**DATA MINING ASSIGNMENT-2**

**DECISION TREE REPORT**

1. **Describe the Decision Tree methods, and Naive Bayes classifier.**

DECISION TREE:

Decision Tree Mining is a type of data mining technique that is used to build Classification Models. It builds classification models in the form of a tree-like structure, just like its name. This type of mining belongs to supervised class learning. Here we take the dataset and train the data model (divide the dataset) and then test the model using testing data (to check for accuracy of the data model)

A two-step process is followed, to build a classification model.

1. In the first step i.e. learning: A classification model based on training data is built.
2. In the second step i.e. Classification, the accuracy of the model is checked and then the model is used to classify new data.

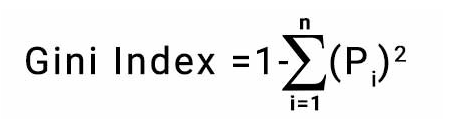
The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified.

DECISION TREE CONSTRUCTION:

A decision tree is a classification scheme which generates a tree and a set of rules representing the model where each internal node denotes a test on an attribute, each branch denotes outcome of the test and each leaf node holds a class label. The topmost node in a tree is the root node.

When data is given the most important work is to select root node and it is selected based on Gain value =P-M where P is impurity measure before splitting and M is impurity measure after splitting.

To calculate impurity measure of each attribute we have few methods like Gini Index, Entropy and misclassification Error



where Pi denotes the probability of an element being classified for a distinct class.



NAÏVE BAYES CLASSIFIER:

A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem.

BAYES THEOREM:

Let X be a data tuple.Let H be some hypothesis such that the datatuple X belongs to a specified class C . We want to determine P(H|X) the probability that the hypothesis H holds that the tuple X belongs to class C, given the attribute description of X. For Example P(H|X) means the probability that customer X will buy a computer given that we know the customers age and income

P(H|X)=P(X|H)P(H)

P(X)

1. **Describe the datasets and code.**

Cardio dataset has 70000 rows and 13 columns. The attributes are id, age, gender, height weight, ap\_hi, ap\_lo , cholesterol , gluc, smoke, alco, active, cardio. Here, cardio is the target variable.

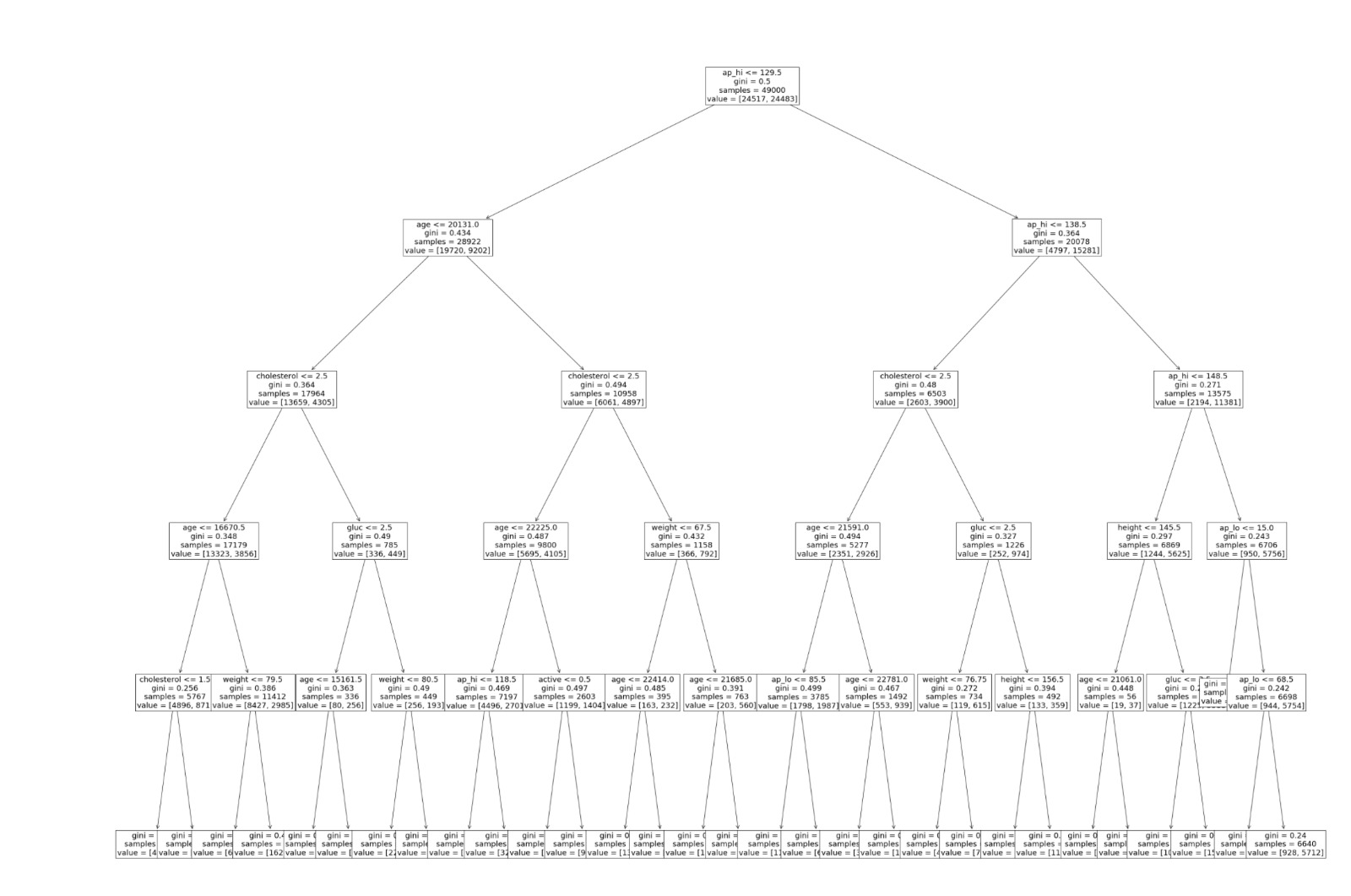
It predicts using training model if a patient has heart disease or not. The dataset is split into 70% training and 30% testing. The training data is used to train a model and this model has to predict whether any patient has heart disease or not of testing data.

Pre-Processing and Code Explanation:

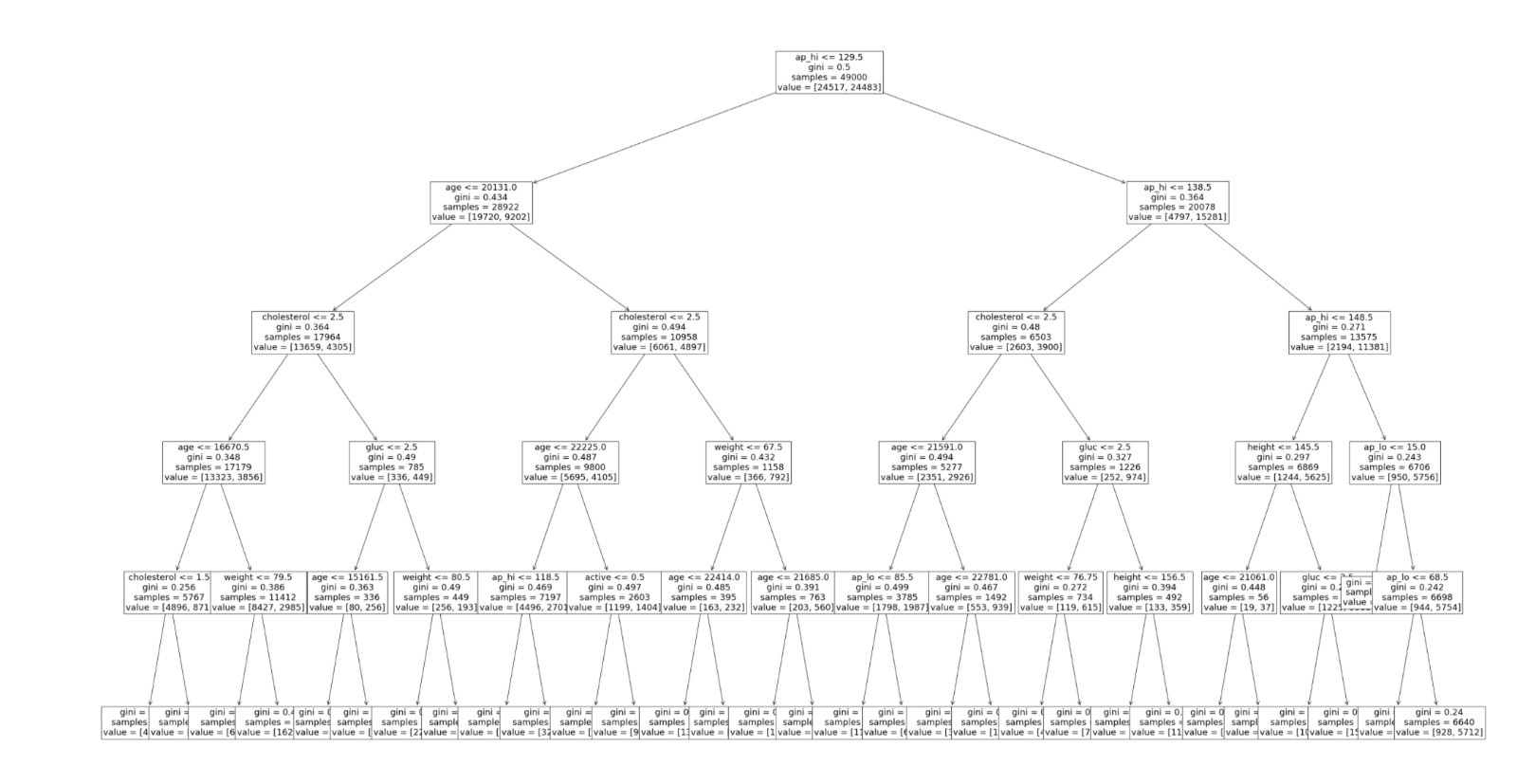
We read the csv file and convert it into dataframe.

**dt\_data= pd.read\_csv(“cardio\_train.csv” , sep =’;’)**

1. **Visualization of the decision tree for Gini and Entropy.**



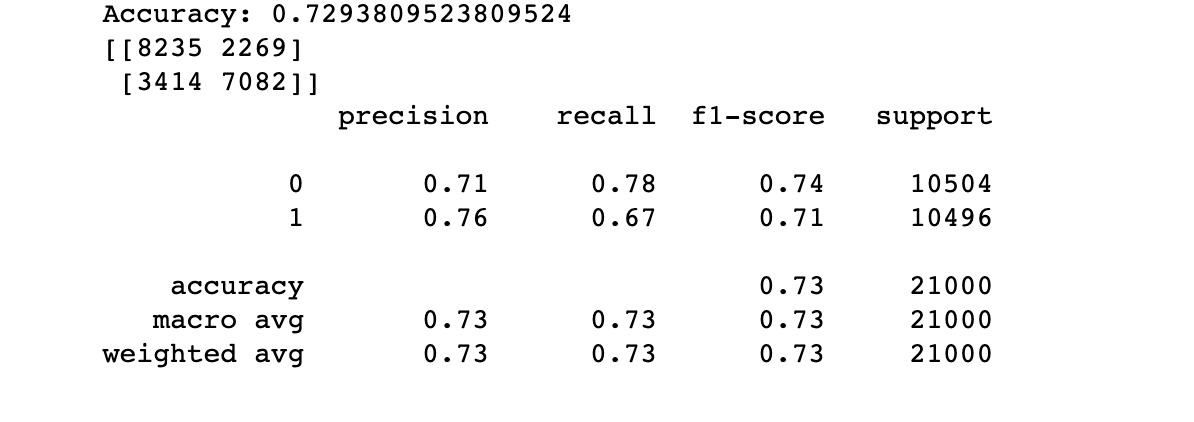
This above figure is the decision tree using Gini Index and has maximum depth = 6 and the root node is ap\_hi <= 129.5



The above figure is the decision tree using Entropy and has maximum depth=6 and the root node is same as entropy decision tree i.e: ap\_hi <= 129.5

4) Interpret your results, and do not forget to compare Gini and entropy.

Using Entropy:



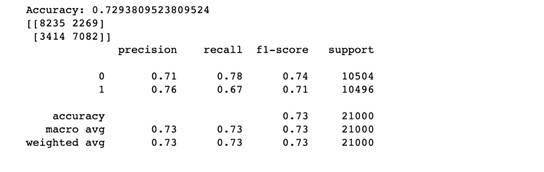
* Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted: NO | Predicted: Yes |
| Actual: NO | True Negative:  8235 | False Positive:  2269 |
| Actual: Yes | False Negative:  3414 | True Positive:  7082 |

Columns represent predicted values and rows represent actual values. Here in the 1st row, there are 10504 values, among which 8235 are correctly classified Negative and 2269 are wrongly classified as Positive.

In the 2nd row there are 10,496 Yes values among which 7082 are correctly classified as Positive and 3178 are wrongly classified as Negative

Using Gini Index:



* Confusion Matrix:

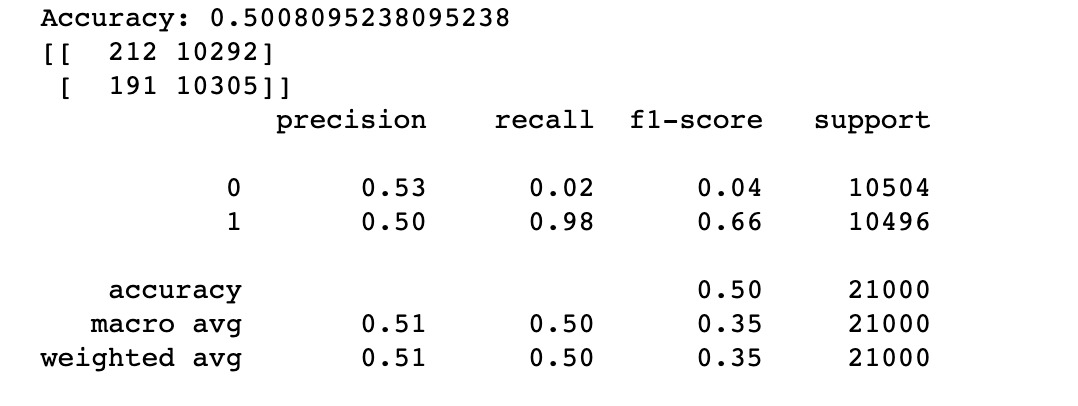
|  |  |  |
| --- | --- | --- |
|  | Predicted: NO | Predicted: Yes |
| Actual: NO | True Negative:  8235 | False Positive:  2269 |
| Actual: Yes | False Negative:  3414 | True Positive:  7082 |

Columns represent predicted values and rows represent actual values. Here in the 1st row, there are 10504 values, among which 8235 are correctly classified Negative and 2269 are wrongly classified as Positive.

In the 2nd row there are 10,496 Yes values among which 7082 are correctly classified as Positive and 3178 are wrongly classified as Negative.

Both Gini Index and Entropy have the same accuracy values.

Naïve Bayes Classifier:



* Confusion Matrix:

|  |  |  |
| --- | --- | --- |
|  | Predicted: NO | Predicted: Yes |
| Actual: NO | True Negative:  212 | False Positive:  10292 |
| Actual: Yes | False Negative:  191 | True Positive:  10305 |

Columns represent predicted values and rows represent actual values. Here in the 1st row, there are 10504 values, among which 10292 are correctly classified Negative and 212 are wrongly classified as Positive.

In the 2nd row there are 10,496 Yes values among which 10305 are correctly classified as Positive and 191 are wrongly classified as Negative.