**#PA 2: Classification - Decision Tree Report**

**Describe the Decision Tree methods, and Naive Bayes classifier in details in your own words. Don’t copy paste it from the internet. Write it on your own. [5 points]**

A Decision Tree is a Supervised Machine Learning algorithm which looks like an inverted tree, where each node represents a predictor variable, the link between the nodes represents a Decision and each leaf node represents an outcome.

**Decision Tree methods:**

1. **Gini index:** This is used to calculate the amount of probability of a specific feature that is classified incorrectly when chosen at random. It is calculated by subtracting the sum of squared probabilities of each class from one. It has values between 0 and 1.

Gini Index is calculated as:

A close up of a clock

Description automatically generated

1. **Entropy**: Entropy is used to assess a dataset's degree of impurity or unpredictability. The entropy is used to determine the homogeneity of a data. Subsets of the data set are created based on related instance values. The entropy of a data collection with identical values is put to 0, whereas the entropy of a data set with equally distinct values is assigned to 1. After a dataset is split based on an attribute, entropy can be used to assess information gain.

Entropy is calculated as:

**Entropy = -p log2 p — q log2q**

**Naïve Bayes classifier:**

A naive Bayes classifiers are a set of algorithms that share a common fundamental as they are probabilistic classifiers that uses Bayes theorem and strong (naive) independence assumptions to classify data.

**Gaussian Naive Bayes:** A Gaussian Naive Bayes algorithm is a special type of Naïve Bayes classifier. It is specifically used when the features have continuous values. It is also assumed that all the features are following a normal distribution.

**Multinomial Naive Bayes:** Multinomial naive Bayes classifier is used for multinomial distributed data sets. Using the Bayes theorem, the program estimates frequency of words. It computes the likelihood of each word in a given sample and returns the data with the highest chance.

**Bernoulli Naive Bayes:** This is Similar to multinomial naive bayes, it uses Binary variables as predictors. The parameters we use to forecast the class variable only accept yes or no answers, such as if a word appears in the text or not.

**Complement Naive Bayes:** This algorithm essentially implements a modified multinomial naive bayes which is useful for imbalanced data sets. Data set can be considered imbalanced where the iteration of certain classes is more than other classes present which does not provide a uninform distribution.

**2) 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. [6 points]**

**Description:**

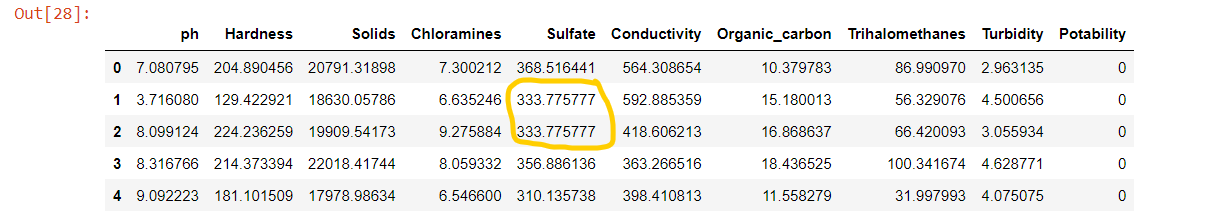
The dataset consists of 3276 records and 10 attributes. Some of the records have NaN values as well in the dataset

1. ph – It has values of type float
2. Hardness – It has values of type float
3. Solids – It has values of type float
4. Chloramines - It has values of type float
5. Sulfate - It has values of type float
6. Conductivity – It has values of type float
7. Organic\_carbon - It has values of type float
8. Trihalomethanes - It has data values of either 0 or 1
9. Turbidity - It has data values of either 0 or 1
10. Potability - It is used as a Target Variable and the data values are either 0 or 1

**Data Pre-processing:**

Since some of the values in the dataset have NaN values, we replace those values with the mean value of that attribute column. This will provide more accuracy while predicting the dataset. We Use Mean() function in Sklearn to update the NaN values in the dataset

We have split the dataset as: 70% for training,10% for testing and 20% for validation of the classifier. Below we could see the Mean values filled in place of NaN values



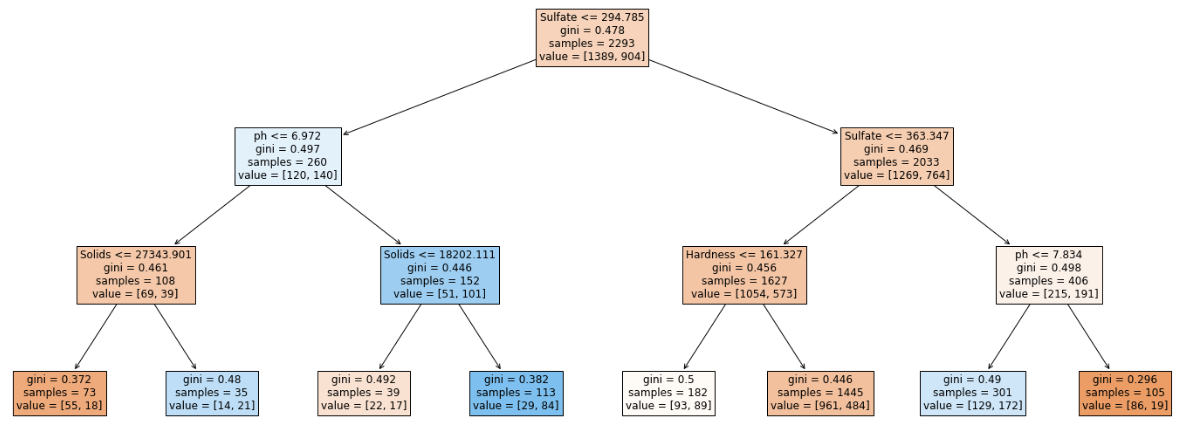
**Code:**

In the given dataset, “Potability” is chosen as a target variable because it is the only attribute which contains either 0 or 1.

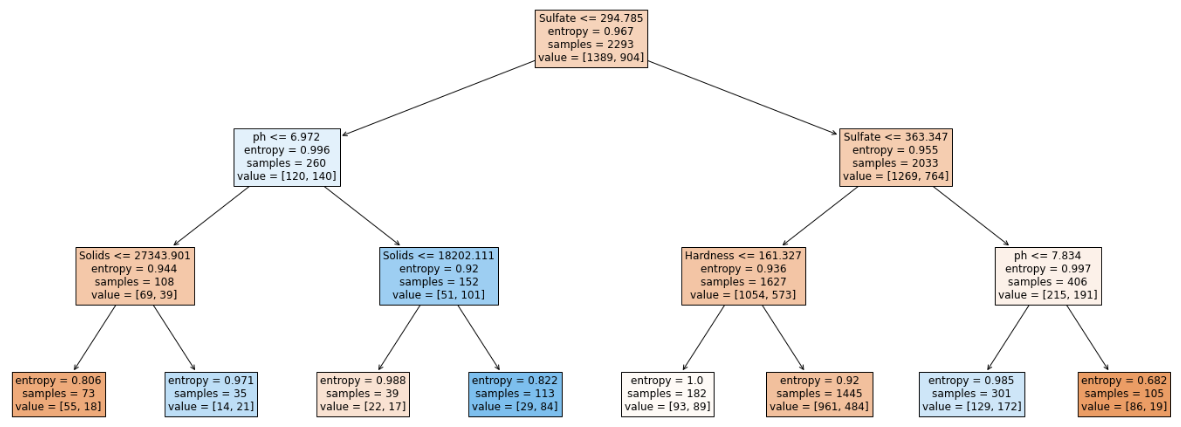
For the decision tree we use Decisiontreeclassifier() to predict the value of the target variable (Potability) by adapting simple decision rules. We generate both **Gini** and **Entropy**  
criterion to measure the quality of the split into test\_train\_valid.

**4) Visualization of the decision tree for gini and entropy.[4 points]**

**The Visualization for Gini with Max depth 3:**

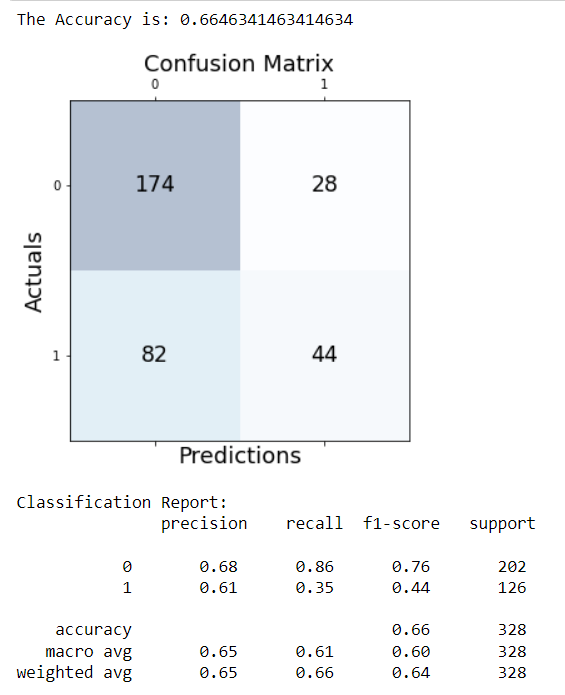


**The Visualization for Entropy with Max depth 3:**



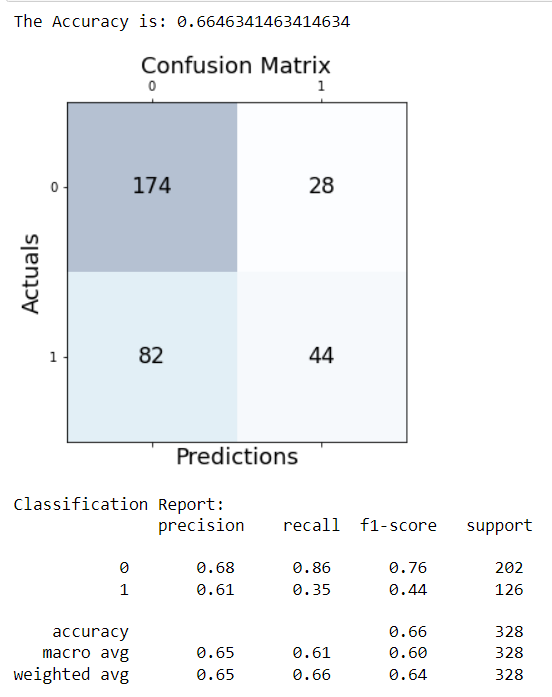
We have also calculated the accuracy score, confusion matrix and classification report using both Gini and entropy criterion. Both of them gave similar results.

**Accuracy score, Confusion matrix and Classification report using “gini” criterion:**

****

Here we could see that the accuracy is 0.667 for Gini and in the classification report it has more recall and f1-score value.

**Accuracy score, Confusion matrix and Classification report using “entropy” criterion:**

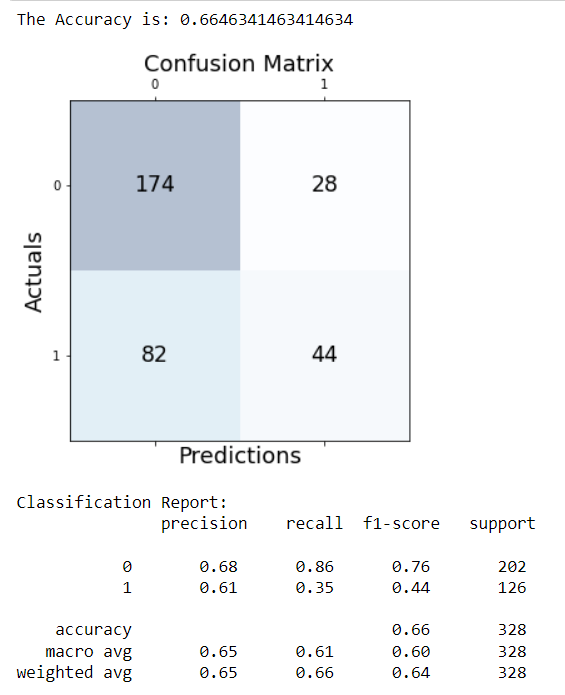
****

From the above two reports we could see that both Gini and Entropy predicts the values same to the target variable.

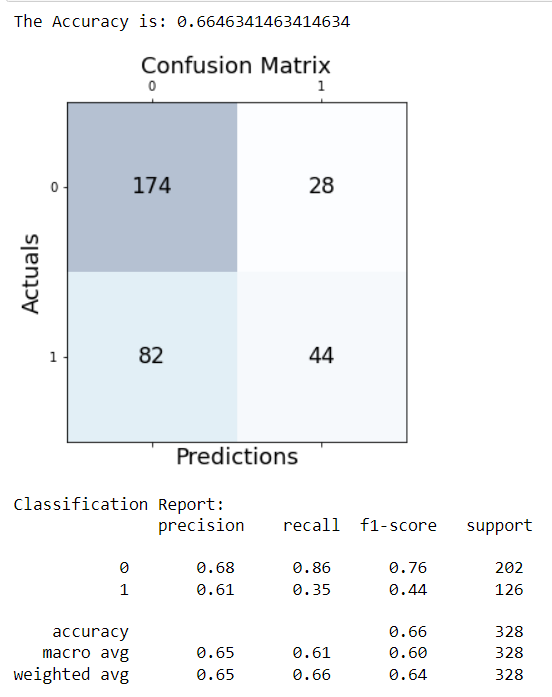
**5.) Interpret your results, compare gini and entropy [4 points]**

We used Decision Tree classifier with both gini and entropy functions and three Naive Bayes classifier to predict the test data. From the results, the accuracy score using Decision Tree classifier with both gini and entropy is the same. However, Naive Bayes classifier gives a comparatively lower accuracy score. Thus, Decision tree is a better classifier for the given data set.

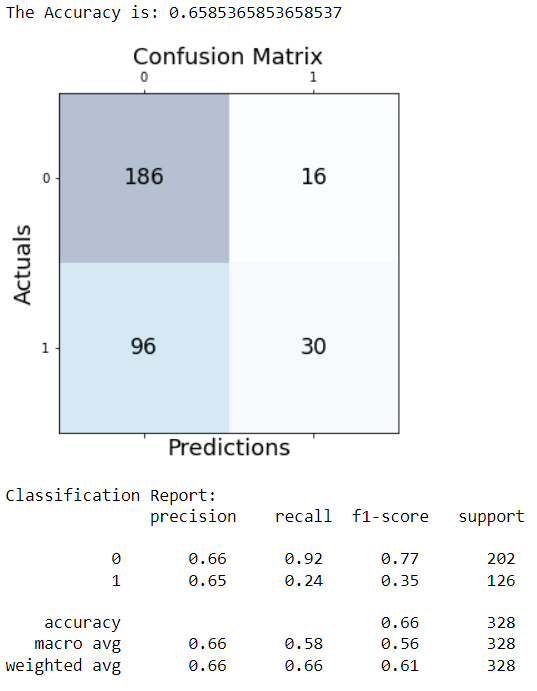
**Results using Decision Tree (gini criterion):**

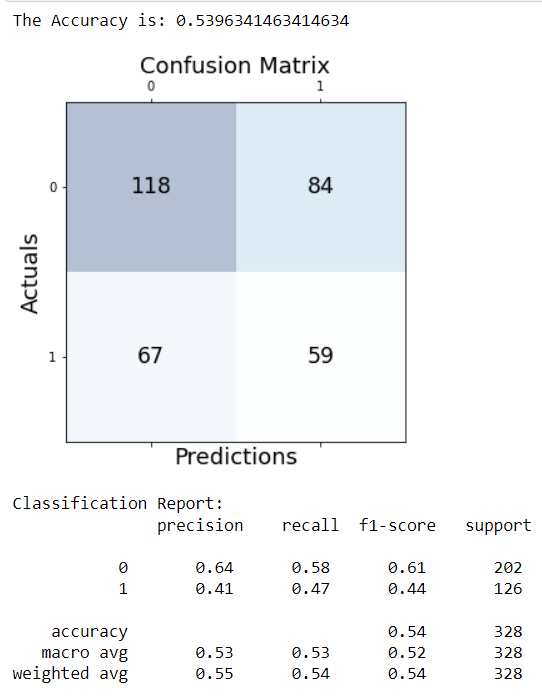
****

**Results using Decision Tree (entropy criterion):**

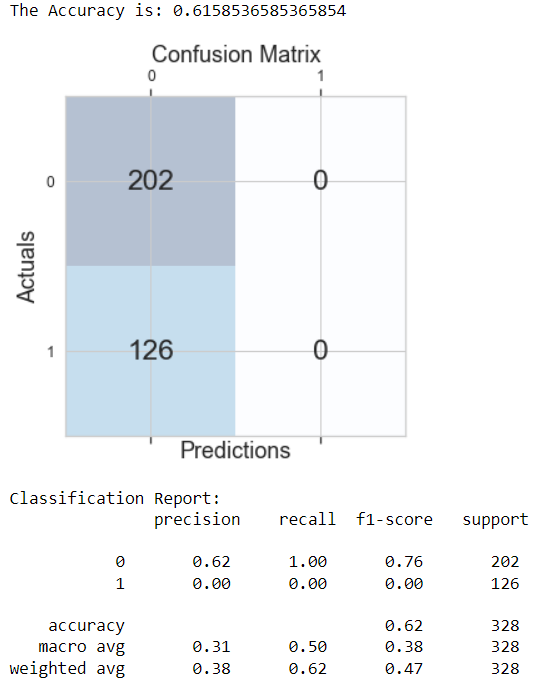
****

**Results using Gaussian Naive Bayes classifier:**

****

**Results using Multinomial Naive Bayes classifier:**

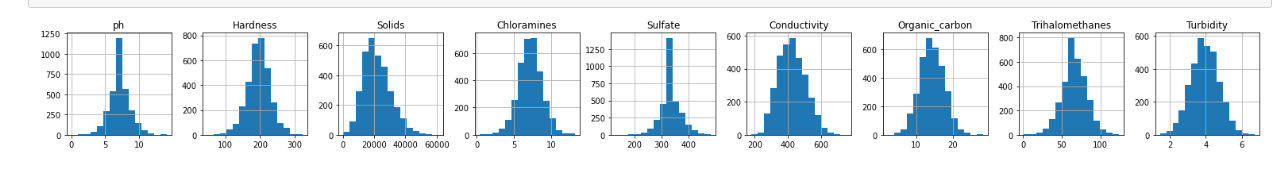
**Results using Bernoulli Naive Bayes classifier:**

****

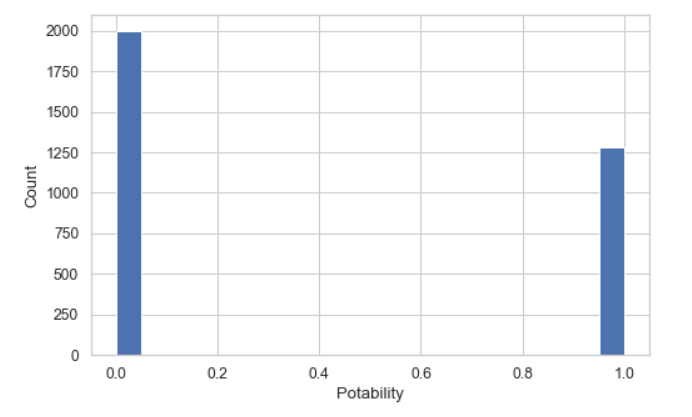
**6) Visualize the dataset, for the target variable - 2 graphs [4 points]**

We have split the data for testing, training and validating and we took the target variable as test data.

**Visual representation of train data for the target variable:**



**Visual Representation of test data for the target variable:**



**References:**

* [**https://www.w3resource.com/pandas/dataframe/dataframe-drop.php**](https://www.w3resource.com/pandas/dataframe/dataframe-drop.php)
* [**https://medium.com/@contactsunny/how-to-split-your-dataset-to-train-and-test-datasets-using-scikit-learn-e7cf6eb5e0d**](https://medium.com/@contactsunny/how-to-split-your-dataset-to-train-and-test-datasets-using-scikit-learn-e7cf6eb5e0d)
* [**https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html**](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)
* [**https://medium.com/@seocodingvidya/python-how-to-draw-confusion-matrix-using-matplotlib-61cd891a3c7f**](https://medium.com/@seocodingvidya/python-how-to-draw-confusion-matrix-using-matplotlib-61cd891a3c7f)
* [**https://towardsdatascience.com/how-to-impliment-a-gaussian-naive-bayes-classifier-in-python-from-scratch-11e0b80faf5a**](https://towardsdatascience.com/how-to-impliment-a-gaussian-naive-bayes-classifier-in-python-from-scratch-11e0b80faf5a)
* [**https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67**](https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67)
* [**https://scikit-learn.org/stable/modules/generated/sklearn.naive\_bayes.BernoulliNB.html**](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html)