Mandeep Report

The coding section for this report is divided into two parts:

1. Data Pre-processing
2. Machine Learning

# Data Pre-processing

This section is responsible for the data in the code. This section imports the dataset into the code, and analyzes it for any anomalies like duplicate datasets or null instances in the dataset and removes them. It also analyzes the dataset for any categorical features and encodes them using a Label Encoder which helps the dataset fit on the algorithm easily.

The first 5 instances of the dataset used in this research is shown in the figure below. The dataset has a total of 250 instances. Since the project requires an algorithm that can identify whether a facebook post is fashion related or not, this makes the problem a binary classification problem. The features in the dataset are as follows:

link\_name: the name of the facebook post. This feature will be used to extracting the bag of words for Natural Language Processing.

num\_reactions: total number of reactions on the post.

num\_comments: total number of comments on the post.

num\_shares: total number of shares for the post.

num\_likes: total number of likes on the post.

num\_loves: total number of love reacts on the post.

num\_wows: total number of wow reacts on the post.

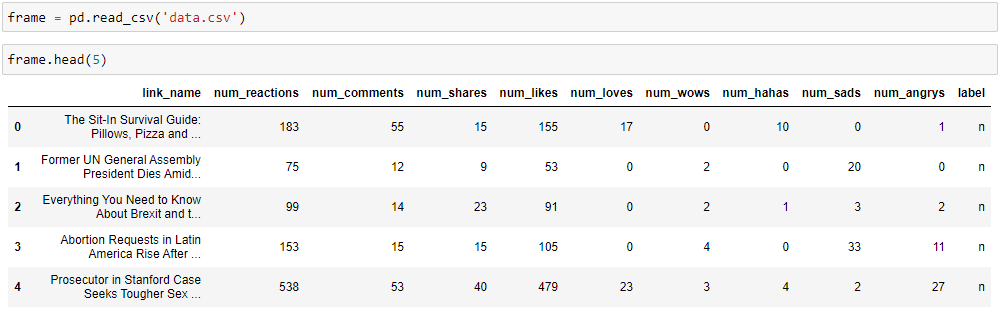
num\_hahas: total number of haha reacts on the post.

num\_sads: total number of sad reacts on the post.

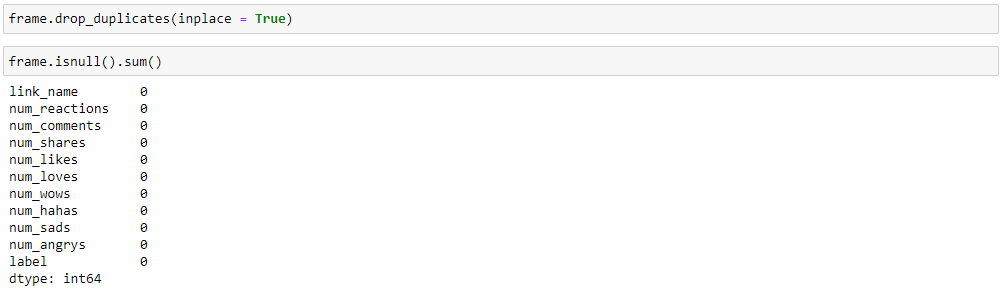
num\_angrys: total number of angry reacts on the post.

Label: denoting whether a post is fashion related or not using “n: for not fashion related” and “y: for fashion related”.

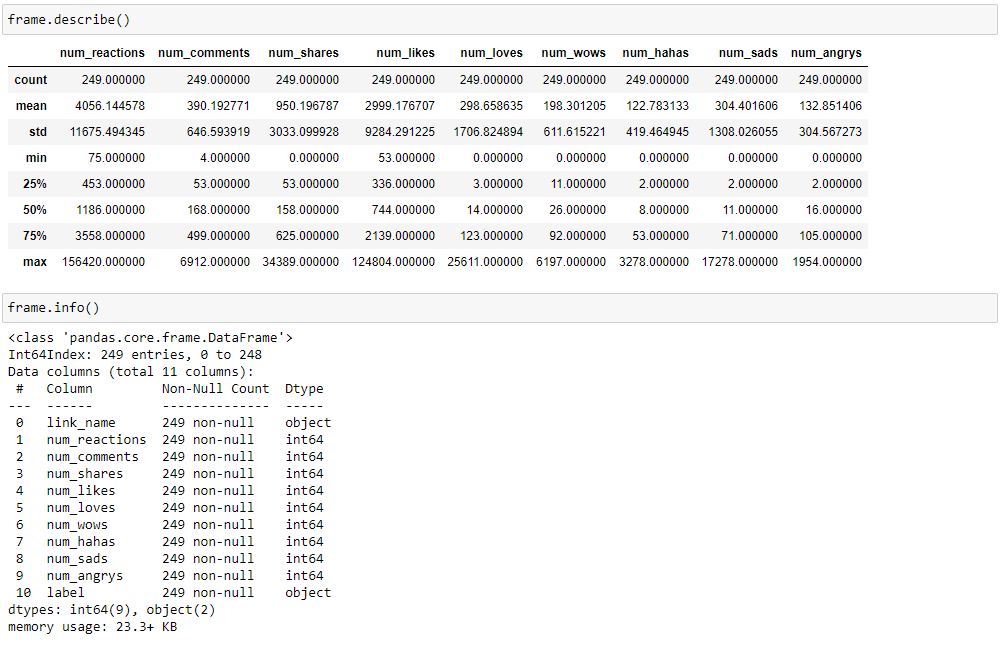
The dataset is divided using sklearn’s test-train-split function. The split ratio is 70:30 where 30% goes to test dataset and the rest goes to training set.



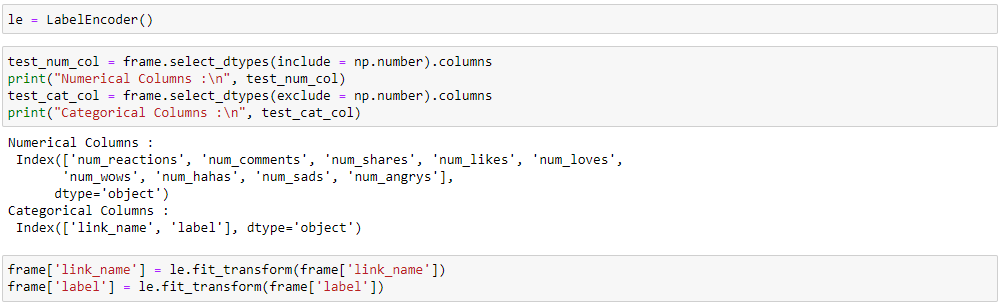
The figure above shows the code responsible for importing the dataset into the code. This is verified by the first five instances of the dataset shown in the figure.



The code above shows the duplicates and the null instances being removed from the dataset. The code below shows the statistical analysis on the dataset. Central tendency measures are used to understand the dataset. The metadata about the dataset is also displayed in the figure below that helps study about the features in the dataset.



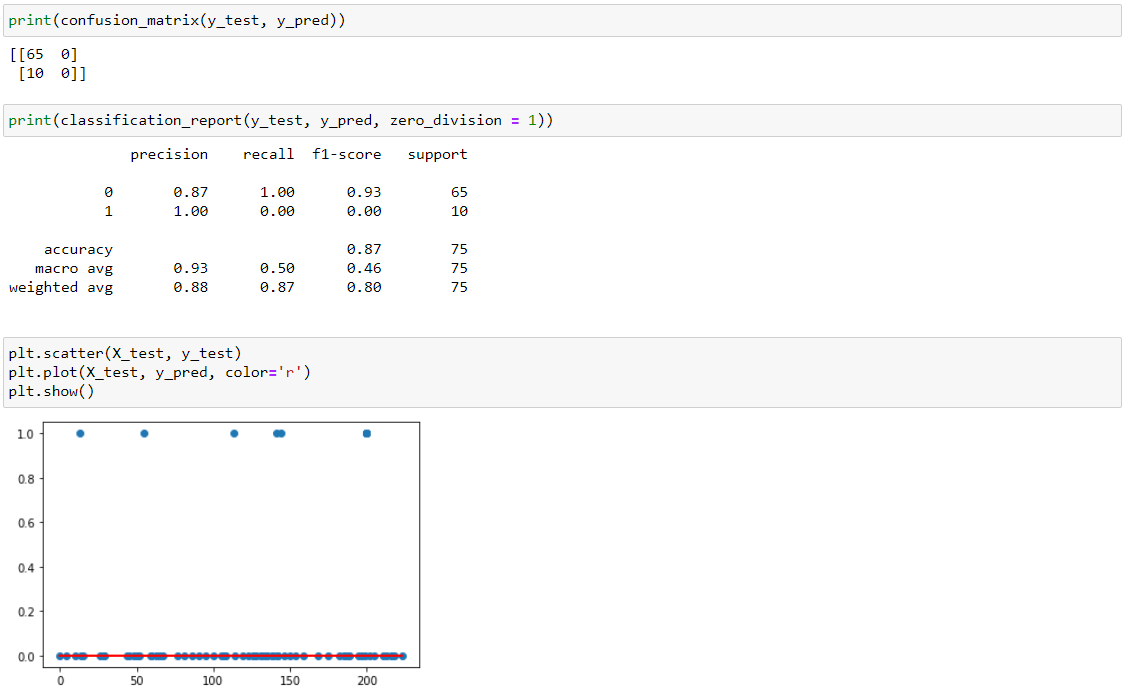
The code below is responsible for the Labe Encoding section of data pre-processing which is critical for training and testing the algorithm.



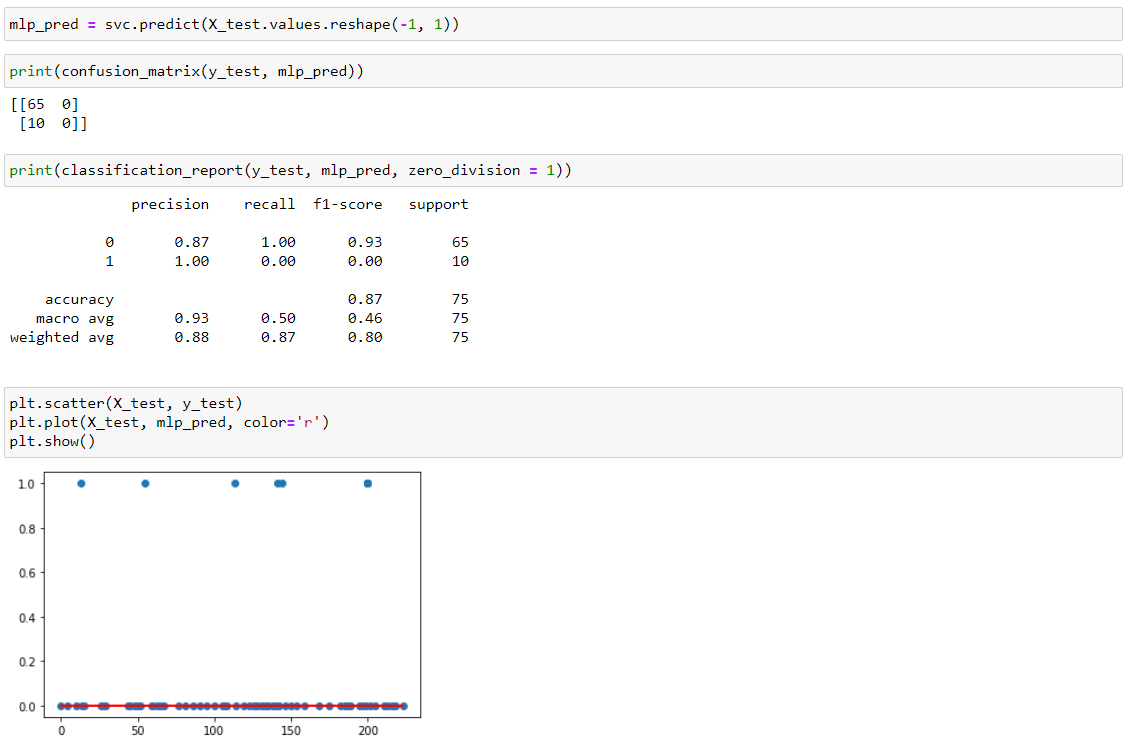
This completes the data pre-processing section.

# Machine Learning

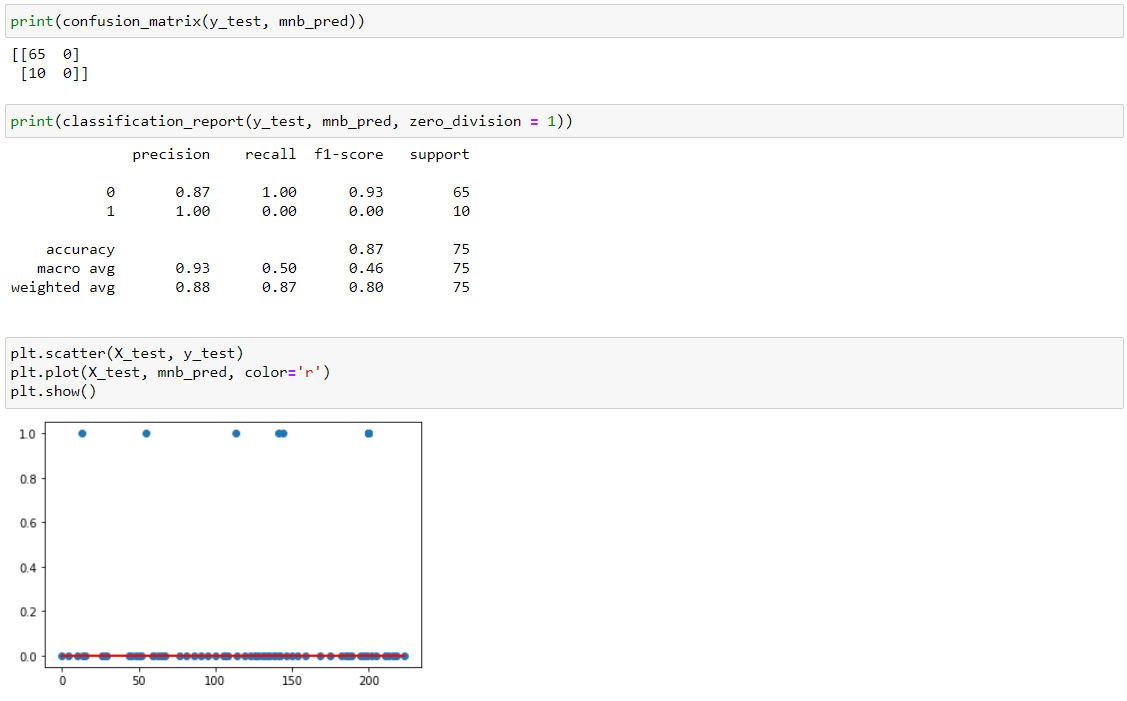
The paradigm for this research in coding would be classification. The algorithms targeted for this research are SVM, Multinomial Naïve Bayes and Multilayer perceptron. The performance of the algorithms is measured using 2-fold cross validation and other classification performance metrics.



The above image shows the performance of the SVM algorithm. The blue scattered dots in the graph above is the dataset and the red line shows the fitting of the SVM algorithm on the dataset.



The above image shows the performance of the MLP algorithm. The blue scattered dots in the graph above is the dataset and the red line shows the fitting of the MLP algorithm on the dataset.



The above image shows the performance of the Multinomial NB algorithm. The blue scattered dots in the graph above is the dataset and the red line shows the fitting of the Multinomial NB algorithm on the dataset.