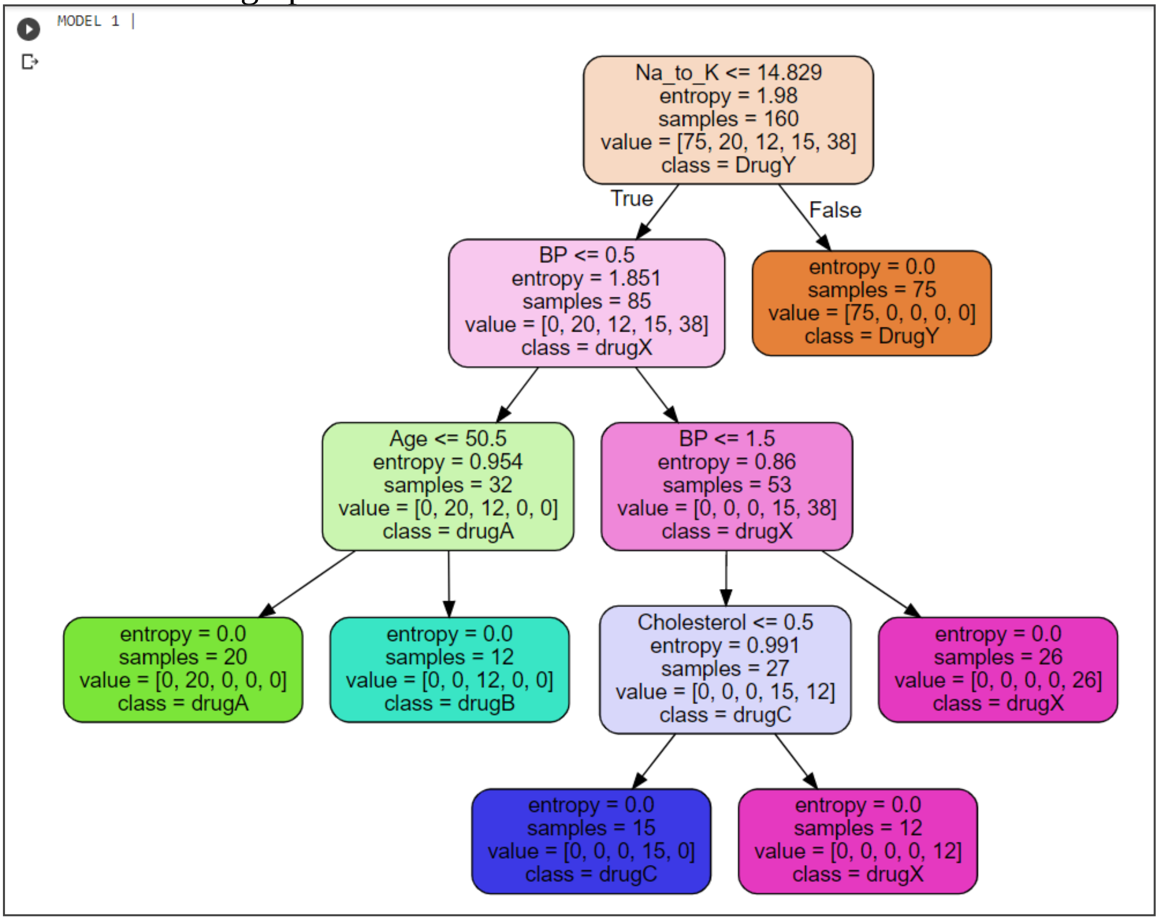
Results & Observations

The features for the given dataset are: Age, Sex, BP, Cholesterol, Na to K columns and the target is the Drug column.

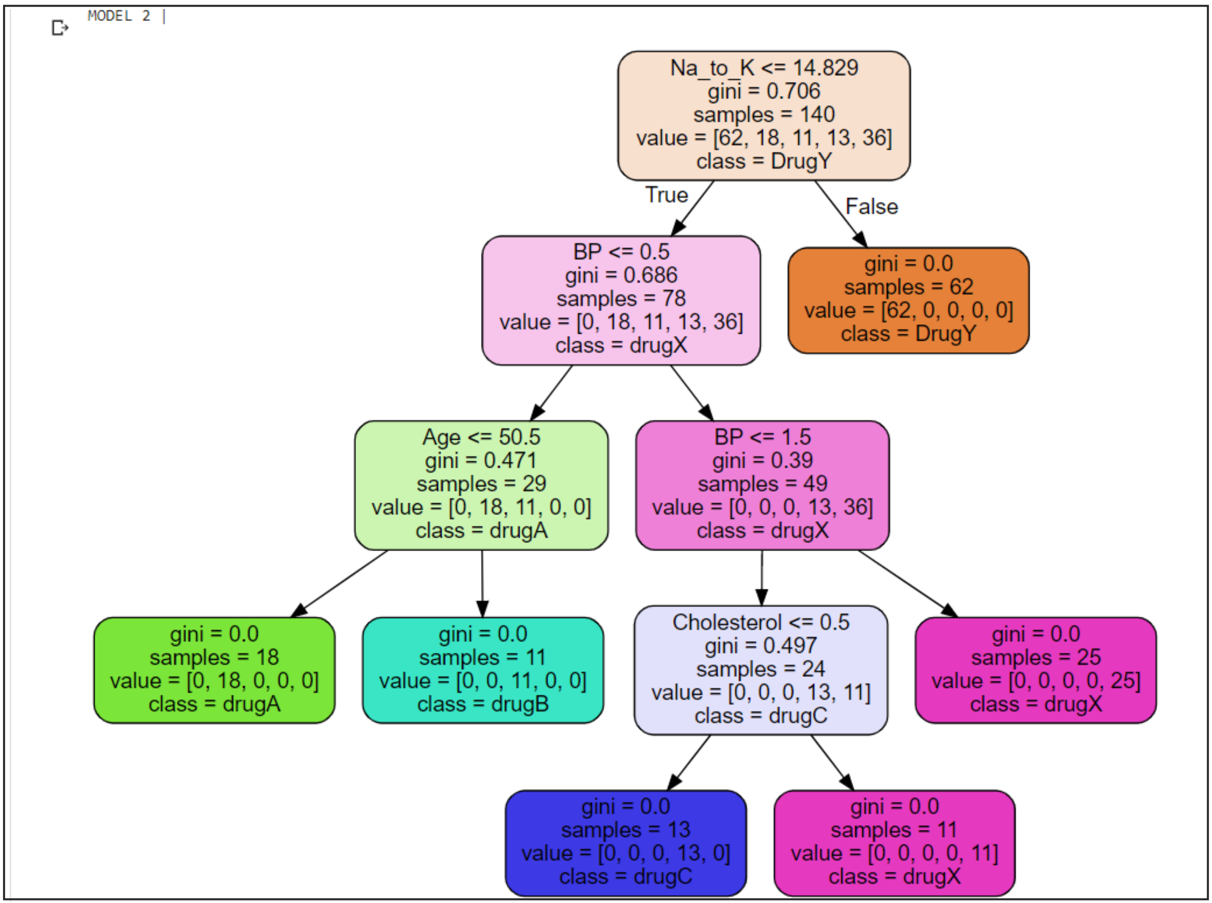
No missing values found.

We got accuracy equal to 1. In multilabel classification, the function returns the subset accuracy. If the entire set of predicted labels for a sample strictly matches with the true set of labels, then the subset accuracy is 1.0, otherwise it is 0.0

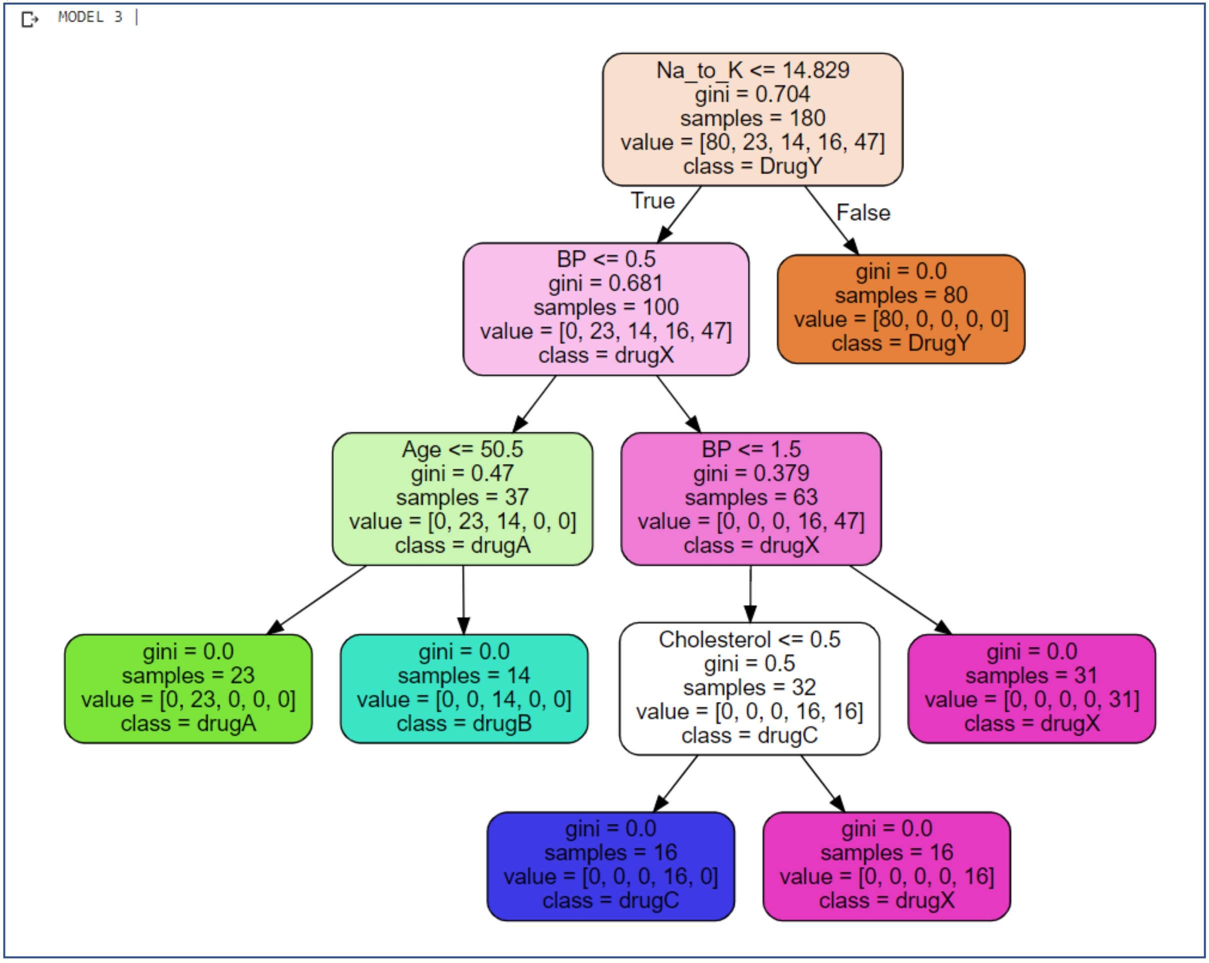
Since our accuracy is consistent thus there is no over fitting or under fitting in the model. The graphical visualization of the models is shown below:



```
MODEL 1
```



```
MODEL 2 |
```



```
MODEL 3 |
```

Task 2: Decision Tree Regressor

Method

Import necessary libraries for performing mathematical calculations and reading the file. Read the data using pandas library and name it as df.

The features of the dataset are: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age columns and the target is Concrete compressive strength.

Define X and Y as X = df.iloc [:, :-1] and Y = df.iloc [: ,-1].

Scale and fit the data.

Split the dataset into train and test set in the ratio of 70:30 with reproducibility set as 2021 throughout the task.

Train the decision tree regressor and with evaluation metrics such as MSE and node selection strategy to be set as random.

Report the accuracy of the model and MSE. And MAE.

Plot the graphical visualization of the decision tree.

Results & Observations

The features of the dataset are: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, Age columns and the target is Concrete compressive strength.

We get the value of MSE and MAE as follows:

MSE: 51.8, the lower the MSE the higher the accuracy of predictions as there would be excellent match between the actual and the predicted data set.

MAE: 4.7, a low MAE is better

We obtain an accuracy of 0.80

Performance on training data using decision tree: 0.99

Performance of testing data using decision tree: 0.80

There is an overfitting in the model as the dataset is performing 99% on training set and accuracy on test set drops.

We cannot print the confusion matrix and classification report of this model since it is not classifier, it is regressor.

Task 1:
https://colab.research.google.com/drive/1Eed0mdtF2GqMyiEwQ_i1asoPGJWprGI#scrollTo=aCH9Lnu9qCeZ

Task 2:
https://colab.research.google.com/drive/1UyBg6HAGDNcrUp37_bV6a97sbiF2fXdP#scrollTo=E0DbixPNRLL2