**ASSIGNMENT 2**

**Decision Tree and Naïve Bayes Classifier**

Decision Trees are the most remarkable tool that is used for the classification and regression values for a given variable. The decision tree method will construct a tree like structure based on the number of attributes and their values in the given dataset. The decision tree is constructed in a manner where we could understand it very easily because it is sorted from the root node and based on each and every decision in the tree it is split into different subsets. The records are split into sets for further classification:

1. Training set.
2. Test set.

The decision trees can classify the data based on their attributes, all the similar attributes in the data are grouped into subsets so that it makes the work easier to sort the data based on its attributes. The classification of data in the decision trees thus becomes the most important feature as we could easily differentiate and understand large datasets.

The accuracy of the classifier can also be predicted so that we could know the number of the test samples that are correctly classified. Therefore we can build an accurate model using all these attributes and classification model. The decision trees generate easily understandable structures and provide clear information about which of the data is more important for classification.

Gini Index: The gini index will allow us to calculate the probability of a feature from the dataset, the value of gini index is between 0-1 if it is 0 then the attributes belong to the same class else they belong to many other classes.

Gini Index= A screenshot of a cell phone

Description automatically generated

The entropy tells what variable should be split next at every step and it is also called as information gain.

Naïve Bayes Classifier:

The naïve bayes classifier is a method which uses the Bayes method on the features of the class using the naïve assumptions that each feature is independent and contribute at the same level to the outcome.

According to Bayes theorem:

P(X/Y)=P(Y/X)\*P(X)/P(Y)

**Description of the dataset cardio\_train.csv:**

There are no null values in the dataset so there is no need of data preprocessing for cardio\_train.csv. The dataset description is given in the following lines mentioned below:

1. ID: The id is specified using integer type and it denotes the Id assigned to the data in the given dataset.
2. Age: The age is specified using integer type and it denotes the age of the patient in the given dataset.
3. Gender: Gender is mentioned in binary format where it has male and female, 1 for female and 2 for male.
4. Height: The height of the patient it is mentioned using integer.
5. Weight: The weight of the patient it is mentioned using integer.
6. ap\_hi: Value of ap\_hi integer value.
7. ap\_lo: Value of ap\_lo integer value.
8. cholesterol: The cholestrol level of the patient 1-3; 1- normal, 2 - above normal, 3- well above normal.
9. Gluc: The glucose level of the patient is 1-3; if 1- normal, if 2 - above normal, if 3- well above normal.

10)Smoke: The smoke habit of the patient is described in this column 0 – if no smoke, 1- if smoke.

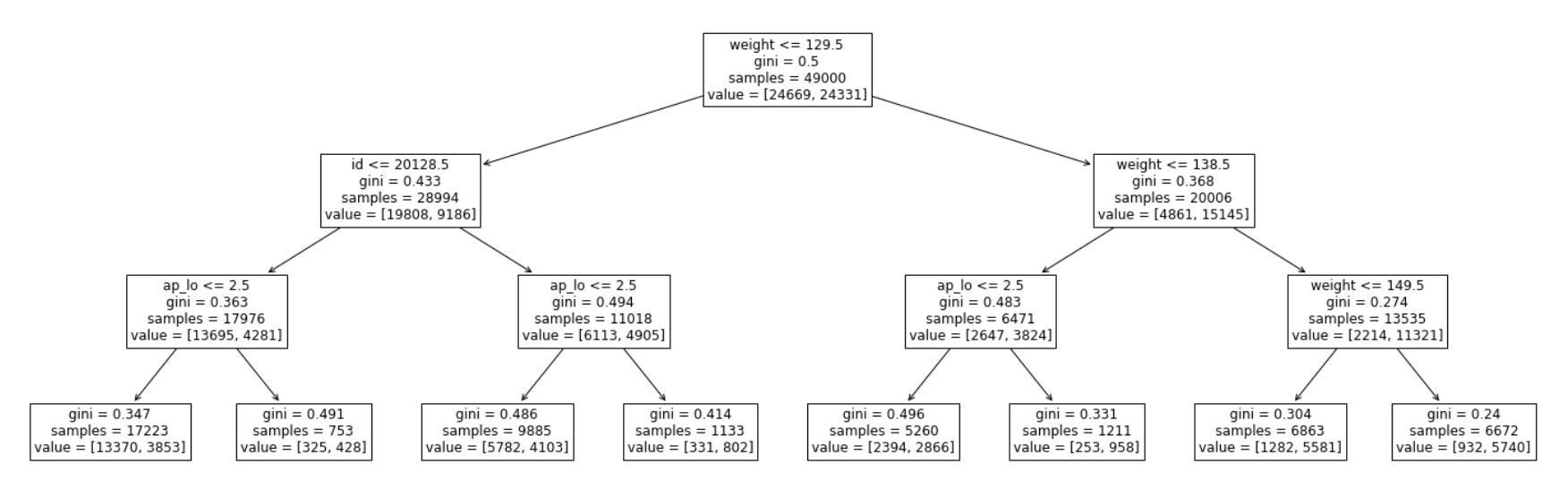
11) alco: The patient consumes alcohol or not is described in .if 1- yes, if 2- no.

12) active: Physically active or not if 1- yes, if 2- no)

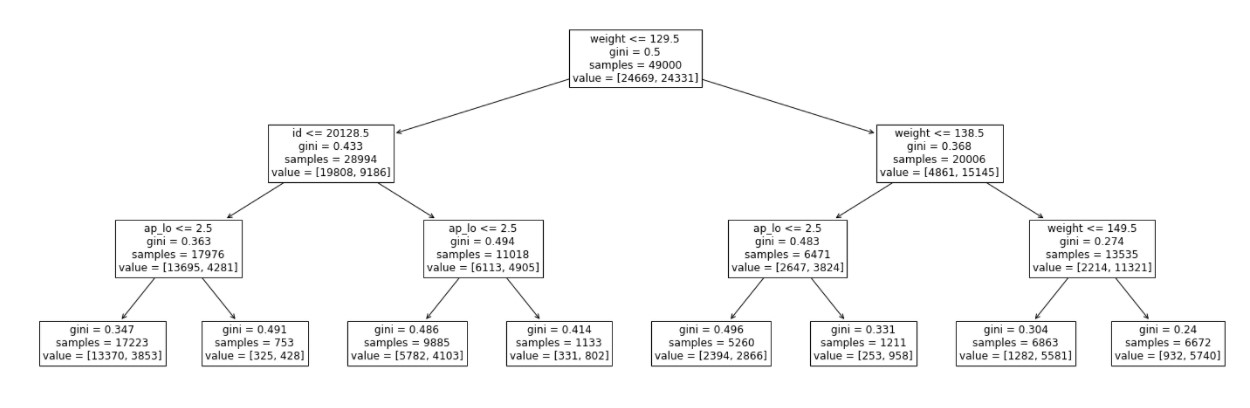
13) cardio: Class label that has to be determined if there is a chance for cardio disease if 1- yes, if 2- no.

**Visualization of gini and entropy using decision tree algorithm:**

**Visualization of decision tree using Gini:**

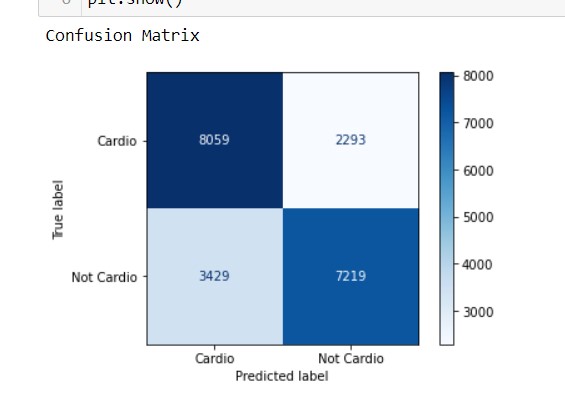
****

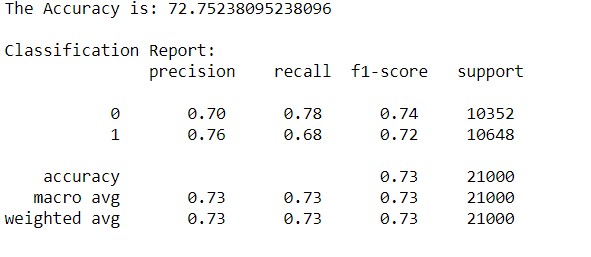
**Visualization of decision tree using Entropy:**



The visualization of the given dataset using Gini and entropy is shown in the above figure.

**Confusion Matrix for the dataset using Gini:**

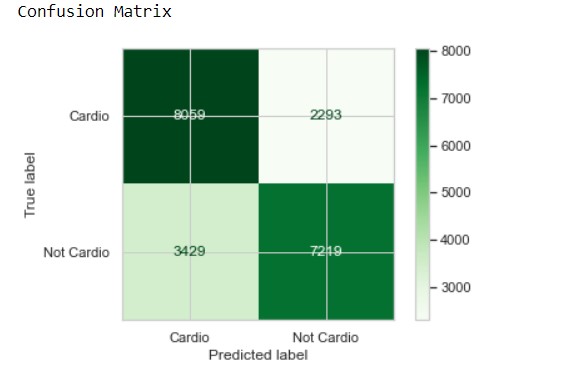


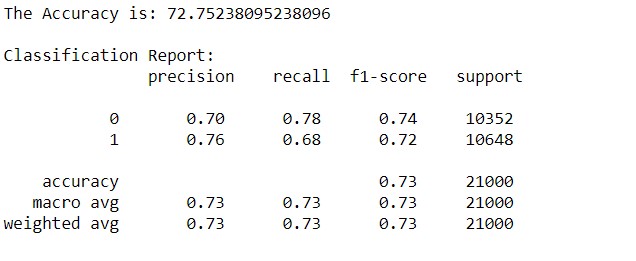


The Accuracy of dataset using gini is: 72.75238095238096.

Rows represent the actual values and the columns represent the actual values of the given dataset. In 1st row actual number of cardiovascular are 10,352 among which 8059 are correctly classified as negative and 2293 are wrongly classified as positive, in the second row there are 10,648 yes values among which 7219 are correctly classified as true and 3429 are wrongly classified as negative. Using these the accuracy of the dataset is calculated as 72.752.

**Confusion Matrix using Entropy:**





The Accuracy using entropy is: 72.75238095238096.

Rows represent the actual values and the columns represent the actual values of the given dataset. In 1st row actual number of cardiovascular are 10,352 among which 8059 are correctly classified as negative and 2293 are wrongly classified as positive, in the second row there are 10,648 yes values among which 7219 are correctly classified as true and 3429 are wrongly classified as negative. Using these the accuracy of the dataset is calculated as 72.752.

Compare Gini and Entropy:

|  |  |  |
| --- | --- | --- |
|  | **GINI** | **ENTROPY** |
| **ACCURACY** | 72.75238095238096 | 72.75238095238096 |

Confusion Matrix with Gini:-

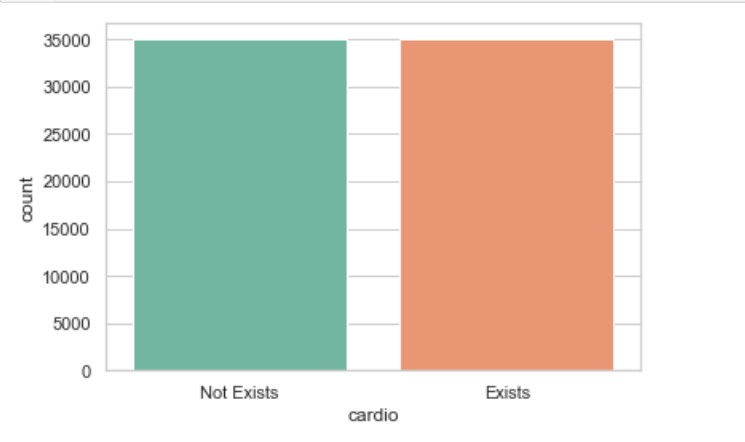
|  |  |  |
| --- | --- | --- |
|  | Predicted:No | Predicted:Yes |
| Actual:No | 8059 | 2293 |
| Actual:Yes | 3429 | 7219 |

Confusion Matrix with Entropy:-

|  |  |  |
| --- | --- | --- |
|  | Predicted:No | Predicted:Yes |
| Actual:No | 8059 | 2293 |
| Actual:Yes | 3429 | 7219 |

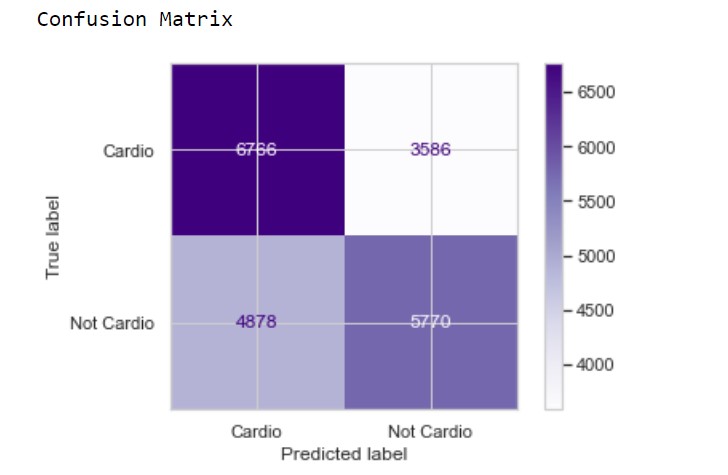
**Visualization of Target Variable:**

The visualization of the target variable is shown:



The confusion matrix for Naïve Bayes Classifier:

|  |  |  |
| --- | --- | --- |
|  | Predicted:No | Predicted:Yes |
| Actual:No | True Negative:  6766 | False Positive:  3586 |
| Actual:Yes | False Negative:  4878 | True Positive :  5770 |



Rows represent the actual values and the columns represent the actual values of the given dataset. In 1st row actual number of cardiovascular are 10,648 among which 6766 are correctly classified as negative and 3586 are wrongly classified as positive, in the second row there are 10,648 yes values among which 4878 are correctly classified as true and 5770 are wrongly classified as negative. The Accuracy is: 59.69523809523809.

**Result:**

The accuracy for gini and entropy for the decision tree is same 72.75238095238096 and for the naïve bayes the accuracy is 59.69523809523809 which is less, so here we conclude that decision tree is a better classifier than naïve bayes.