**REPORT - Decision Tree and Naive Bayes**

1. The Decision tree is a method which we are using to determine the classification and the prediction of the values. We are using the decision tree methodology mainly in classification and prediction. Basically, we are defining some rules visually for simple interpretation and understanding. There are three different nodes in the Decision tree which are root node, branch node and the leaf nodes. We are using GINI and the Entropy to calculate the Decision tree. We are using both in this assignment and Entropy is te measure of the uncertainty. The Naïve Bayes classifier explains to us that once we have too many events and we don’t know how to handle them we will be checking some Naïve assumptions to work easier. It is a classification technique based on the Bayes theorem with an assumption of independence among predictors. It assumes that the presence of feature in the class is unrelated to the presence of the feature even if the feature depends on each other it will independently contribute to the probability whether the classification is done correctly.
   1. This is the formula which will calculate the GINI

A picture containing object, clock

Description automatically generated

* 1. This is the formula for calculating the Entropy



1. The data set which was given had 14 columns we have to predict the output of the person whether the person is affected by cardio diseases of heart and by training the dataset with the factors which is varying from gender ,height, etc. with their medical status, we will be training the model. We will predict the output of the classifier which has been trained under the same dataset. A close up of a keyboard

   Description automatically generated

**Features of the Data Set :**

1. Age – The feature mentioned is an integer value which is a numerical value .
2. Gender-It is a feature which is an Integer where 0 denotes Female and 1 denotes Male It is a numerical value.
3. Height-It is a feature where it is measured in centimeters and it is numerical value.
4. Weight- It is a feature where it is measured in kg and it is a numerical value.
5. Ap\_hi-It is Systolic BP which is an integer and have numerical value .
6. Ap\_lo-It is Diastolic BP which is an integer and have numerical value .
7. Cholesterol- It is a feature of the data set and it is having categorical values varying from values 1 to 3 by which 1 is risky 2 is normal and 3 is safe .
8. Gluc- It is a feature of the data set and it is having categorical values varying from values 1 to 3 by which 1 is risky 2 is normal and 3 is safe .
9. Smoke- It is a feature of the data set and it is having categorical value with having binary variation by which 1 smokes and 0 doesn’t smoke .
10. Aloc- It is a feature of the data set and it is having categorical values value with having binary variation by which 1 haves alcohol and 0 doesn’t alcohol.
11. Active- It is a feature of the data set and it is having categorical values value with having binary variation by which 1 does activity and 0 doesn’t do activity.
12. Cardio- It is a feature of the data set which we have to predict which is having the values 0 and 1. 0 being the cardio diseases are not present and 1 being affected by cardio diseases
13. Yes we are doing data Preprocessing by which the data is split into 30 % for splitting and 70% for training the model. The data which we are using is the data which had 70000 values hence we train the model with the training values. Then we predict the model with the test data. The testing the model will assure the precision of the output
14. For visualizing the Decision tree, we have to calculate the impurity for the node and the best ways to calculate the impurity is by the two techniques. Here we are doing both GINI and Entropy and we are training the data initially by which it will now starts to calculating the GINI/Entropy which results in constructing the initial node with minimum GINI/Entropy and then choose the other attributes accordingly. I have used tree.plot\_tree which is a sklearn package which will help in visualizing the whole model. Due to size of the dataset I am displaying only till the 3rd split due to the scaling problems and with the help of the export\_graphviz in the same package I will print the Decision tree. For both GINI and for Entropy
15. **GINI Calculation**

A picture containing clock

Description automatically generated

1. **Entropy Calculation**

A screenshot of a cell phone

Description automatically generated

1. After training the model and predicting the testing data with GINI and Entropy. On comparing the results of the entropy and GINI, which differ in less than 1 percent. Although, there is not much of a difference entropy might be a little slower for computation, GINI is more efficient when it comes to computation because it uses square rather than log and the library is implemented by default. Hence the samples which is recorded for splitting the data is same as well as the split is nearly same for both of them the GINI is having much more higher probability of getting better results compared to entropy.

**Results**

**GINI Calculation**

A screenshot of a cell phone

Description automatically generated

**Entropy Calculation**

A screenshot of a cell phone

Description automatically generated

From the above results it is clear and evident that the entropy calculation is having slightly less accuracy than GINI so we can say that the GINI calculation is the best and optimal method for calculating the impurity.

Reference:

<https://dataaspirant.com/decision-tree-algorithm-python-with-scikit-learn/>

https://www.statsdirect.com/help/nonparametric\_methods/gini\_coefficient.html/