

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Major Project

|  |  |  |
| --- | --- | --- |
| ***Classroom Name:ML11B3*** | ***Ml-Major-Nov*** | ***09-01-2021*** |

# Project Report

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

Answer:

## SUMMARY:

1. A data about twitter statistics of twitter users was given which was studied by performing EDA and Visual EDA on it.
2. Further, the data was cleaned to get a better and accurate results on the data frame that we were working on.
3. As required 2 questions were framed on the given data and were answered by performing various actions.
4. Feature selection and Feature engineering was done.
5. Machine Learning modelling was ensembled using three different algorithms (Name 1, Name 2, Name 3)
6. At last, Accuracy calculation was done

**The details of the above summary can be seen below:**

## EDA and visual EDA

* 1. Rows and columns and other information about the data set were analyzed.
  2. Plenty graphs were plotted to analyze the data:
     1. Distplot
     2. Countplot
     3. Boxplot
     4. scatterplot

1. **Cleaning Data**
   1. Unwanted columns were deleted/dropped.
   2. Null values were removed or replaced.
   3. Columns with maximum NULL values were deleted
   4. Outliers were removed.
2. **Questions that were Framed and their Answers**
   1. **Question:** What are the average words used by male and female users in their description to describe themselves?
      1. **Answer:** On an average 13.616104181951577 words are used by male users and 12.093870402802102 words are used by female users.
   2. **Question:** On an average how much tweets to both the genders do?
      1. **Answer:** Male users have 33879.69754218635 whereas female users have 27527.38721541156 tweets on an average
3. **Feature selection and Feature engineering**

4.1 figuring out features that should make an impact on the prediction by what they showed

4.2 studying their values(many columns had only one value , so they don’t really affect the predictions

4.3 now studying the final columns and removing values weren’t possible (them being too low or too high)

1. **Machine Learning modeling**

5.1 We have used 3 models they are:

5.1.1 Logistic Regression

5.1.2 Decision tree

5.1.3 K Nearest Neighbors

5.2 Decision Tree obtained highest accuracy and least accuracy was obtained by accuracy

1. **Ensemble Modeling**

6.1 Simple Ensemble techniques:

1. Max Voting

2. Averaging

3. Weighted Averaging

* 1. 6.2 We have used first method i.e Max voting

### Team members:

* Jois Prabhanjan
* Atishaya Jain
* Chaitanya Apate
* Shashank K S
* Siddharth
* Akshat Parashar
* sirisha K
* Sirisha Kodukula
* Sai Kumar
* AlekhyaNaravajhula
* Sraddha Reddy
* erukullasaitharun
* Akash Singh
* Kalyani Vidhate
* Meghana Yelleti

Some of the specific works are listed below :

\*) Prabhanjan Jois :- Ml algorithms , Data visualization and Ensymbel modeling.

\*) Siddharth Thewani :- Data preprocessing.

\*) Atishaya Jain :- Ml algorithms , Questions to the dataset.

\*) Akshat parashar :- Report Preparation.

[Everything was done on the dataset provided to us, no other dataset was selected]