**PA#2 Report for DT NB**

Aaditya Damle – 1001955625

Akshay Sadalgekar – 1001944713

Aishwarya Muchandi - 1001

1. Describe the Decision Tree methods, and Naïve Bayes classifier.

Answer: **Decision Tree method:**

* Sklearn.tree.DecisionTreeClassifier is the method provide by sklearn to perform Decision Tree for classification.
* Some parameter passed for this method are criterion which chooses between “**gini**” or **“entropy”**, **max\_depth** defines the maximum depth of the tree, **min\_samples\_leaf** sets the min. number of samples required to be at leaf node. There’s also a **random\_state** parameter to reproduce same output during multiple runs.
* This method is used to define the model. The **fit()** method applied to the model will train the model using the attributes which are genrally stored in Train\_X variable  
  and the outputs Train\_Y variable.
* To get prediction for checking accuracy we use the **predict()** method.

**Naïve Bayes Classifier:**

* Here, we used the **Gaussian Naïve Bayes** method for classification. The likelihood of the features is assumed to be Gaussian.
* GaussianNB() is the method used to define the model.
* The **fit()** method applied to the model will train the model using the attributes which are genrally stored in Train\_X variable  
  and the outputs Train\_Y variable.
* To get prediction for checking accuracy we use the **predict()** method. This method has only one parameter “Test\_X” which contains the attributes in the testing set.

1. Describe the datasets like what do you understand from the dataset? And if you have done any pre-processing, and your code, please write down your observation.

Answer:

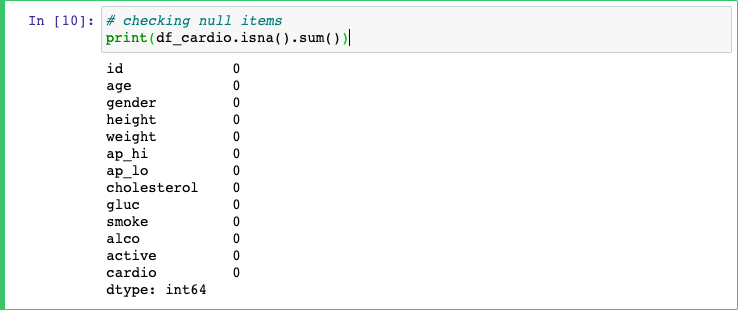
* This dataset consists of heart related information and some other information which can affect the occurrence of a cardiovascular disease in a human being.
* Some of the attributes in the dataset are age, gender, cholesterol, glucose, smoke, alcohol, active etc.
* This dataset can be used to build a classifier to predict the occurrence of cardiovascular disease.
* In the preprocessing step,

1. We have check whether there are any null values in the dataset by using the following code:

checking null items

print(df\_cardio.isna().sum())

Output:



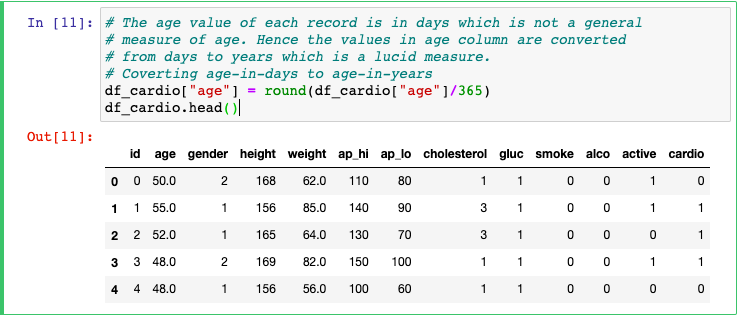
As seen in the output, there are no null values present and thus no further action is needed.

1. We also changed the value of a variable for all the records. The value of “age” attribute was in days which is very unintelligible measure of age. So we changed it to years by dividing the values by 365(Number of days in one year).

df\_cardio["age"] = round(df\_cardio["age"]/365)

df\_cardio.head()

Output:

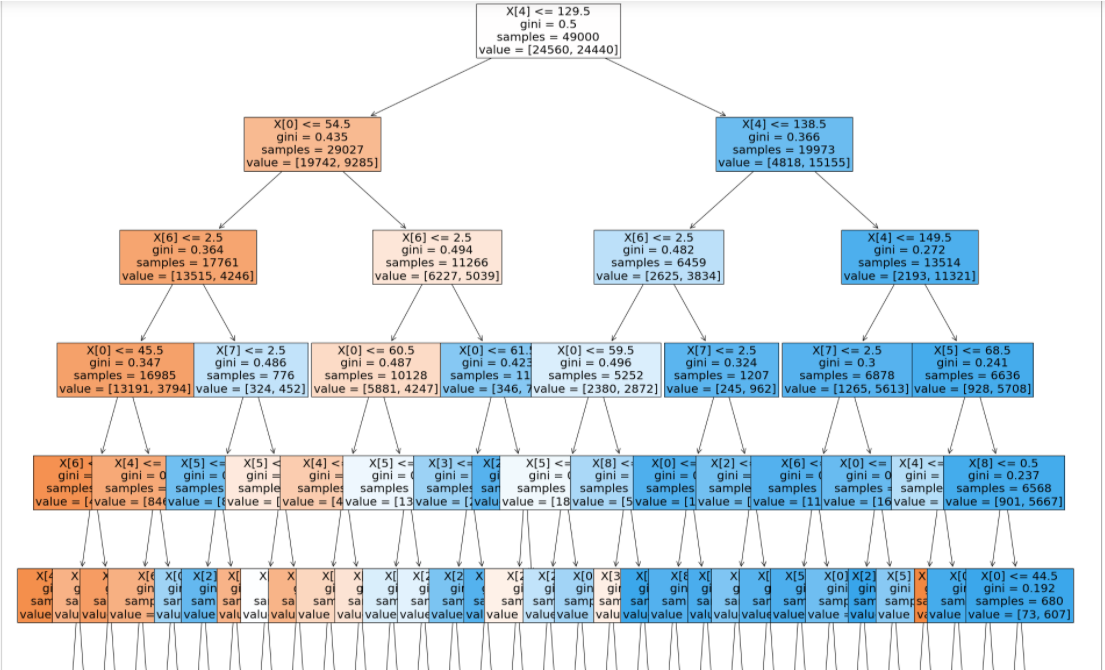


1. Visualization of the decision tree for gini and entropy.

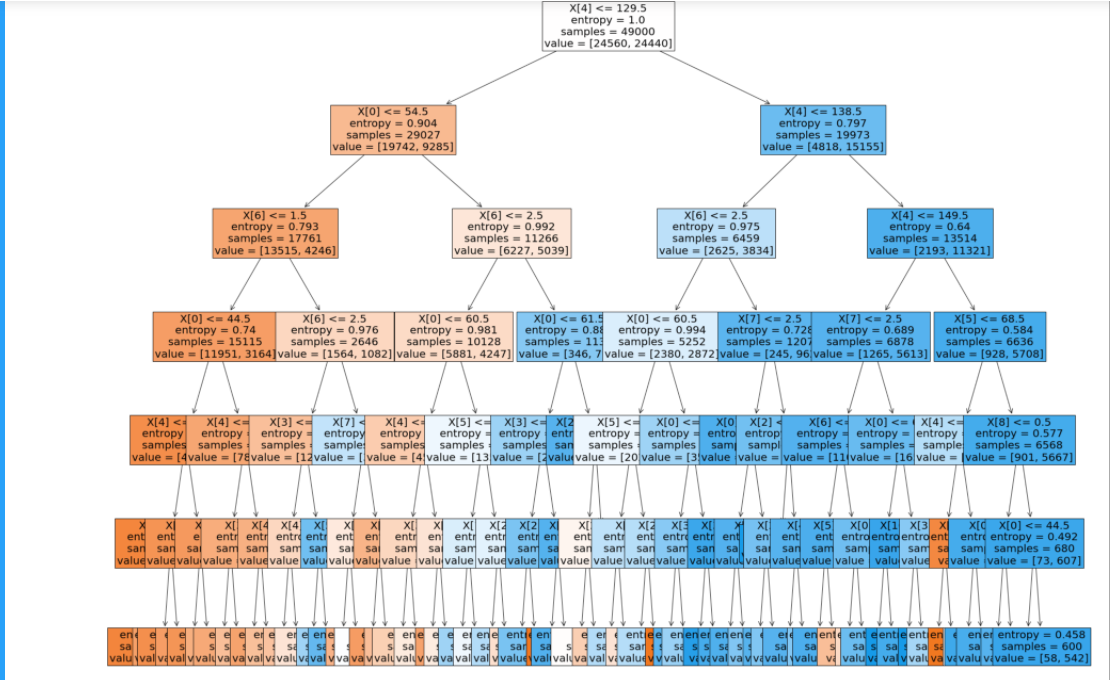
Answer: In the following visualizations, the nodes in higher depths are merged with each other because the font size is greater. Smaller font size will set them apart.

**Visualization of Decision Tree:**

**Gini:**

****

**Entropy:**

****

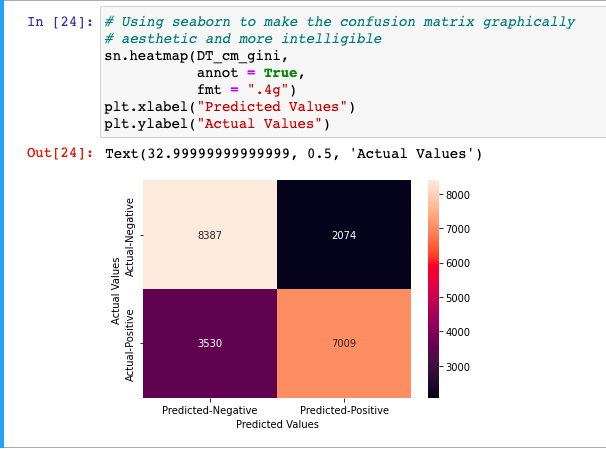
1. Interpret your results, compare gini and entropy.

Answer:

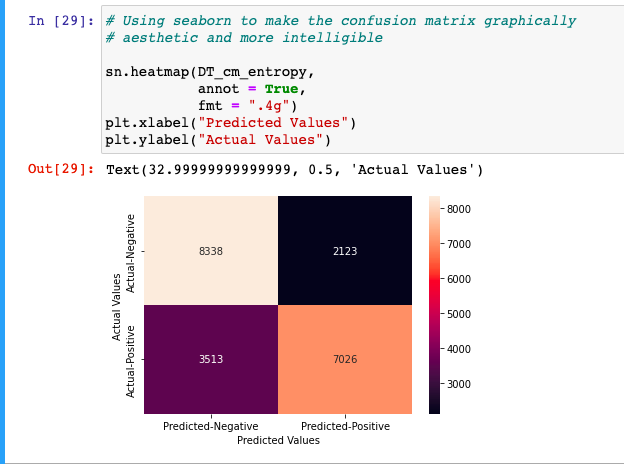
* Both the trees have used X[4] attribute to make decision at the root node.
* Both trees have mostly similar samples and values. As we go in depth of the trees the values start showing some differences.
* The value of gini at root node is 0.5 while that of entropy is 1.0.
* Let’s consider confusion matrix to examine the results:

In Gini more records are correctly predicted negative while in Entropy more values are correctly predicted positive.

1. Gini



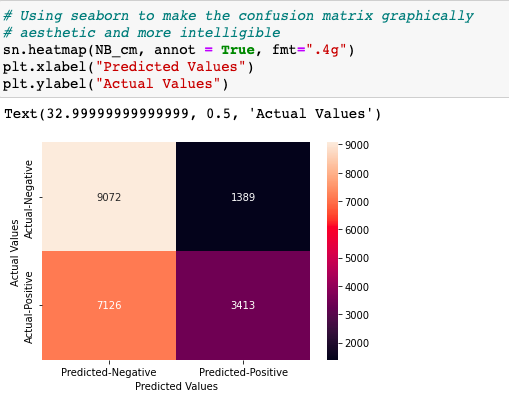
1. Entropy



1. Compare the results of DT(gini), DT(entropy) and Naïve Bayes.

Answer:

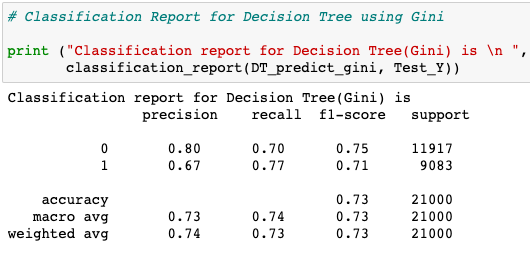
* The accuracy score for Gini is 73.314% and entropy is 73.162% which is approximately similar but accuracy score for Naïve Bayes is very low 59.452%.
* Let’s consider the confusion matrix of Naïve bayes and compare it with both matrices of DT which we plotted above:



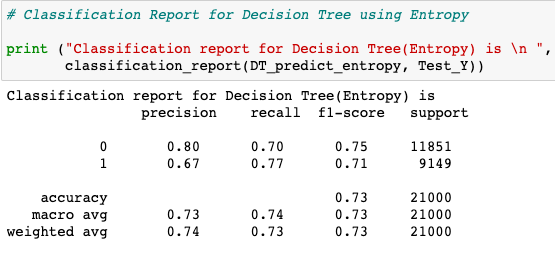
Here, as the accuracy is comparably low, more values which are actually positive are wrongly predicted as negative.

Now let’s compare the classification report of all 3 classifiers.

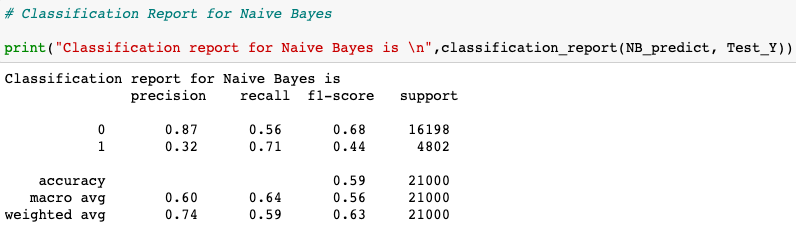
Gini:



Entropy:



NaiveBayes:



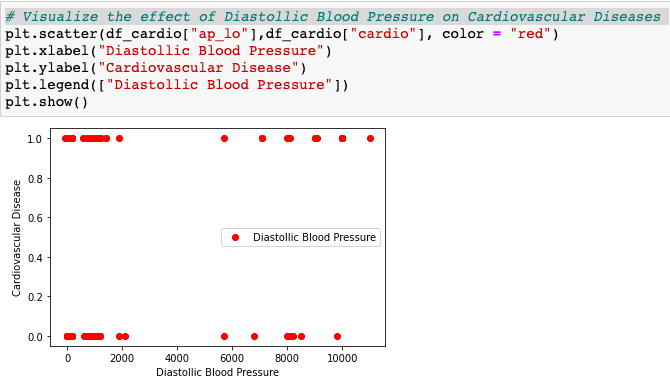
The values of precision, recall, f1 and support are almost same in the classification report for Gini and Entropy.

For Naïve Bayes we get, less precision for value 1 (0.32), less recall for value 0 (0.56) and less f1 score for both the values. The values of macro avg and weighted avg is affected accordingly.

1. Visualize the dataset, for the target variable – 2 graphs

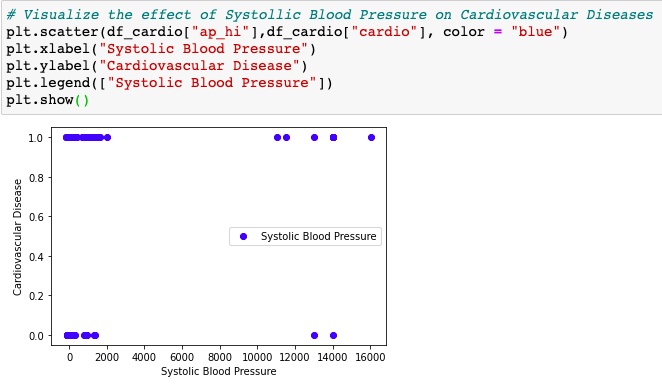
Answer:

Visualize the effect of Diastolic Blood Pressure on Cardiovascular Diseases:



So, when the Diastolic pressure decrease below a particular value the chance of cardiovascular diseases increases.

Visualize the effect of Systolic Blood Pressure on Cardiovascular Diseases:



Similar to Diastolic pressure, when the Systolic pressure decrease below a particular value the chance of cardiovascular diseases increases.

References:

<http://education.abcom.com/admit-prediction/>

<https://mljar.com/blog/visualize-decision-tree/>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html>

<https://towardsdatascience.com/visualizing-decision-trees-with-python-scikit-learn-graphviz-matplotlib-1c50b4aa68dc>

<https://seaborn.pydata.org/generated/seaborn.pairplot.html>

<https://www.youtube.com/watch?v=cpZExlOKFH4>