Assignment 3

Student ID: 300318105 **Student Name: Snehal Sudhir Bhole**

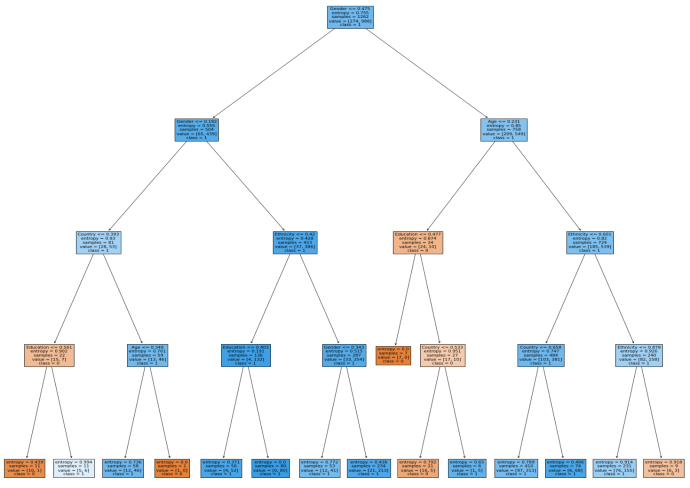
Report:

For this assignment, I explored all the Decision Tree Models trained in Assignments 1 and 2. The three models considered were for the drug 'Cannabis'. The first model is generated with the holdout method, the second model with oversampled data and the third model is generated with under-sampled data.

Model	Accuracy
Model 1 (Hold-Out Method on Original dataset)	0.7752808988764045
Model 2: (Oversampled dataset)	0.6208347405155916
Model 3: (Under sampled dataset)	0.7274456827648317

Thus, the model used for this assignment is the model generated in Assignment 1 using the hold-out method.

1. Display/visualized the resultant model created by the decision tree. [20 marks]



2. Explain how, and why, the algorithm made a specific decision. [10 marks]

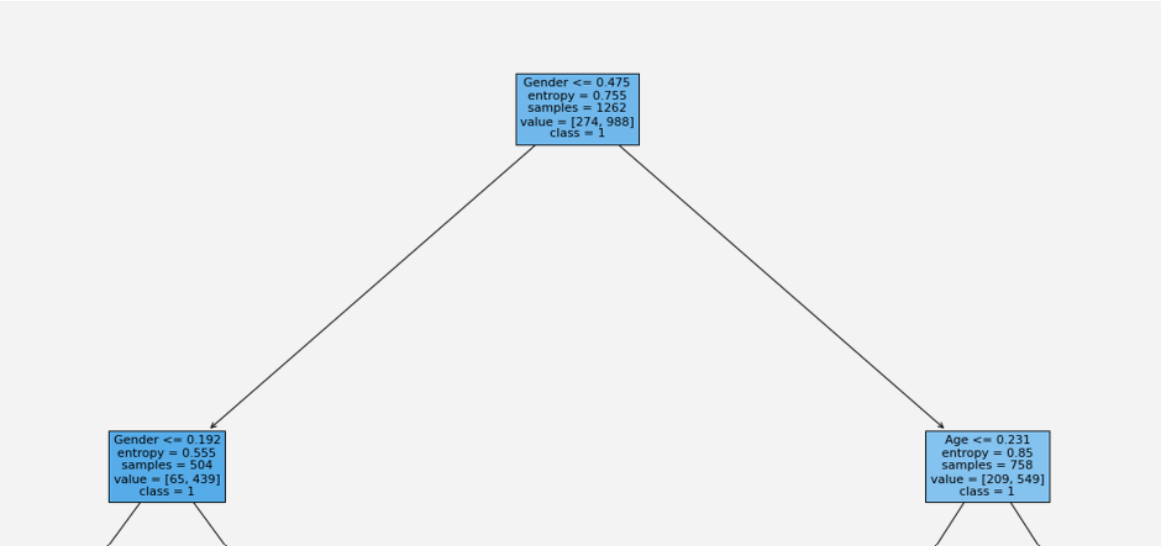
During feature selection, the following input features are selected as important features: 'Ethnicity', 'Education', 'Extraversion', 'Neuroticism', and 'Agreeableness'. The algorithm is trained with the following parameters:

max_depth = 4, criterion = Entropy, random_state = 0

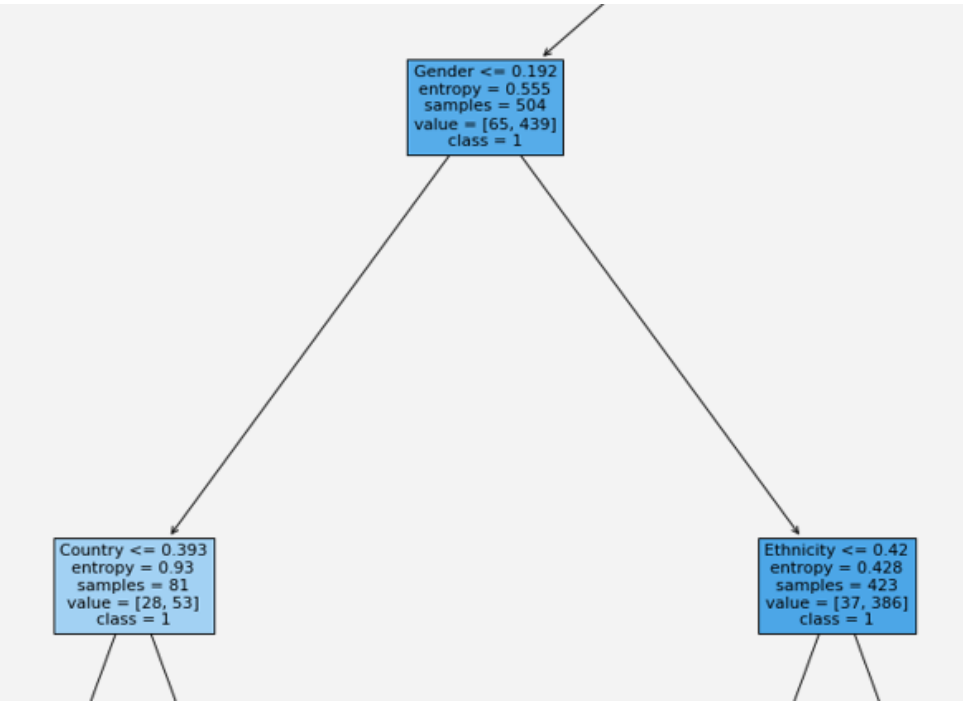
Decision Tree algorithm optimizes the cost function and learns from the dataset. For classification problems, the cost or loss function is the measure of impurity in the leaf nodes of a root node. Impurity in data is nothing but the uncertainty available in the dataset. Sometimes impurity is referred to as heterogeneity as it is a measure of a mixture of different classes at a particular node. The goal of the Decision tree is to minimize impurity. There are two metrics used: entropy and the Gini index. Entropy measures the amount of information present in the variable, i.e., the more the value of entropy, the variable is more important. The entropy value ranges from 0 to 1.

Entropy can be calculated using the formula:

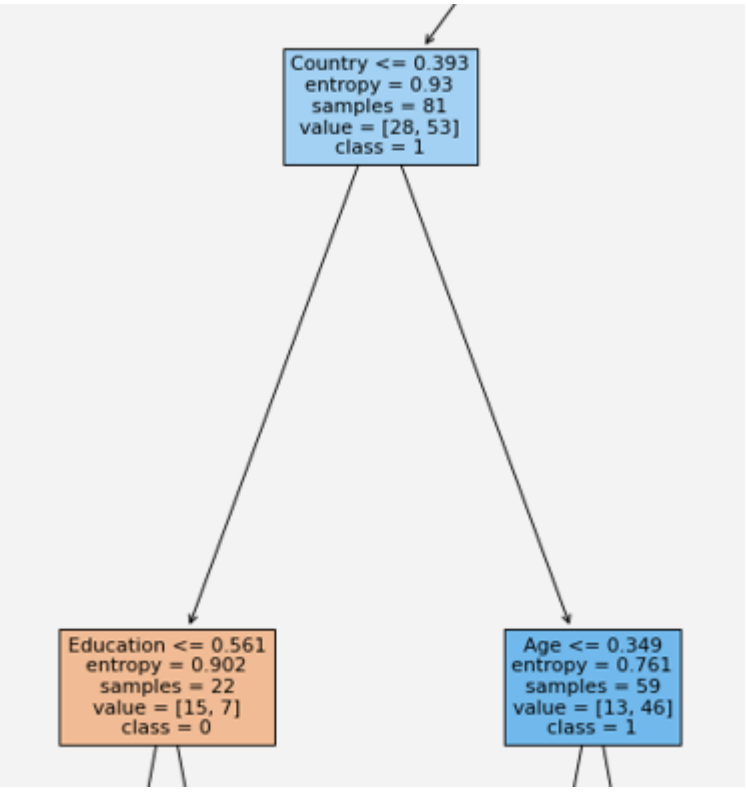
$$\mathbf{H(X) = - \Sigma (p_i * \log_2 p_i)}$$



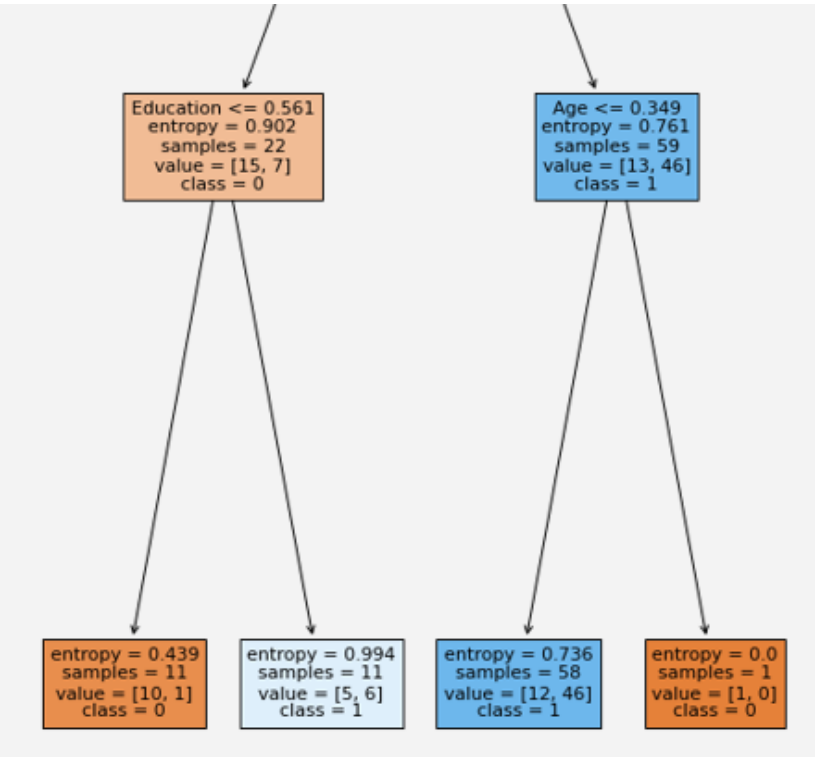
At the root node, the entropy value for the variable Gender is 0.755 which is the maximum for all the inputs, thus the tree is split on variable gender. 1262 training samples have a User: Non-User ratio of 274:988. The tree is split in the ratio of 504:758 samples based on gender value with "User" as the majority class, the right-hand and left-hand side representing values lesser than equal to 0.475 and greater than 0.475 respectively.



In the second split, 504 samples are split in the ratio of 81:423 records based on gender value less than or equal to 0.192 with entropy 0.555 which is calculated by considering all the samples with gender value lesser or equal to 0.475. Similarly, the next split is on the input feature Country which divides sample space into the ratio of 22:59 samples for country value <=0.393.



The next right-side split is on Education<=0.561 with an entropy of 0.902, where “non-User” is the majority class with 15 samples.



As the maximum depth for the Decision Tree is 4, it does not perform further split with entropy value = 0.439 representing pure leaf node and classifies 10 samples with education value <= 0.561 as “non-User” and 6 samples with education >0.561 as “User” with entropy 0.994.

In a similar manner, the decision tree learns and develops the path to leaf nodes, where a path from the root to a leaf node represents a decision rule to determine the class of the test sample.

For example, the first branch in the decision tree gives the rue as

If Gender<=0.475 and Gender <=0.192 and Country <= 0.337 and education <=0.571 then the sample belongs to Class 0 (Non-User).

3. Explain why the algorithm didn’t do something else. [10 marks]

The input features given to the algorithm and dataset used determines the development of the model to great extent. Since we performed the Chi-Square test to select the features which are of greater importance to determine the target class, the module produced obtained results. Moreover, the dataset is biased, as there are 274 records belonging to the class “Non-User” and 988 records for the class “User”. Because of skewed data, the precision and recall values are high.

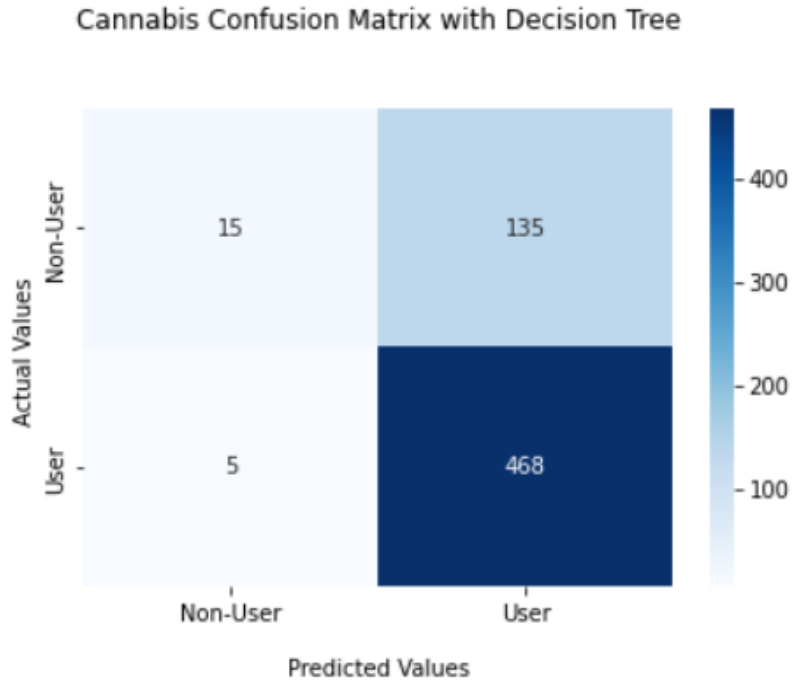
The algorithm is parameter tuned to perform tasks in a particular manner. The input parameter max_depth is set to 4 thus the tree was trained till level 4, had it been something else, the tree could have trained till more levels. Similarly, the split is performed based on the formula entropy for the number of samples considered for the split. The random state is set to 0, which represents that there will not be any shuffling of records performed for the data samples. By experimenting with these parameters, the model's behaviour can be modified.

4. Discuss when the algorithm succeeded and when it failed. [10 marks]

The algorithm has an accuracy of 0.7753 which is considered moderate. As the input data was skewed recall value for the model is considerably high because the model is well-trained to detect the majority class as compared to the minority class. Moreover, the nature of data is an important point to consider while using a decision tree. It is not suitable to handle continuous data. Moreover, with the minute changes in input data, the model forms a new model with major changes.

In the confusion matrix, you can see that there are 468 samples correctly classified as “User” but only 15 as “Non-User” with 135 records falsely classified as “User”. This clearly shows that model is biased toward the majority class. Here the model may not be a great suit if the minority class is equally important as the majority class. On the other hand, the algorithm is performing well for the Majority class.

```
[Text(0, 0.5, 'Non-User'), Text(0, 1.5, 'User')]
```



precision 0.7761194029850746

Recall 0.9894291754756871

5. Explain how you would decide if the resultant model can be trusted. [10 marks]

There are various metrics like accuracy, precision, recall, and f1- score. When the dataset is balanced, these metrics are sufficient to provide the decision-making criteria to select a particular model. But with the imbalanced dataset, these metrics often fail to give sufficient clarification on the results.
For Highly imbalanced datasets, the model might face an accuracy paradox where accuracy can be misleading and the model is efficient to predict the majority class but fails to provide good results for the minority class. In such a scenario higher accuracy cannot be the final goal. In this case, the input data needs a resampling and it may generate a better or more reliable model.
In the drug consumption data used here, it is slightly imbalanced thus we are getting moderate accuracy but a very high recall value.
ROC curve can be another metric which can show the model growth throughout time and graphically represented. The area under the curve gives the accuracy of the model which is more reliable than other metrics.

6. Explain how the algorithm could potentially improve its predictions. [10 marks]

To improve the algorithm's prediction, the parameters of the model can be carefully selected using techniques like gridSearchCV, etc. It is a widely used method to find the optimal parameters that generated more accurate and efficient models. Maintaining the quality of data also alters the model's accuracy.
Another way to improve the decision Tree's performance is by using ensemble methods. Ensemble methods combine several base models and provide an optimal model. Some models are trained more efficiently on some data, thus multiple models can be generated and combined to provide the best model that works well with all kinds of data.