|  |
| --- |
| ML MAJOR PROJECT  MAY-2020 |
|  |
| July 23  SUMMARY OF MAJOR PROJECT  By: G.L.K RAGHAVENDRA VARMA |





# SUMMARY OF MAJOR PROJECT :

The given problem statement is:

Problem Statement : For a given dataset (problem) which is the best classification algorithm (as per accuracy)?

So as per the statement we took the dataset provided in the mail. In this dataset, “gender” is the dependent variable.

We first cleaned the data by removing unnecessary columns. Then we categorized gender into male, female, brand and unknown and found out the number users under each category. Then we used feature engineering on link bar and side bar colors.

With that data, we also raised two questions and answered them. They are…

**1. What are the most used words by each gender category on the tweets posted?**

a) A. Males:

1. just

2. like

3. don't

B. Females:

1. just

2. like

3. love

C. Brands:

1. weather

2. channel

3. updates

2. Which colour is most used for links and sidebars by each gender category?

A. Links:

1. Male - Olive

2. Female - Amethyst

3. Brand - Curoius Blue

B. Sidebars:

1. Male – Wood smoke

2. Female – Trade wind

3. Brand - Algae Green

* Coming to ensemble learning, we used ‘logistic regression’, ‘svm’ and ‘knn’.
* Out of three, we found that “logistic regression” has more accuracy.
* In the program, we tried to find the accuracy of determining genders using link and sidebars as well as words in description and texts. In those two, description and texts gave more accurate output.