

**Malignant Comment Prediction**

**Submitted By: -**

Pankaj Pagare

# ACKNOWLEDGMENT

I would like to express my gratitude towards Flip Robo Technologies for their kind co- operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons and my mentor (Ms. Swati Mahaseth) for giving me such attention and time as and whenever required.

# INTRODUCTION

## Business Problem Statement

* The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users.
* Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection. Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.
* There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

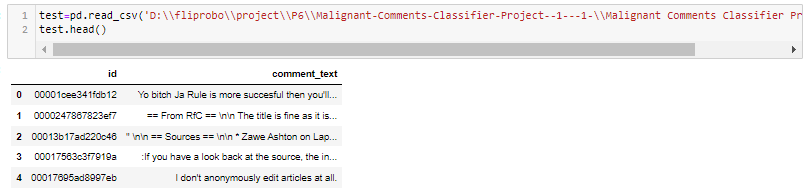
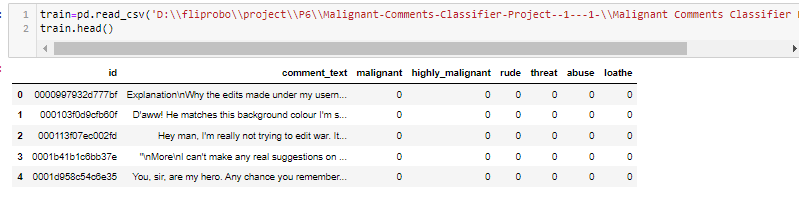
## Conceptual Background of the Domain Problem

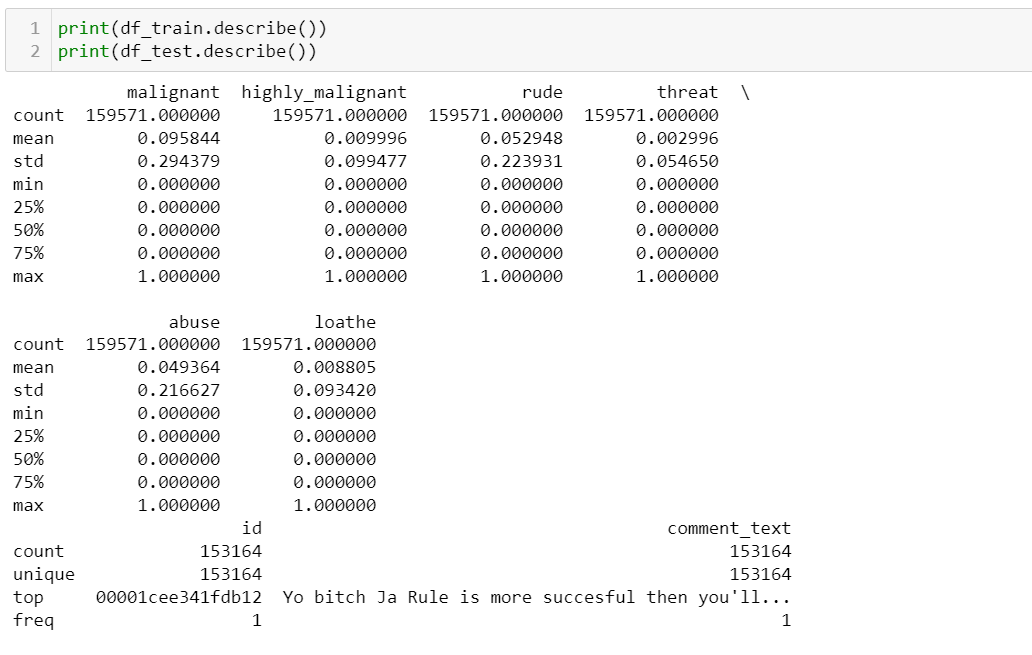
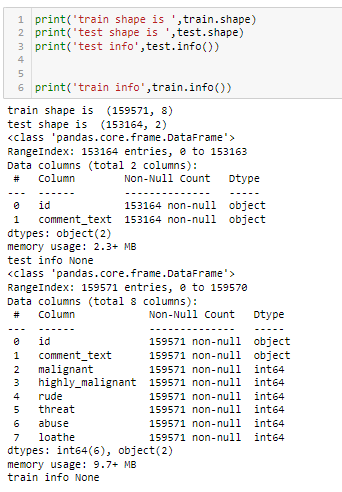
## Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

## Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

# Analytical Problem Framing

Dataset Representation:





**So clearly it is a Multi Class classification problem.**

**In *multi-class classification*, we have one basic assumption that our data can belong to only one label out of all the labels we have.**

## Data Sources and their formats & inferences

* Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
* Highly Malignant: It denotes comments that are highly malignant and hurtful.
* Rude: It denotes comments that are very rude and offensive.
* Threat: It contains indication of the comments that are giving any threat to someone.
* Abuse: It is for comments that are abusive in nature.
* Loathe: It describes the comments which are hateful and loathing in nature.
* ID: It includes unique Ids associated with each comment text given.
* Comment text: This column contains the comments extracted from various social media platforms.

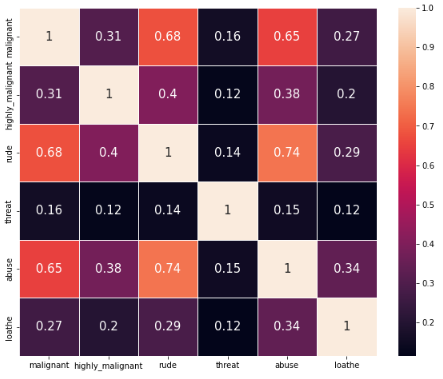
**Observation:**

* 2 Independent variables with 6 target variables.

* From the dataset we can infer that it is clearly classification problem.

## Data Inputs- Logic- Output Relationships

### Correlation :

****

#### Observation:

## Malignant, rude and abusive words are highly correlated words for my dataset.

## Assumption for the problem:

From the dataset we can infer that it is clearly a Classification problem.

* Hardware and Software Requirements and Tools Used

Software Used:

* Jupyter Notebook
* Ms-Paint
* MS-PowerPoint
* MS-Word

Hardware used:

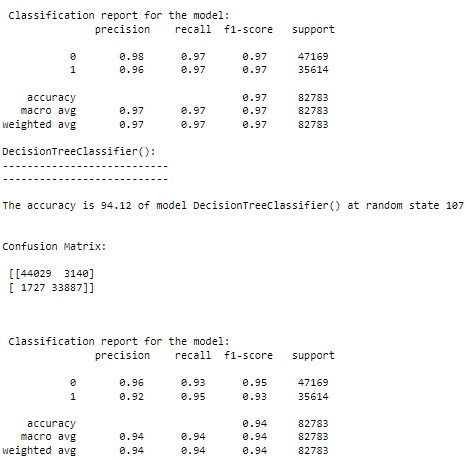
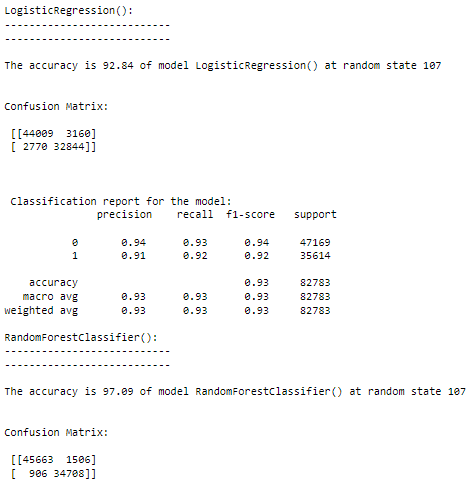
* Laptop
* Good internet connectivity

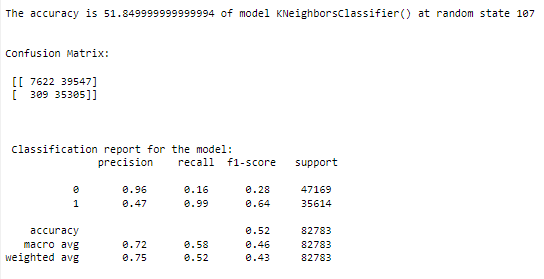
# Model/s Development and Evaluation

## Testing of Identified Approaches (Algorithms)

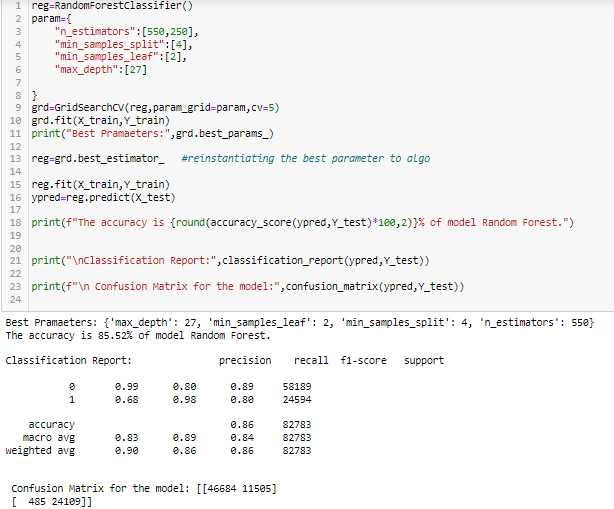
* LogisticRegression()
* DecisionTreeClassifier()
* KNeighborsClassifier()
* RandomForestClassifier()

## Running the selected Models:



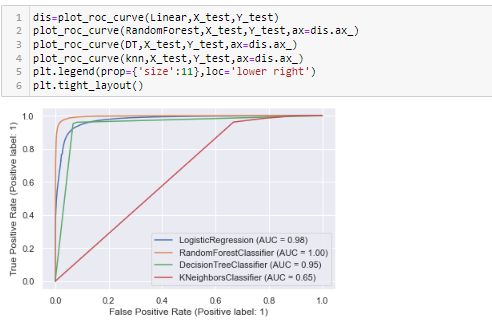


HYPERTUNED THE BEST MODEL



### OBSERVATION:

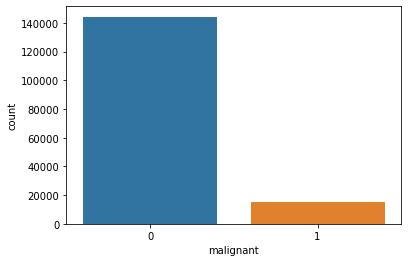
* Finally we have saved the Random Forest Classifier Model.



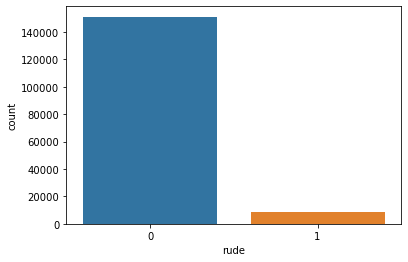
Here for model Random Forest we get the highest AUC score. So, we finalized Random Forest model.

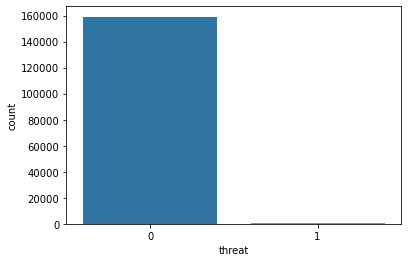
# Visualizations

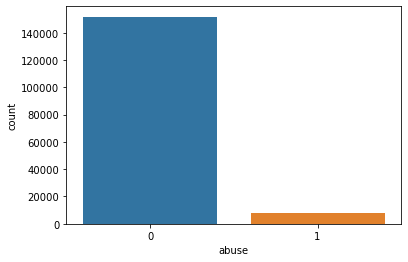
## Univariate Analysis of categorical variables:

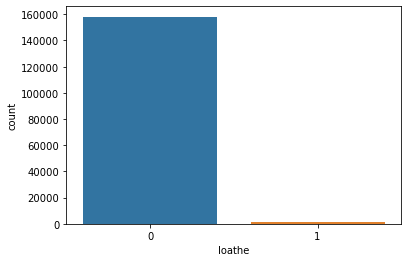
****

#### 

****



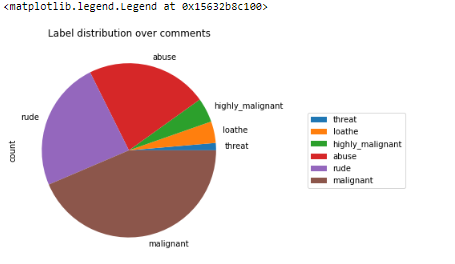




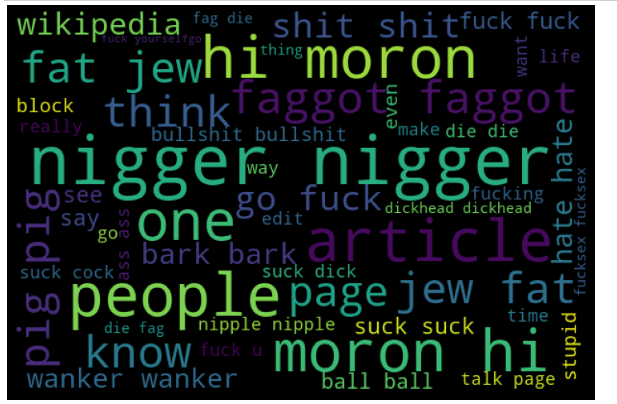
**Observation :**

* The malignant,highly malignant and rude data are highly imbalaced and seems that there are not much comments with loud data.
* Malignant data are most as compared to rude and highly malignant data.
* Similarly for threat,abuse and loathe also data are highly imbalaced and seems that there are not much comments with loud data.
* Abusive data are more than threat or loathe.
* So, we did oversampling as we concluded that data is imbalanced.





* Here we can clearly see the amount of comment we got for each type in the dataset.



* Here you can see the words for the malignant types.

**Interpretation of the Results**

* + Hence we can go with normal Random Forest Classifier model with is also giving good accuracy.
  + This model may help to solve real life problems by identifying Harsh ,hurting and loud words and preventing it hurting to the readers.

# CONCLUSION

Finally, we have saved the Random Forest Classifier Model and hyper tuned but accuracy got reduced So we save Random Forest with default parameters.

## Learning Outcomes of the Study in respect of Data Science

* Hence we can go with normal Random Forest Classifier model which is also giving good accuracy.
* Finally we have saved the Random Forest Regressor Model.

## Limitations of this work and Scope for Future Work

* In future this machine learning model using NLP may bind with various website which can provide real time data for malignant comment prediction.
* This project predicts the loud words and may predict it reaching to their user so that they are not hurt also build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.
* It may avoid the readers to spread negativity with their harsh words.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*