

**MALIGNANT COMMENTS CLASSIFICATION**

Submitted by:

Ankita Prashant Chaudhari

**ACKNOWLEDGMENT**

[1] Macleans.ca. 2020. Online Hate Speech In Canada Is Up 600 Percent. What Can Be Done? - Macleans.Ca. [online] Available at: [Accessed 28 October 2020].

[2] R. Hatzipanagos, “Perspective | How online hate turns into real-life violence,” The Washington Post, 30-Nov-2018. [Online]. Available: https://www.washingtonpost.com/nation/2018/11/30/how-online-hate-speech-is-fueling-real-life-violence /. [Accessed: 28-Oct-2020].

[3] “Court documents identify 13 injured in deadly van attack | CBC News,” CBCnews, 25-Apr-2018. [Online]. Available: https://www.cbc.ca/news/canada/toronto/injured-van-attack-1.4633308. [Accessed: 28-Oct-2020].

[4] “Potentials of NLP: Techniques, Industry-Implementation and Global Market Outline,” Analytics Insight, 26-Jan-2020. [Online]. Available: https://www.analyticsinsight.net/potentials-of-nlp-techniques-industry-implementation-and-global-market -outline/. [Accessed: 28-Oct-2020].

**INTRODUCTION**

**Business Problem Framing**

The background for the problem originates from the multitude of online forums, where-in people participate actively and make comments. As the comments sometimes may be abusive, insulting or even hate-based, it becomes the responsibility of the hosting organizations to ensure that these conversations are not of negative type. The task was thus to build a model which could make prediction to classify the comments into various categories.

**Conceptual Background of the Domain Problem**

Given a group of sentences or paragraphs, used as a comment by a user in an online platform, classify it to belong to one or more of the following categories id', 'comment\_text', 'malignant', 'highly\_malignant', 'rude', 'threat','abuse', 'loathe'],

dtype='object' with either approximate probabilities or discrete values (0/1).

**Review of Literature**

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 cyber bullying.

.

**Motivation for the Problem Undertaken**

Harmful or toxic comments in the social media space have many negative impacts to society. The ability to readily and accurately identify comments as toxic could provide many benefits while mitigating the harm. Also, our research has shown the capability of readily available algorithms to be employed in such a way to address this challenge.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

**1.What is Statistical Modeling and How is it Used?**

Statistical modelingis the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

When [data analysts](https://www.northeastern.edu/graduate/blog/what-does-a-data-analyst-do/) apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically. Rather than sifting through the raw data, this practice allows them to identify relationships between variables, [make predictions](https://www.northeastern.edu/graduate/blog/predictive-analytics/) about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.

“When you analyze data, you are looking for patterns,” says Mello. “You are using a sample to make an inference about the whole.”

## Important Statistical Techniques in Data Analysis

Before any statistical model can be created, an analyst needs to collect or fetch the data on a database, clouds, social media, or within a plain excel file. To do this, analysts must also have a solid grasp of data structure and management, including how and where data is stored, fetched, and maintained. Those working in this field should thus share a passion for facts and data, and understand the basics of data manipulation, as well.

Once it comes time to analyze the data, there are an array of statistical models analysts may choose to utilize. According to Mello, most common techniques will fall into the following two groups:

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### Regression Models

Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

“The most traditional regression models that have been used for a long time are logistic regression, linear regression, and polynomial regression,” Mello says. “These are the most common.”

Other examples of regression models can include stepwise regression, ridge regression, lasso regression, and elastic net regression.

### Classification Models

**Classification** is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

“Classification models are a form of supervised machine learning which is often used when the analyst needs to understand how they got to a certain point,” Mello says. “They give you more than just an output; [they give you] more information that you can use to explain the results of the prediction to your boss or stakeholder.”

Some of the most common classification models include decision trees, random forests, nearest neighbor, and  Naive Bayes.

There are also the neural networking models that are more used in AI. “These are very powerful models, and they can make accurate predictions very well,” Mello says, “but you typically cannot explain what is happening behind the scenes.”

**Types of the Model:**

Here it is the Classification modeling technique use because in this dataset loathe is target variable and in this columns there are only two values 0 and 1 values are present so use classification models.

**Problem Definition:**

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.

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.

**Data Set Description**

The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes ‘Id’, ‘Comments’, ‘Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

The data set includes:

* **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.

**2. Data Analysis:**

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making. There are several methods and techniques to perform analysis depending on the industry and the aim of the analysis.

**DataSet:**

Classification analysis is a data analysis task within [data-mining,](https://indstaging.wpengine.com/defined/data-mining/) that identifies and assigns categories to a collection of data to allow for more accurate analysis. The classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

Classification analysis can be used to question, make a decision, or predict behavior through the use of an algorithm. It works by developing a set of training data which contains a certain set of attributes as well as the likely outcome. The job of the classification algorithm is to discover how that set of attributes reaches its conclusion.

There are two steps in the construction of a classification model.

* **Learning Step** – this is where different algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.
* **Classification Step:** this is where the model used to predict class labels, tests the constructed model on test data. Which in turn estimates the accuracy of the classification rules.

**Types of classification model:**

* Logistic Regression
* Naïve Bayes
* Stochastic Gradient Descent
* K-Nearest Neighbours
* Decision Tree
* Random Forest
* Support Vector Machine

**Introduction**

Online forums and social media platforms have provided individuals with the means to put forward their thoughts and freely express their opinion on various issues and incidents. In some cases, these online comments contain explicit language which may hurt the readers. Comments containing explicit language can be classified into myriad categories such as [‘id', 'comment\_text', 'malignant', 'highly\_malignant', 'rude', 'threat','abuse', 'loathe'],. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions.

To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying.

**Objective**:

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 cyber bullying.

**Data Loading and Visualisations:**

The first and foremost step involves importing necessary libraries and packages and loading the dataset as a pandas dataframe. Data visualization is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](https://www.tableau.com/learn/articles/data-visualization/glossary), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Our eyes are [drawn to colors and patterns](https://www.tableau.com/learn/whitepapers/tableau-visual-guidebook). We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we [quickly see trends and outliers](https://www.tableau.com/reports/business-intelligence-trends). If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

**Importing libraries**

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, and Seaborn.

### Reading data:

We will now read the data from a CSV file into a Pandas DataFrame in this there are two dataset first is train dataset and second is test dataset.

Install This project requires anaconda python, because below libraries already available.

• Numpy

• Matplotlib

• Seaborn

• Skit-learn

• Pandas

Also need to have software.

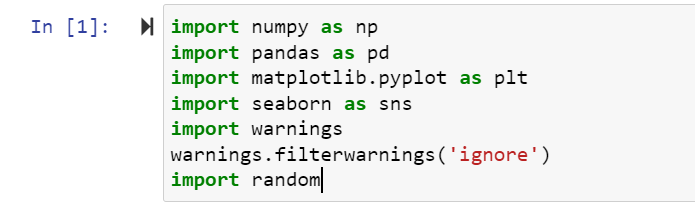
Install, run and execute a Jupyter notebook.

**EDA**:

There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

## The Importance of Exploratory Data Analysis (EDA): There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

**Loading dataset :**

**** ****

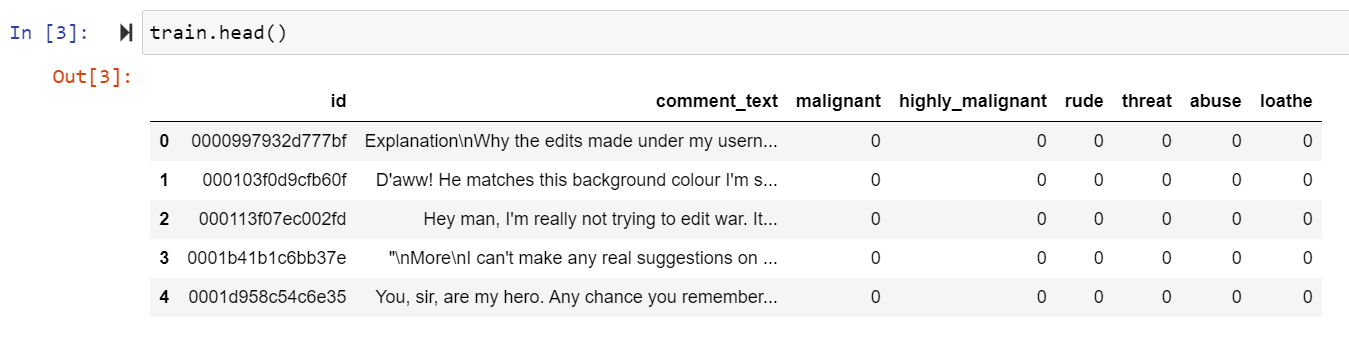
Here importing all necessary library and also load Train and Test dataset.

**Data Sources and their formats**

Then need to read and load house dataset. And Then this below code isusefor

display all the columns in dataset. Let us have a look at how our dataset The structure and details of the data are given below in this there are dataset in which no of row and columns are present:

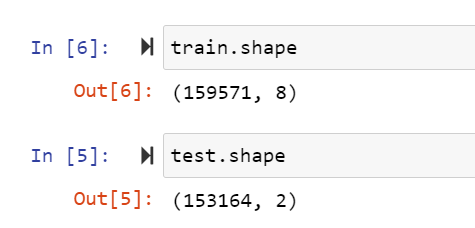
**train.head()**



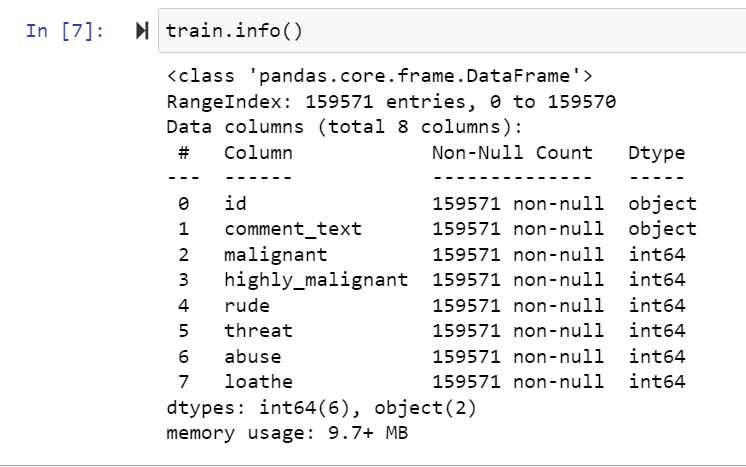
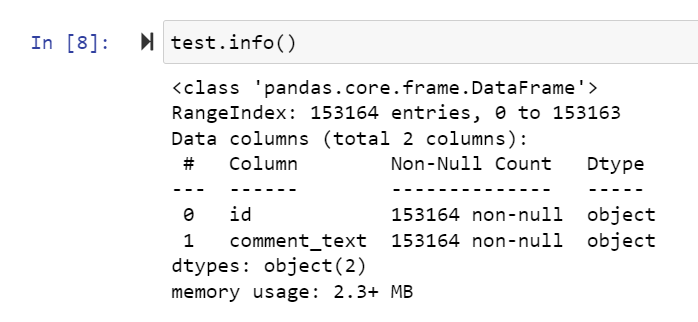
**test.head()**



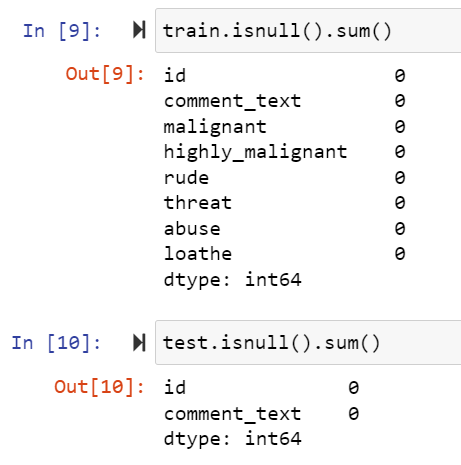
Here loathe is our Target Variable in Train dataset and in Test dataset there was no target variable and in Train dataset Loathe are target variable and there was 2 numeric values 0 and 1 are present so this is **Classification Problem** so Need use Classification Algorithm.



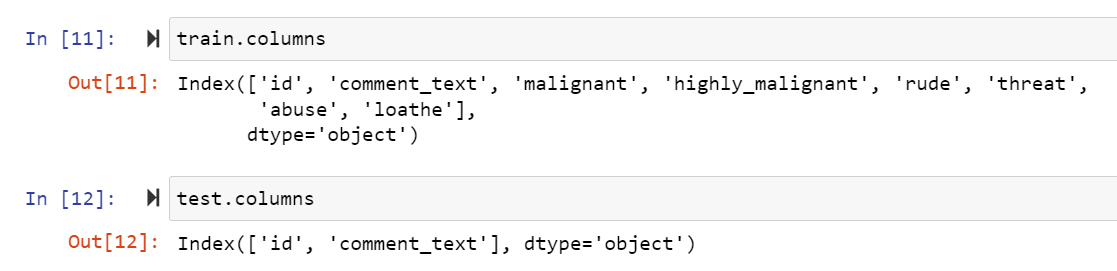
Check the how much Columns and rows present in dataset using df.shape() is use to check the rows and columns count in this above dataset there are 159571 rows and 8 columns are present in train dataset and 153164 rows and 2 columns are present in test dataset.

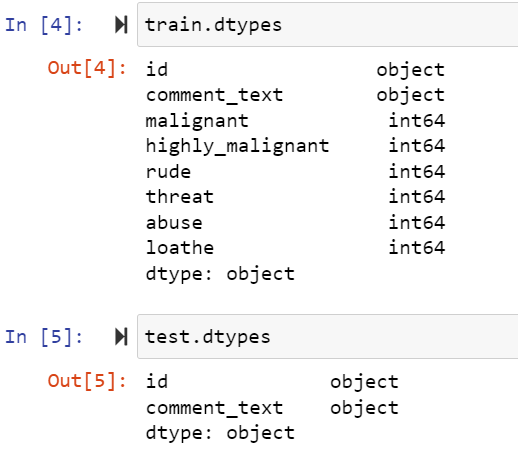
In above using df.info() is used to get the information all the columns in dataset and its data types. In above dataset there are 8 columns in train dataset and 2 columns are in test dataset.



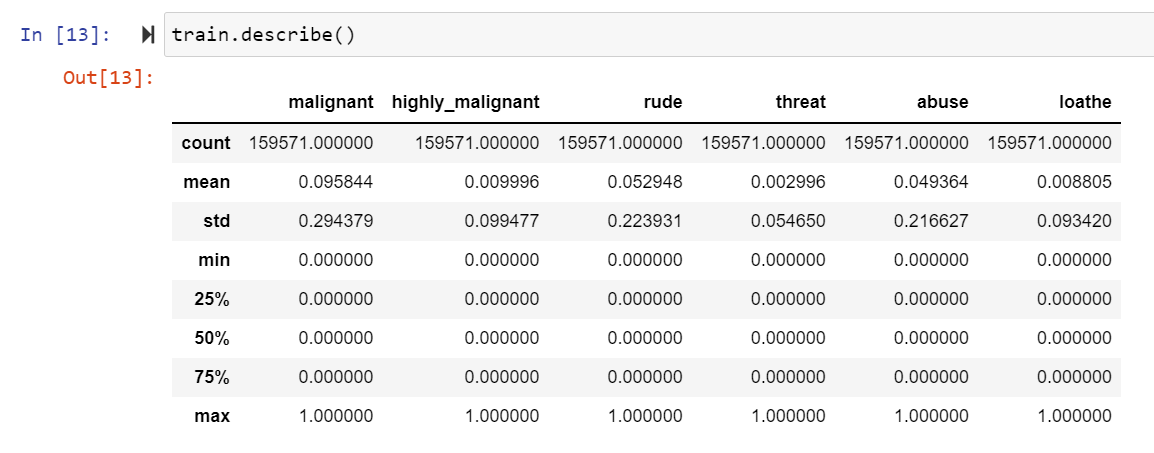
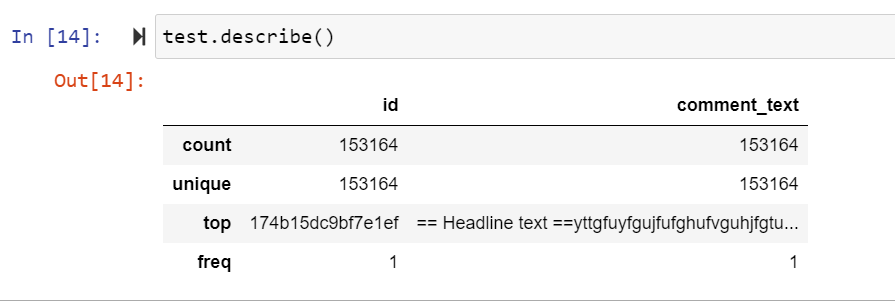
Check the how much missing values present in dataset using df.isnull() is use to check the missing values in dataset in this train dataset there was no missing values missing values are present and no missing are present in test dataset.



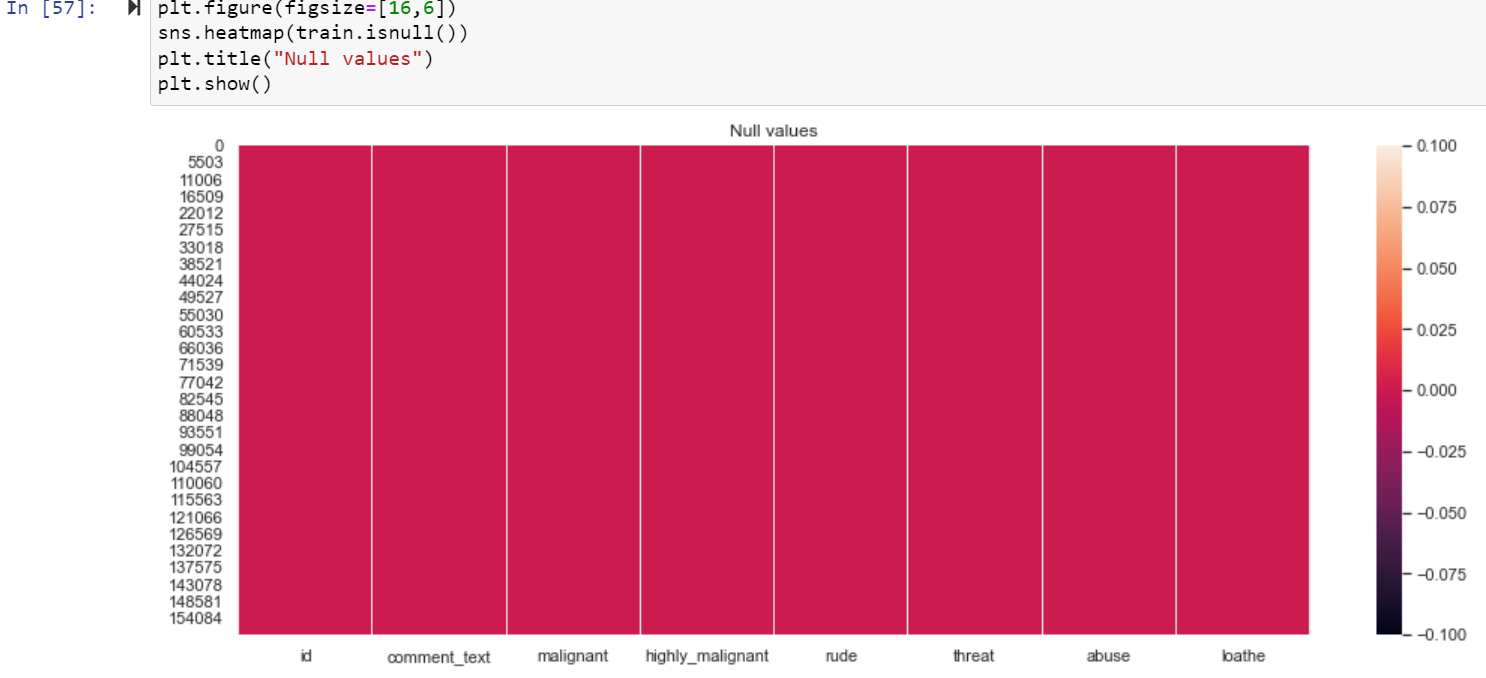
In above df.colums is use to get all the column names in train dataset and test dataset.



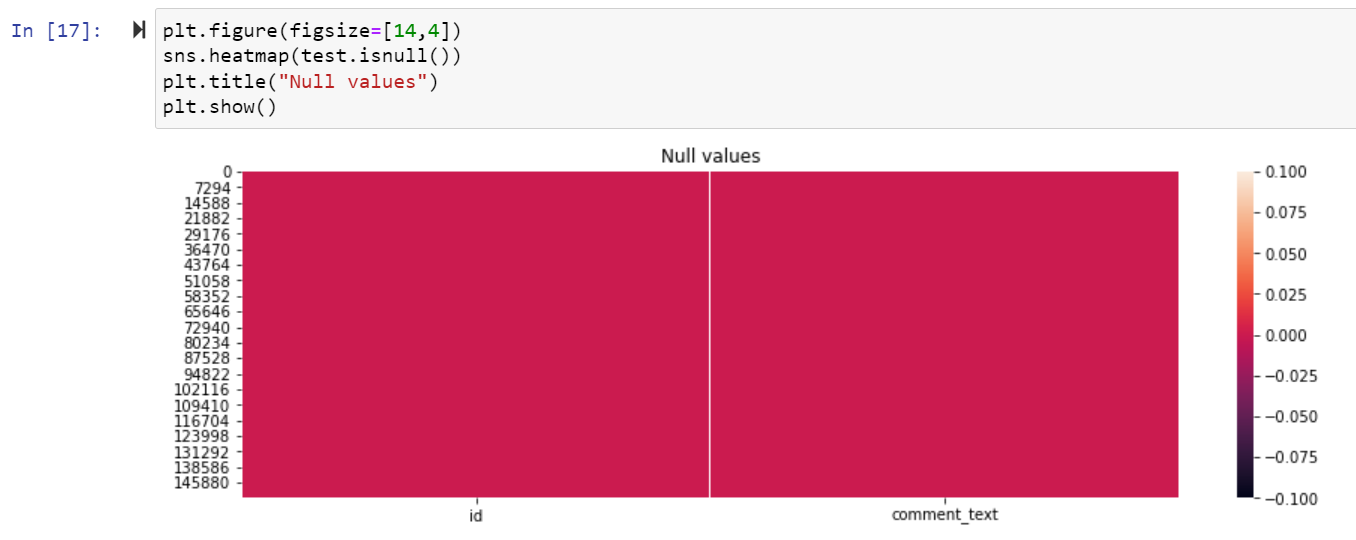
In above code df.dtypes is used to check the data types of all the column In above train and test dataset.

Above describe the information of train and test dataset using df.describe().

****

Graphically checking null values in train dataset there was no null values are present.



Graphically checking null values in test dataset there was no null values are present.

**Data Preprocessing Done:**

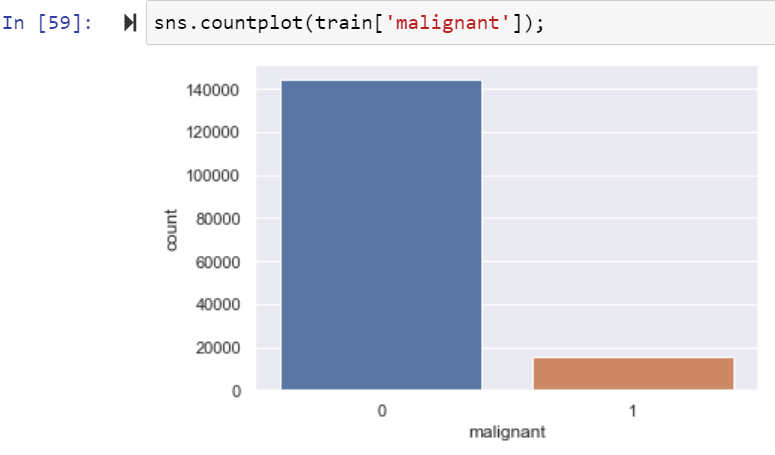
**Graphical representation**

### We will start with Univariate Analysis. We will be using a bar graph for this purpose.

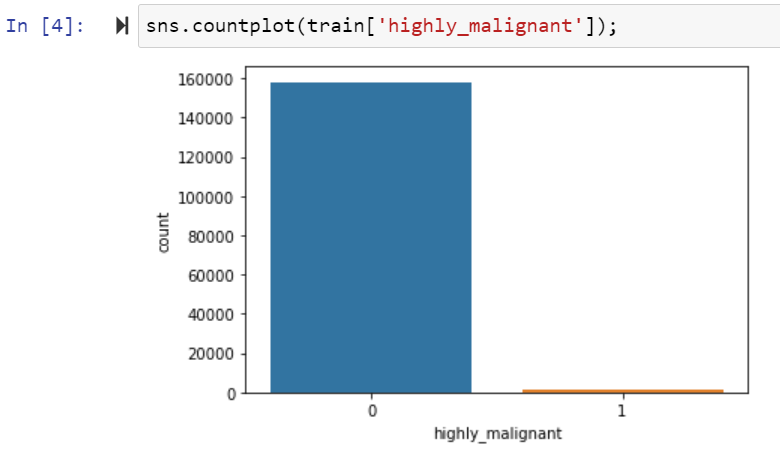
1. Univarient Analysis.
2. Bi Varient Analysis
3. Multivarient Analysis

**Univarient Analysis of Train dataset**:

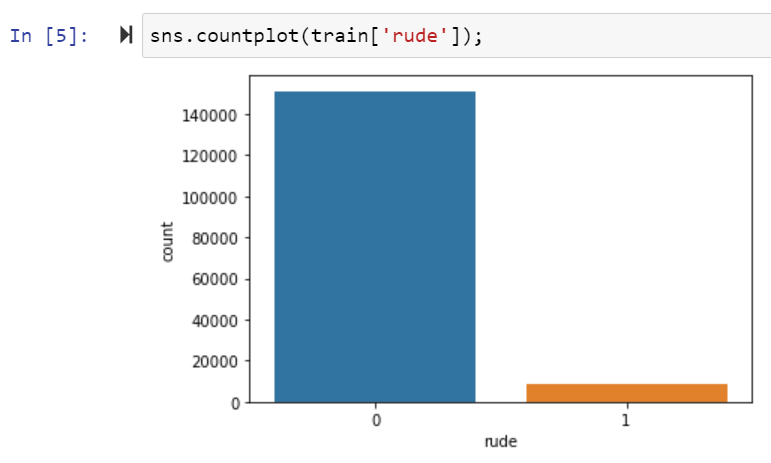
Univariate analysis **explores each variable in a data set**, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.\



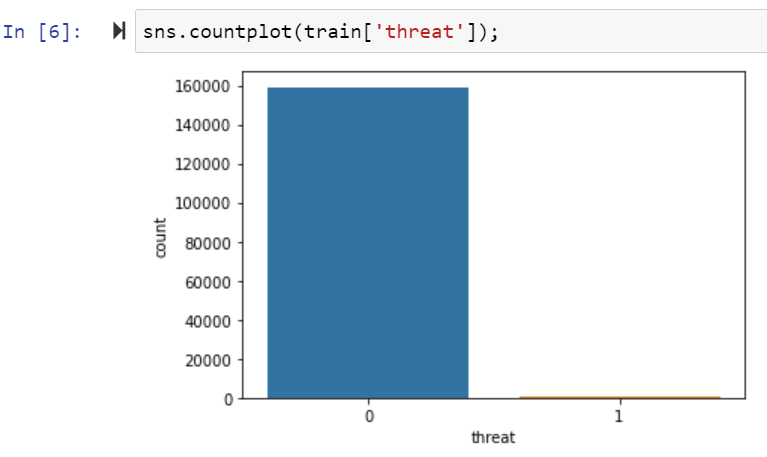
Univarient analysis of ‘malignant’ column in train dataset.



Univarient analysis of ‘highly\_malignant’ column in train dataset.



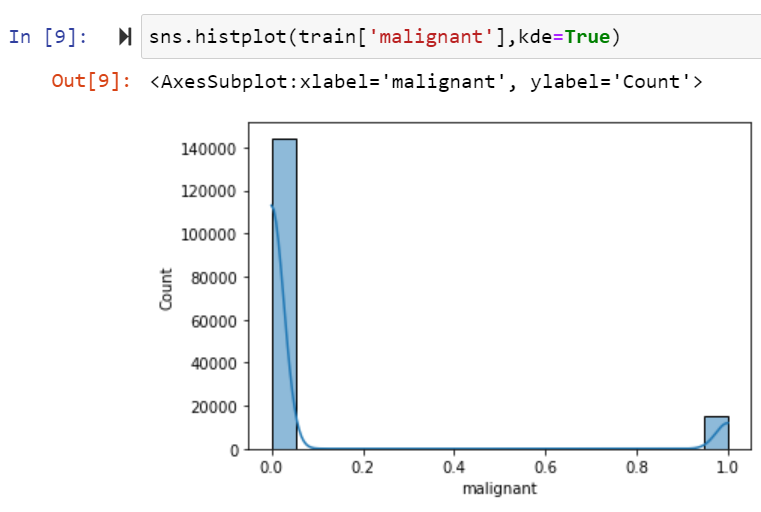
Univarient analysis of ‘rude column in train dataset.

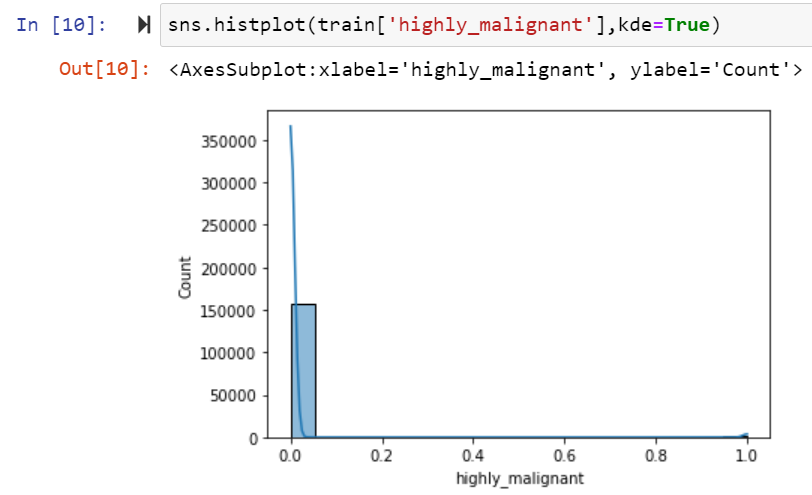


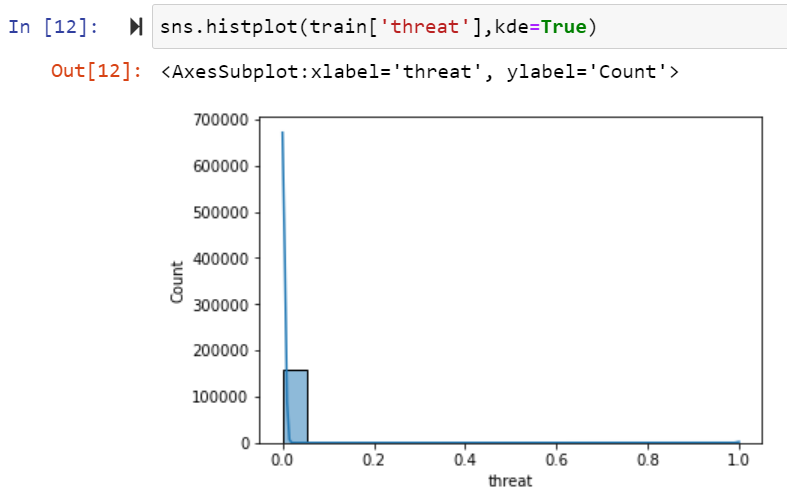
Univarient analysis of ‘threat’ column in train dataset.

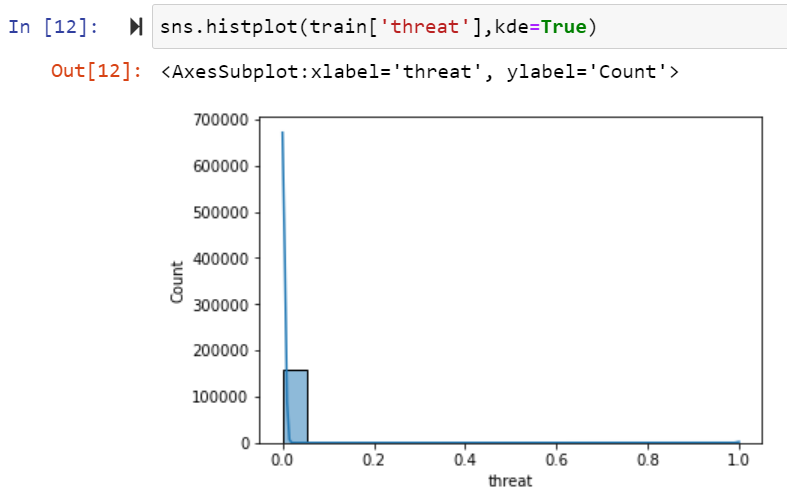
**Checking Distribution of Train Dataset:**

We will now make a **distribution plot** of the math score of the students. A distribution plot tells us how the data is distributed. We will use the distplot function.

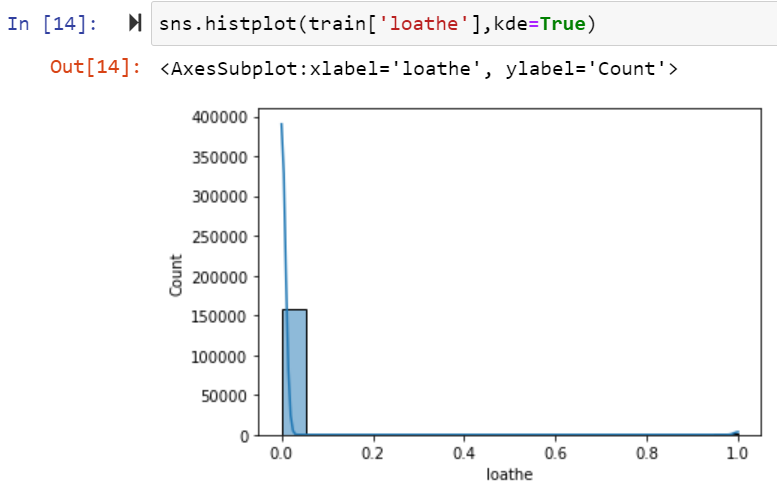




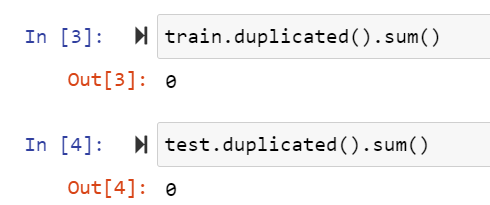




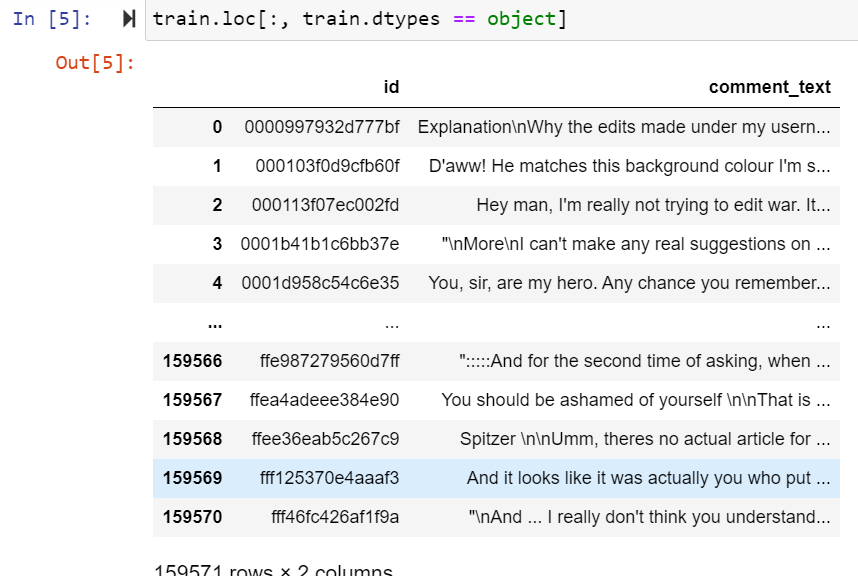




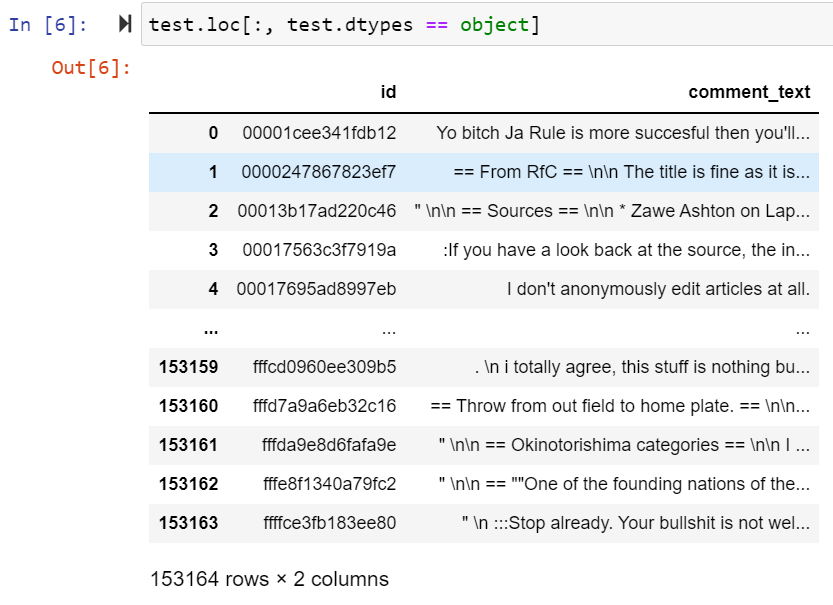
Here data columns are not normally distributed.



Here checking duplicate values in train and test dataset.



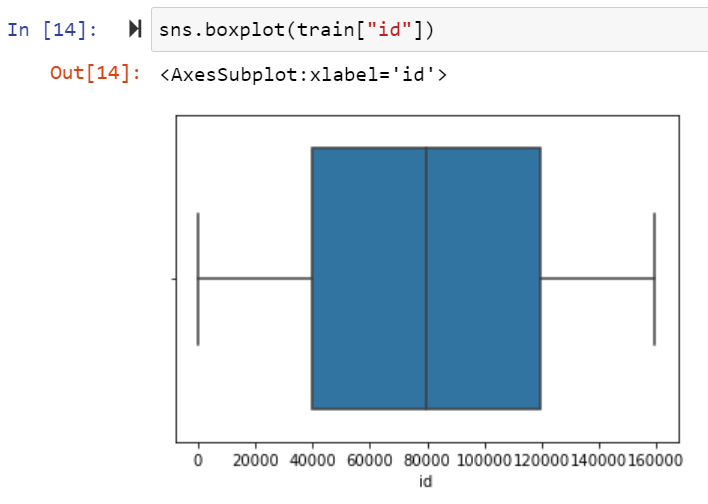
Here above are the object data types columns.

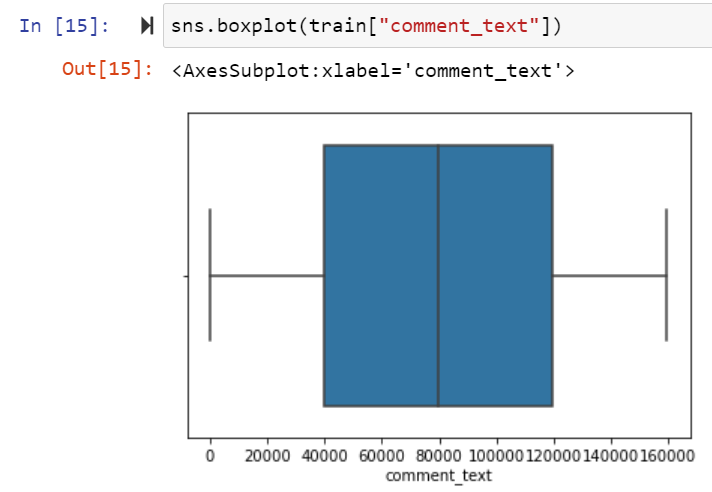


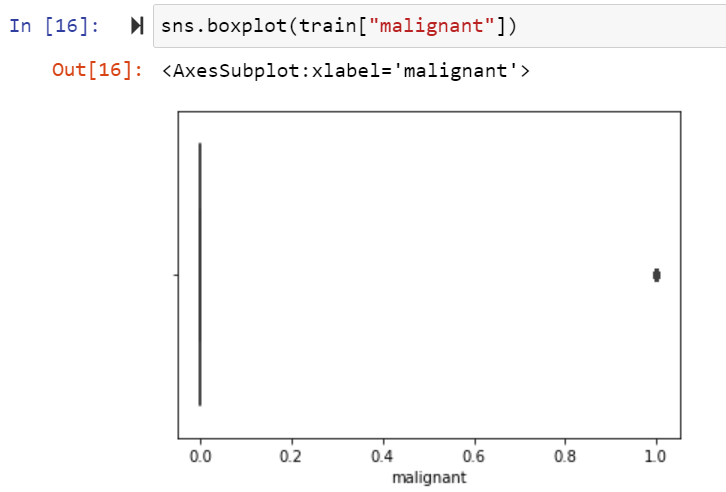
Here above are the object data types columns.

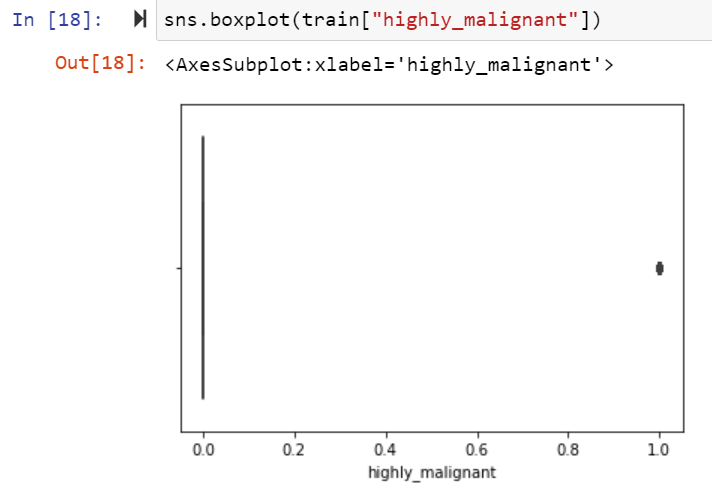
**Checking Outliers of Train Dataset:**

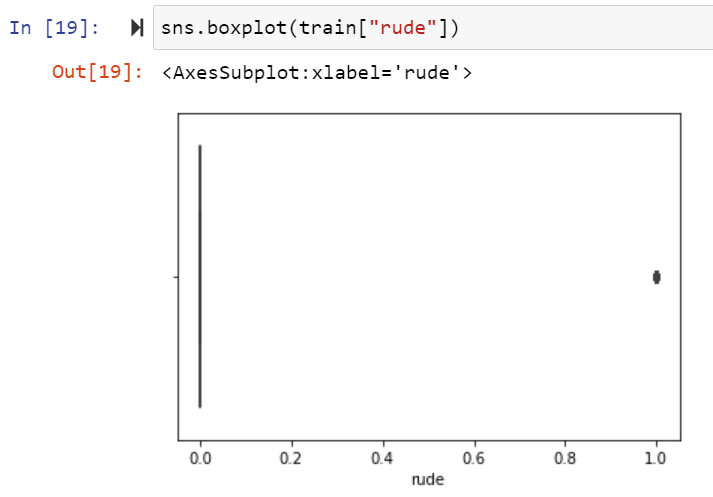
An outlier is an **object(s) that deviates significantly from the rest of the object collection**. It is an abnormal observation during the Data Analysis stage, that data point lies far away from other values. An outlier is an observation that diverges from well-structured data.



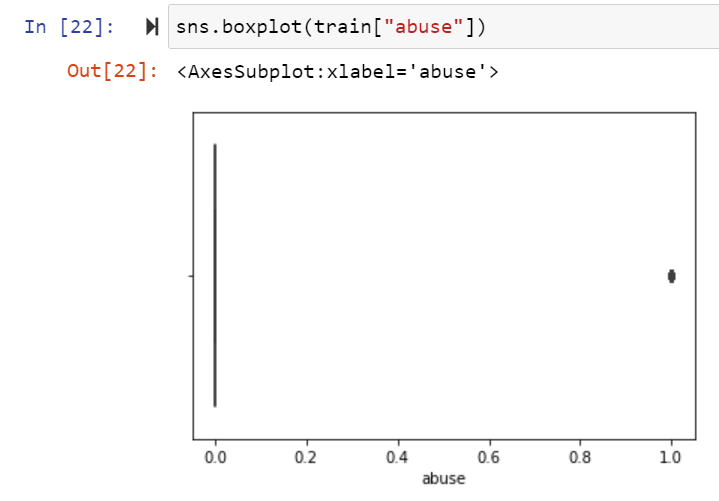


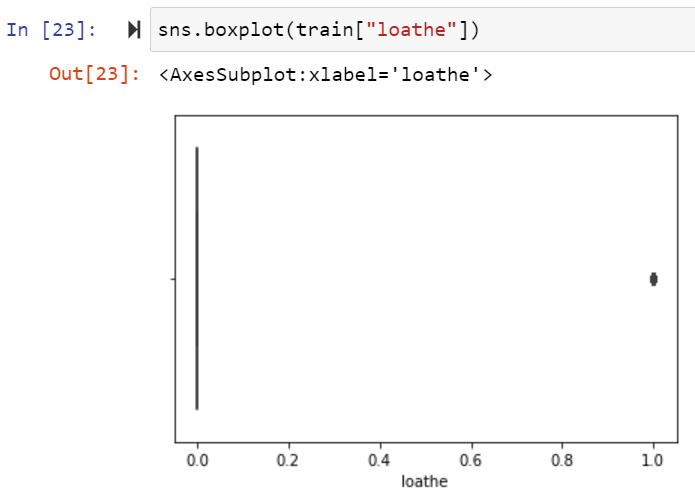
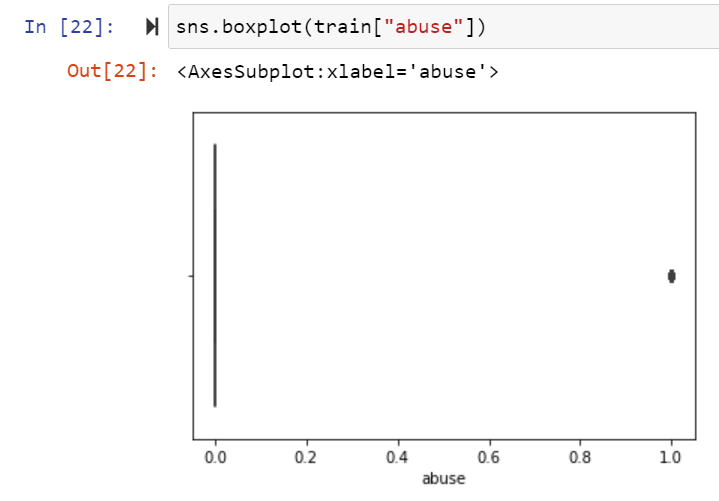


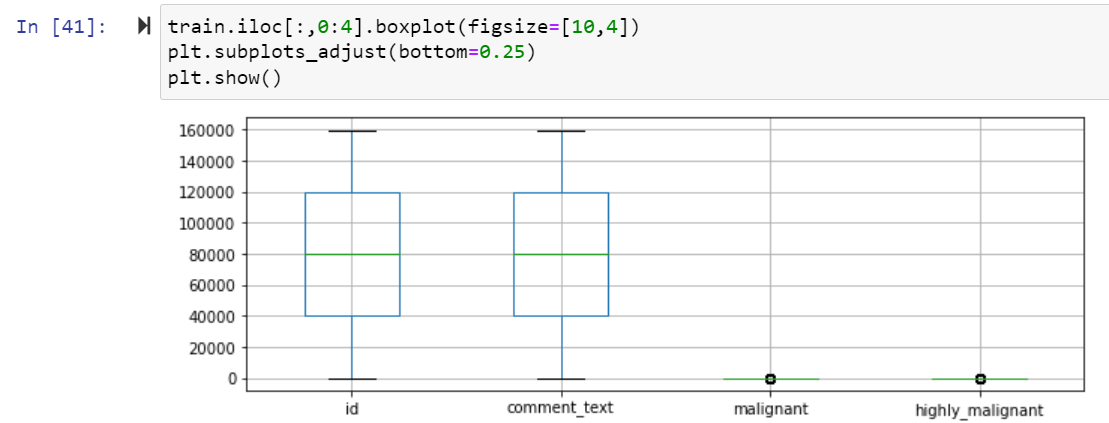


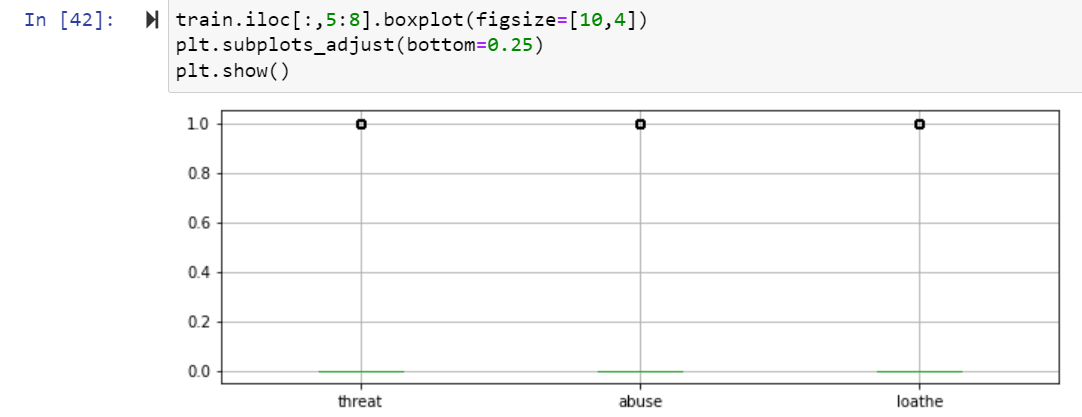








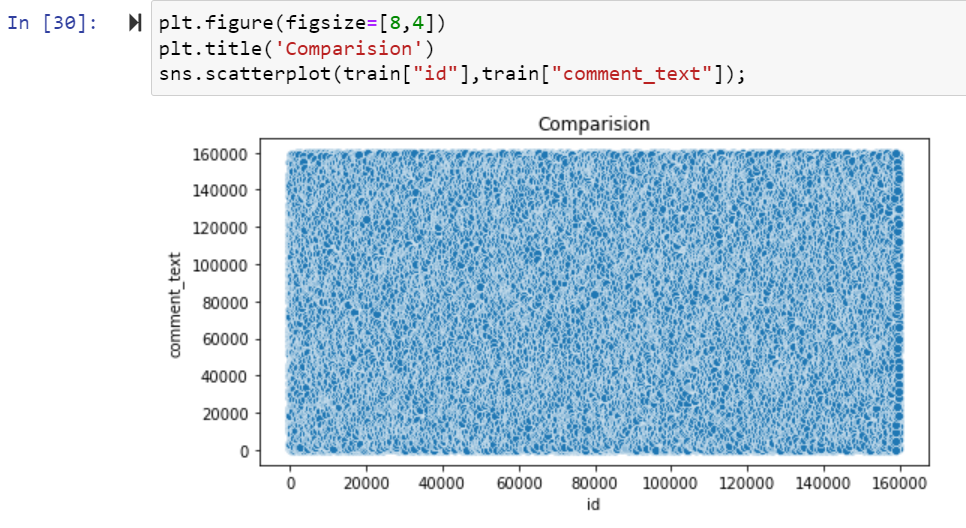


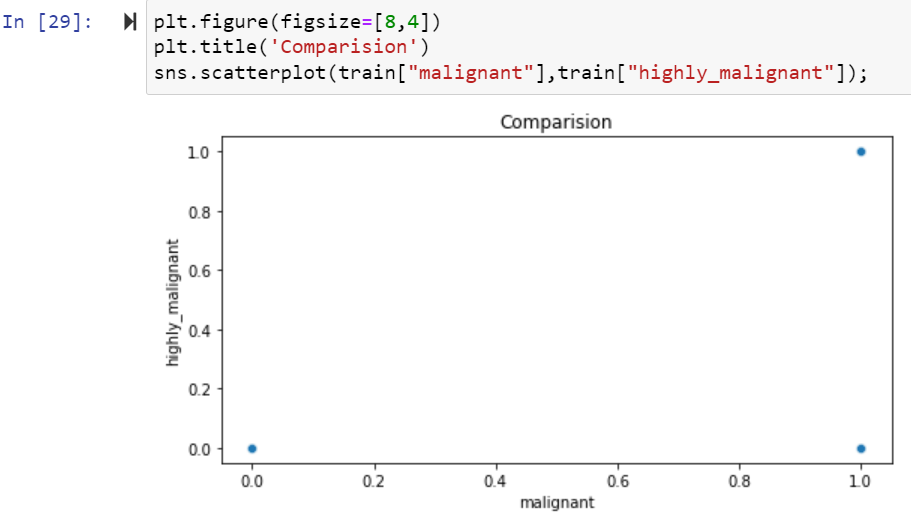


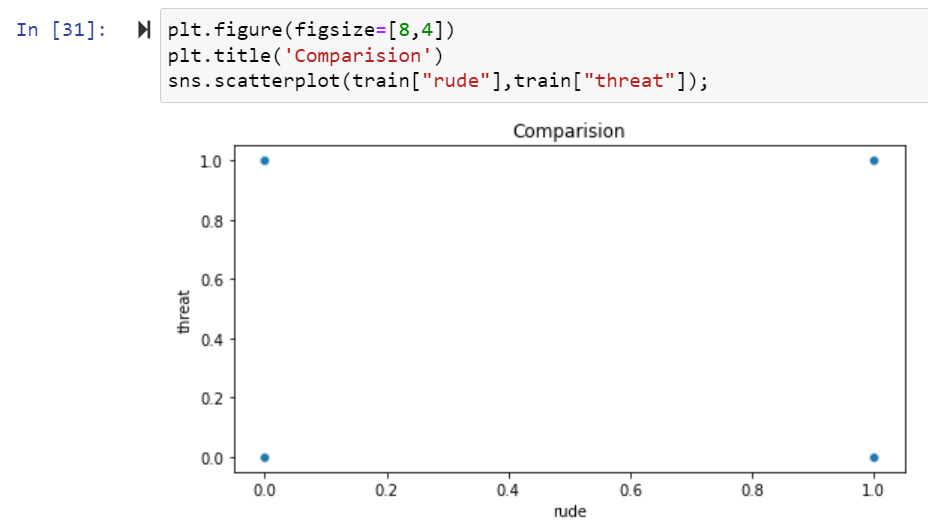
Here many Column outliers are present in Train dataset

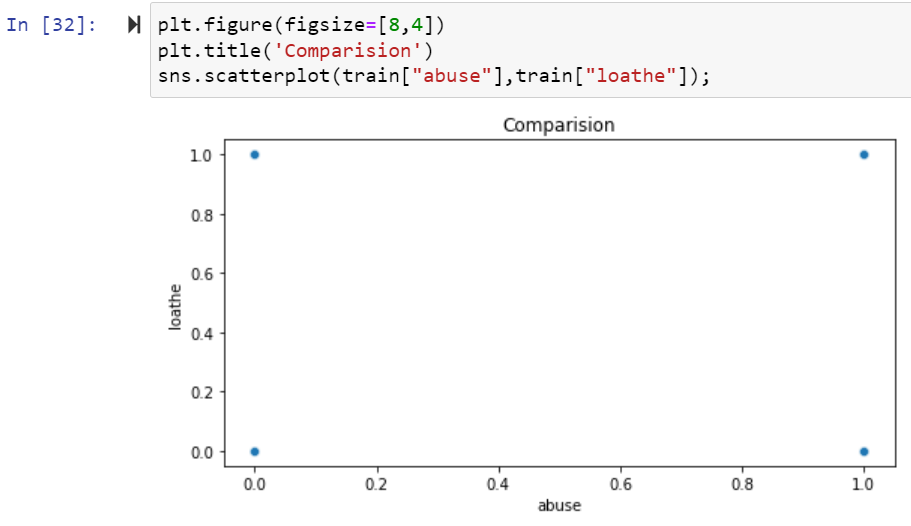
**Bi varient Analysis of Train Dataset:**

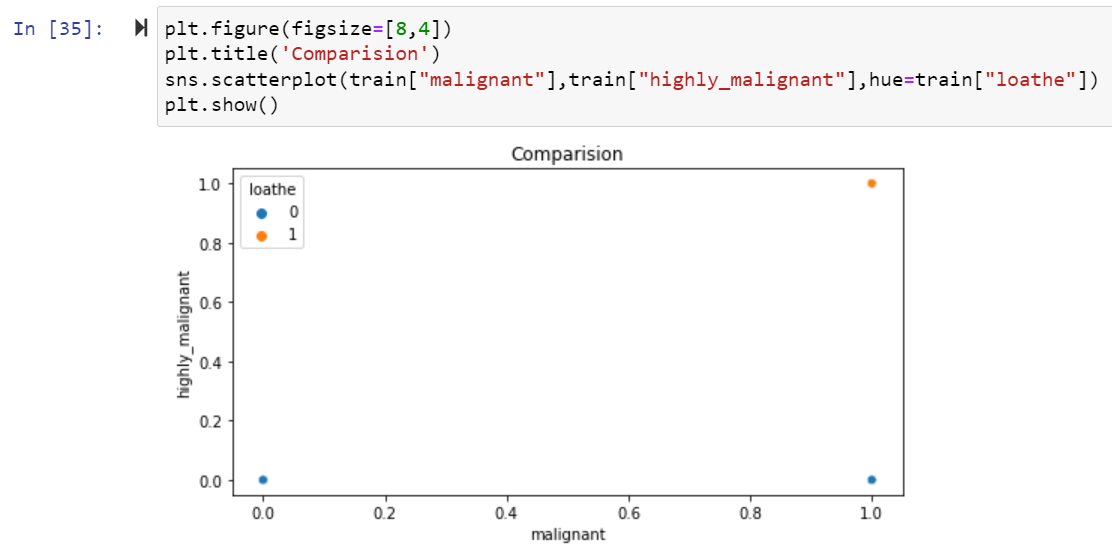
#### Scatter Plot

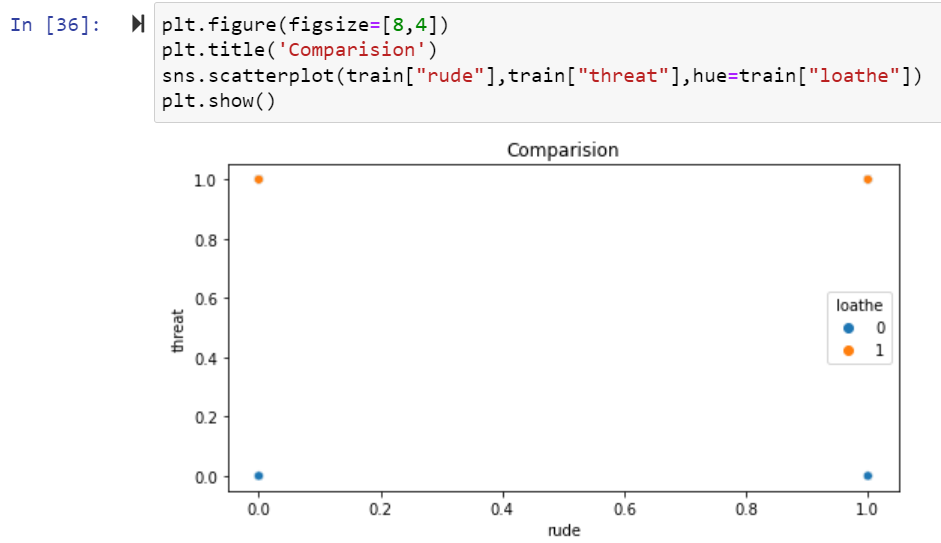


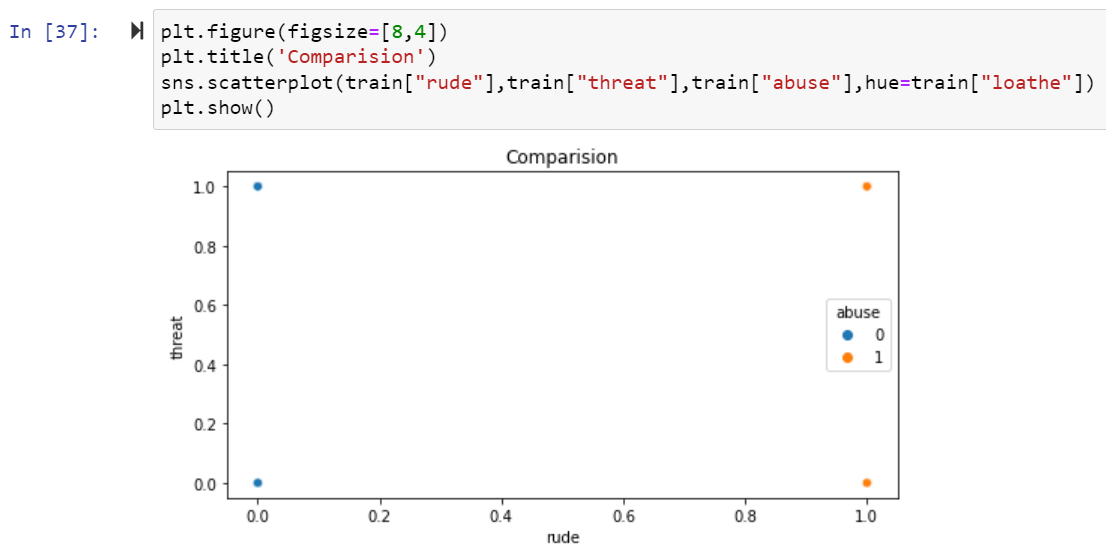












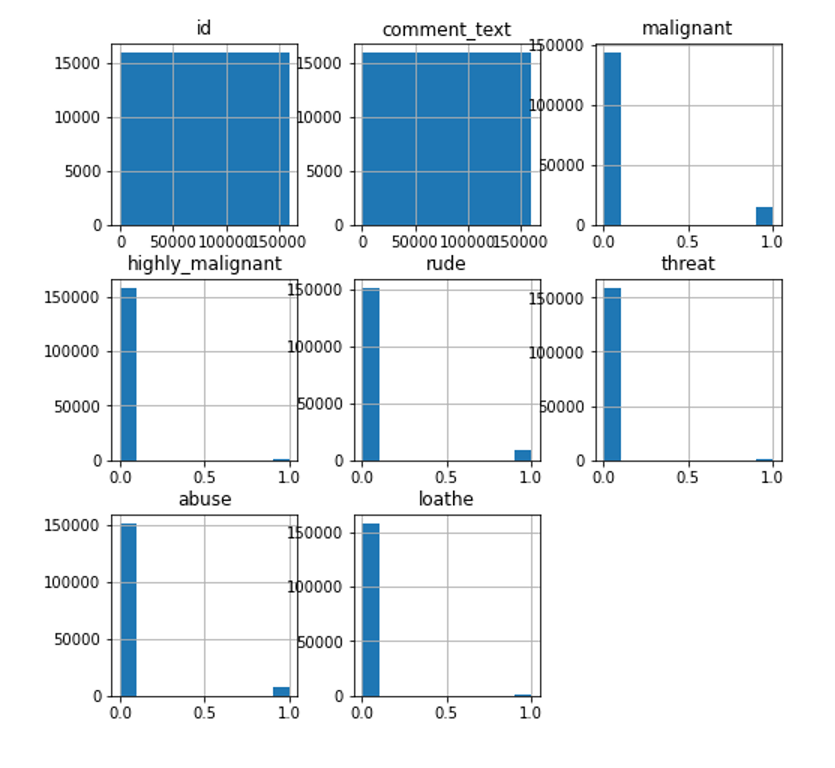
Here Above Scatterplot shows the comparison between two columns in train Dataset.

**Multi Varient Analysis of Train Dataset:**

Multivariate analysis is used **to study more complex sets of data than what univariate analysis methods can handle**. ... Multivariate analysis can reduce the likelihood of Type errors. Sometimes, univariate analysis is preferred as multivariate techniques can result in difficulty interpreting the results of the test.

**HistPlot:**

Plot univariate or bivariate **histograms to show distributions of datasets**. ... Either a long-form collection of vectors that can be assigned to named variables or a wide-form dataset that will be internally reshaped. x, y vectors or keys in data. Variables that specify positions on the x and y axes.

****

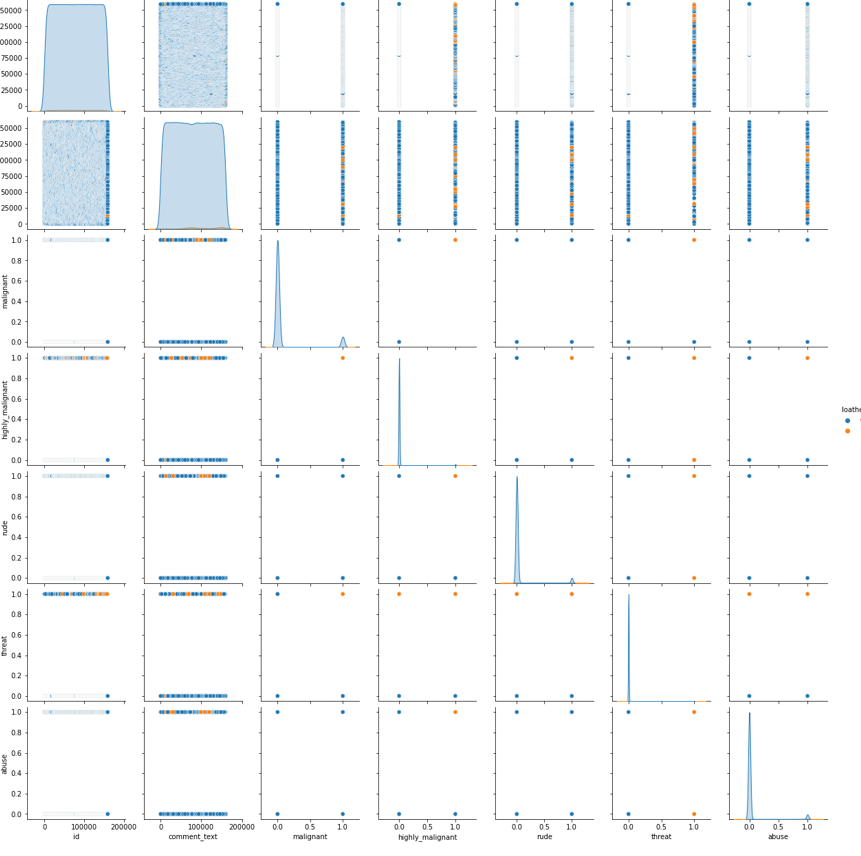
Above are the histplot of train dataset.

**Pair Plot of Train Dataset**:

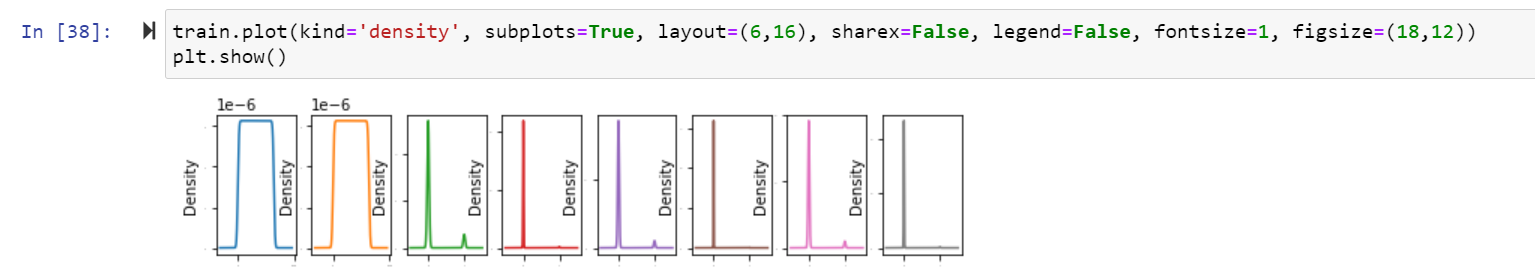
Pair plot **visualizes given data to find the relationship between them where the variables can be continuous or categorical**. Plot pairwise relationships in a data-set. Pair plot is a module of sea born library which provides a high-level interface for drawing attractive and informative statistical graphics.

**sns.pairplot(train, hue='loathe')**

A pairs plot allows us to see both distribution of single variables and relationships between two variables . Pair plots are a great **method to identify trends for follow-up analysis** and, fortunately, are easily implemented in Python!



Above are the pair plot of train dataset.



Here distribution of all columns in Train Dataset.

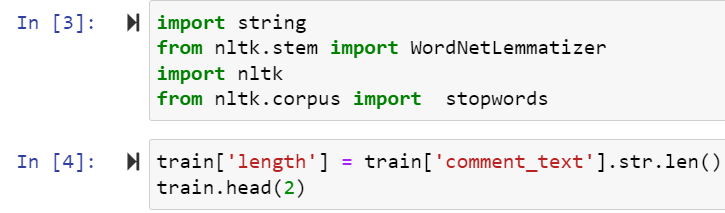
**Data Inputs- Output Logic- Relationships**

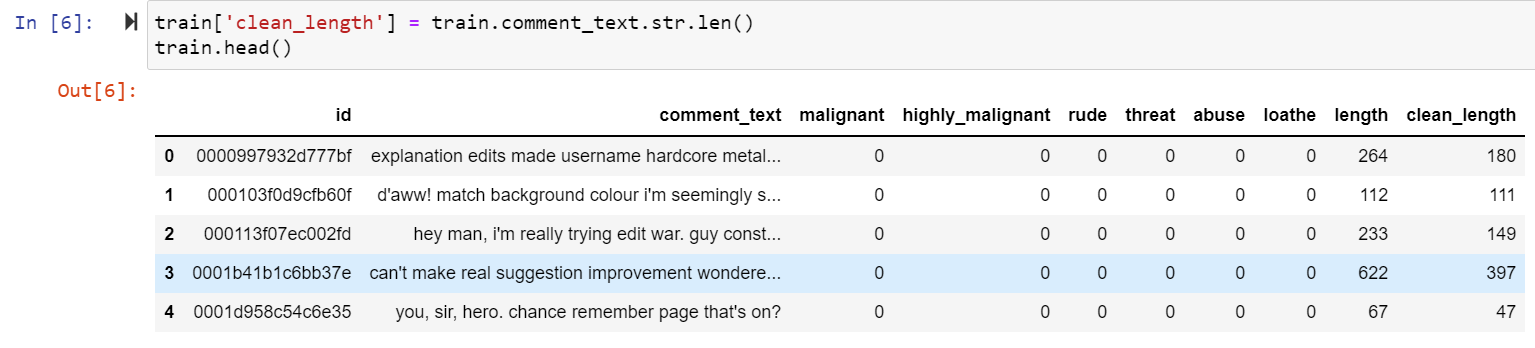
**1.Preparation for removal of punctuation marks:**I imported the string library comprising all punctuation characters and appended the numeric digits to it, as those were required to be removed too.

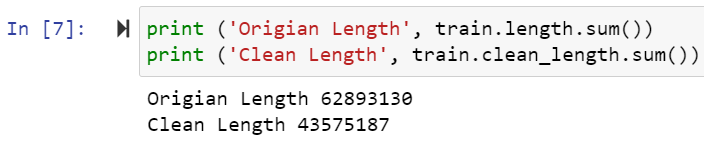
**2.Updating the list of stop words :**Stop words are those words that are frequently used in both written and verbal communication and thereby do not have either a positive/negative impact on our statement. E.g. is, this, us, etc.Python has a built-in dictionary of stop words. I used the same and also appended the single letters like ‘b’, ‘c’ …. to it, which might be pre-existing or have generated during data preprocessing.

**2.Stemming and Lemmatising :**Stemming is the process of converting inflected/derived words to their word stem or the root form. Basically, a large number of similar origin words are converted to the same word. E.g. words like “stems”, “stemmer”, “stemming”, “stemmed” are based on “stem”. This helps in achieving the training process with a better accuracy. Lemmatising is the process of grouping together the inflected forms of a word so they can be analyzed as a single item. This is quite similar to stemming in its working but not exactly same. Lemmatising depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. I used the **word-net library** in nltk for this purpose. Stemmer and Lemmatizer were imported from nltk.

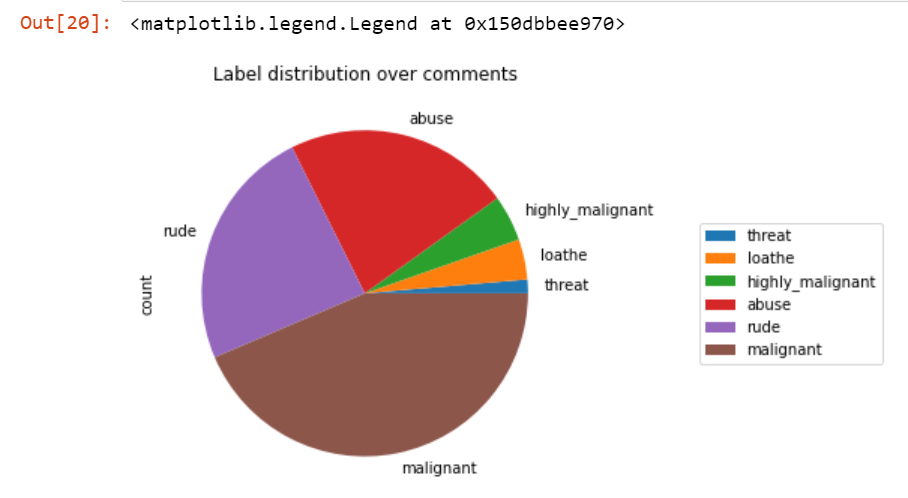
**4. Applying Count Vectorizer :**Count Vectorizer is used for converting a string of words into a matrix of words. Column headers have the words themselves and the cell values signify the frequency of occurrence of the word. I passed the custom list of stop words created earlier as the parameter with default values for ‘lowercase’ and ‘regular expression’ .

****

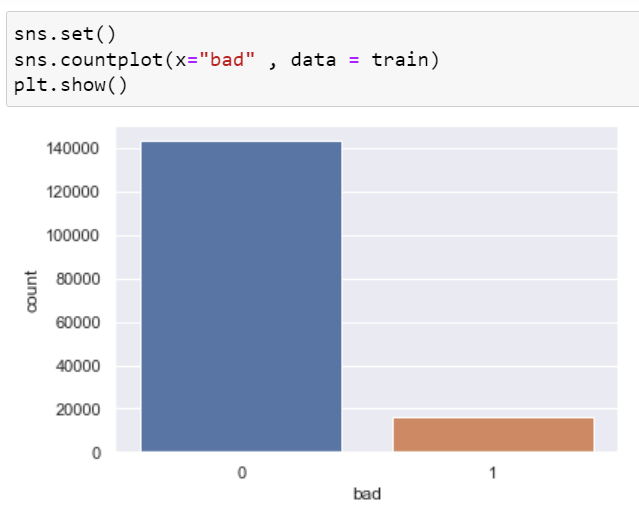
****

****

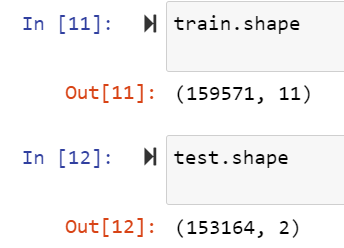
Above using Lemmatising clean the data. After clean data check original length and clean length.

****

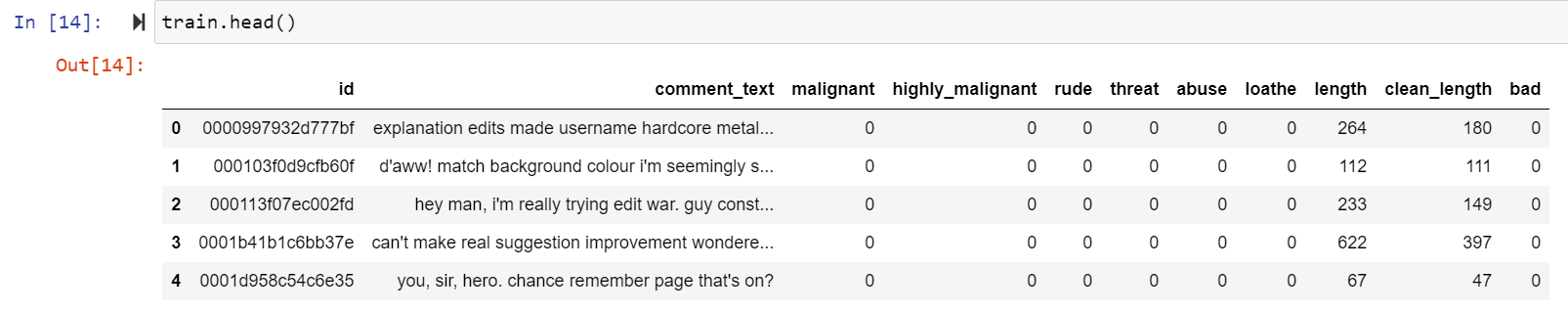
Above are the pie chart distribution of columns in train dataset.

****

Above are the count plot of bad column.



After the checking shape of train and test dataset.



After that checking train dataset.

**State the set of assumptions (if any) related to the problem under consideration**

Several Machine Learning models have been developed and deployed to filter out the unruly language and protect internet users from becoming victims of online harassment and cyberbullying.

**Hardware and Software Requirements and Tools Used**

1. Software Requirements:

* 1. Coding Language: Python3, Python
  2. Coding software : Anaconda, Jupyter Notebook

1. Microsoft Office Word.
2. Snipping Tools (For Screenshots).
3. Microsoft Excel

**Non Functional Requirements:**

1: Platform Independent: The application would be platform independent if all the requirements are installed in the device.

2: Performance: The application should have better accuracy and should provide the information in less time.

3: Capacity: The capacity of the storage should be high so that large amount of data can be stored in order to train the model.

**Hardware Requirements:**

1 GB RAM.

200 GB HDD.

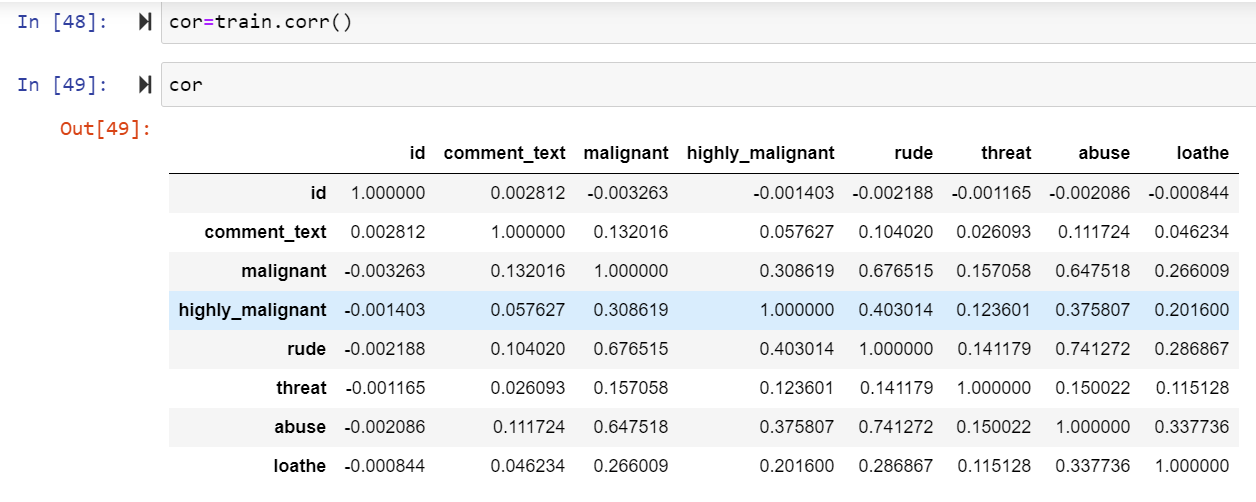
Intel 1.66 GHz Processor Pentium 4

**Visualizations:**

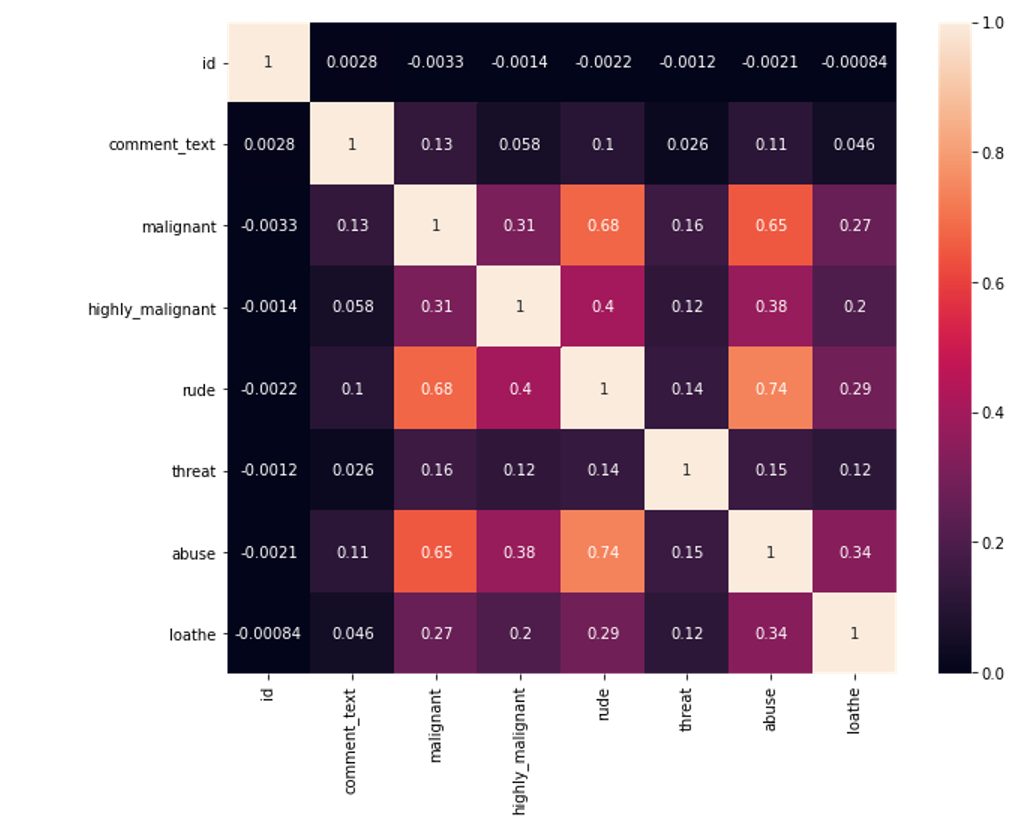
**Heatmap of Train Dataset:**

cor=df\_tr.corr()

cor

****

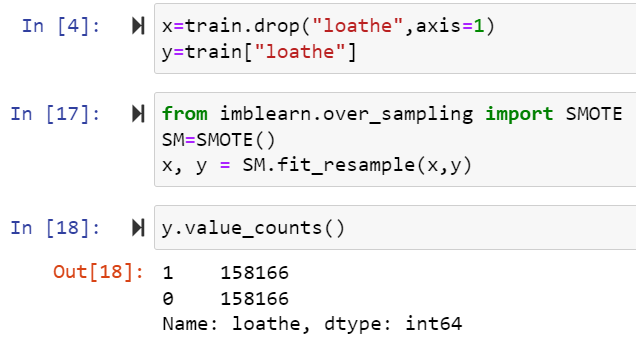
After that Checking Correlation of all independent columns with Target column in train dataset.



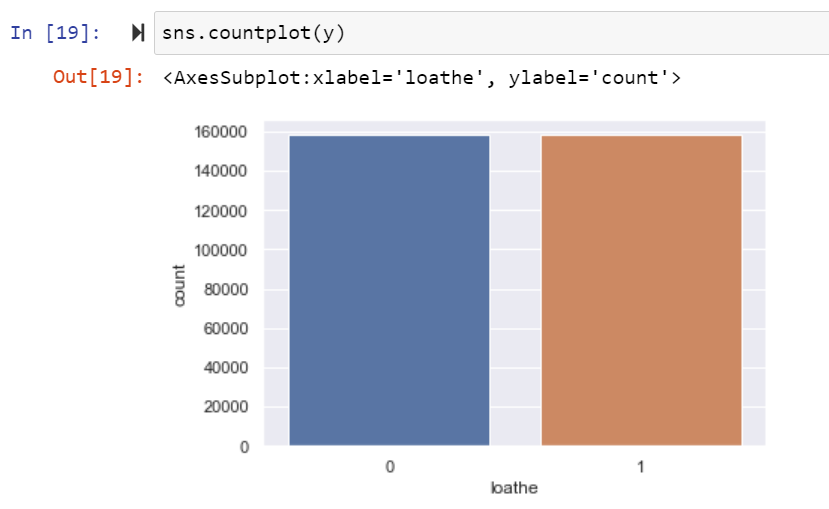
Heat map shows the correlation of every independent variable in dataset with target variable. Here above heatmap the every independent variable check correlation with tar

**Oversampling:**

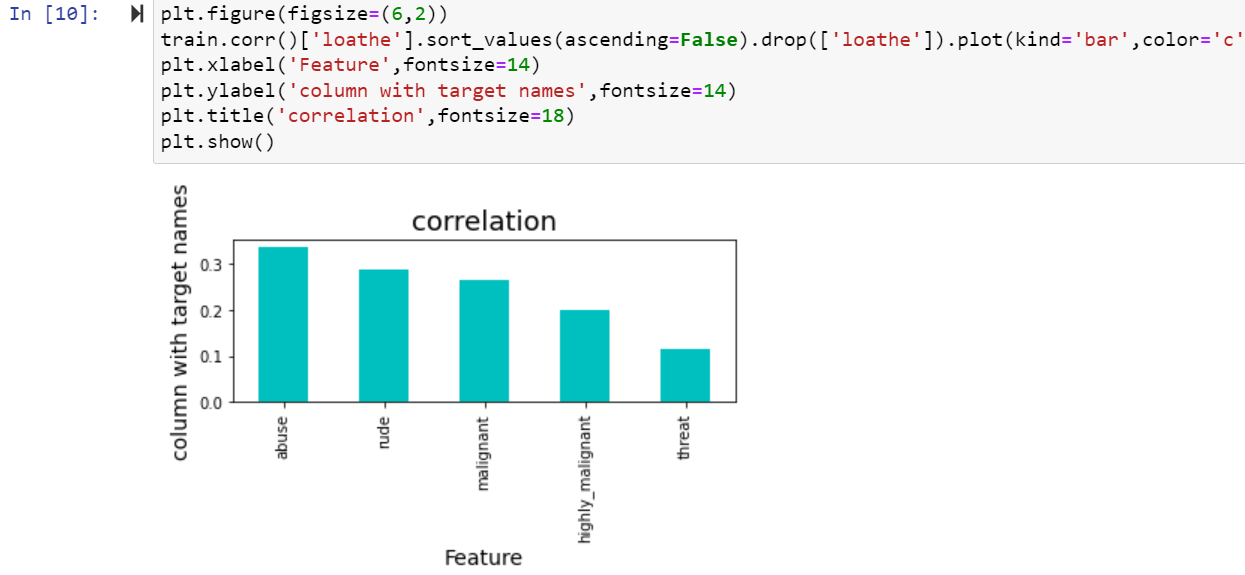
**SMOTE:**

****

After using oversampling SMOTE method now imbalance class are balance.



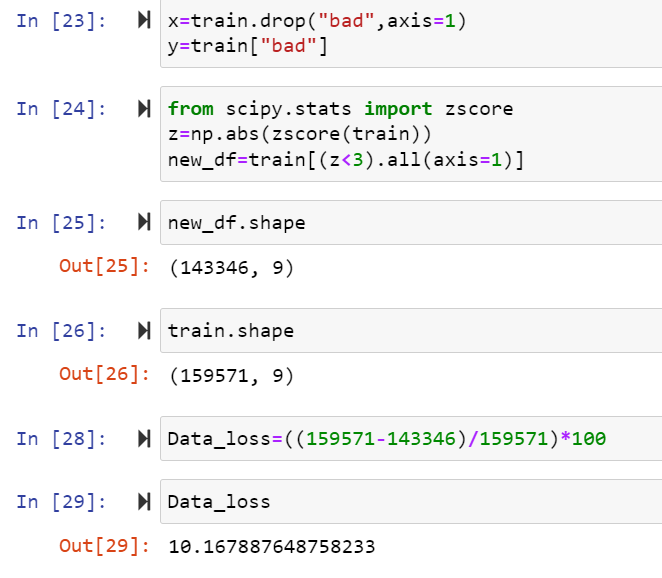
Now class are balance.



Here we shows the correlation in another way in this above we show all columns are positively correlated with target variable.

**Z-score:**

Take your data point, subtract the mean from the data point, and then divide by your standard deviation. That gives you your Z-score. You can use Z-Score to determine outliers.One of the most commonly used tools in determining outliers is the Z-score. Z-score is just the number of standard deviations away from the mean that a certain data point is.In your future data science life, Z-scores are gonna be a really useful way to think about how usual or how unusual a certain data point is. And that’s going to be really valuable once we start making inferences based on our data. In this story, we will take a deep dive into our notebooks and learn how to detect outliers using Z-Score.



After that Cheking Data loss of the dataset after using preprocessing steps here using zscore method checking dataloss here after zscore there 10.17 is dataloss.

**IQR:**

The interquartile range rule is useful in detecting the presence of outliers. [Outliers](https://www.thoughtco.com/what-is-an-outlier-3126227) are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

## What Is the Interquartile Range?

Any set of data can be described by its [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237). These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile Q1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile Q3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

These five numbers tell a person more about their data than looking at the numbers all at once could, or at least make this much easier. For example, the [range](https://www.thoughtco.com/what-is-the-range-in-statistics-3126248), which is the minimum subtracted from the maximum, is one indicator of how spread out the data is in a set (note: the range is highly sensitive to outliers—if an outlier is also a minimum or maximum, the range will not be an accurate representation of the breadth of a data set).

Range would be difficult to extrapolate otherwise. Similar to the range but less sensitive to outliers is the interquartile range. The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-3126245) is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile:

IQR = Q3 – Q1.

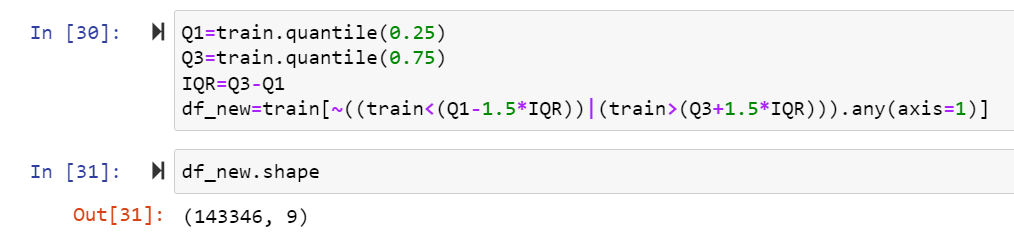
The interquartile range shows how the data is spread about the median. It is less susceptible than the range to outliers and can, therefore, be more helpful.

## Using the Interquartile Rule to Find Outliers

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Remember that the interquartile rule is only a rule of thumb that generally holds but does not apply to every case. In general, you should always follow up your outlier analysis by studying the resulting outliers to see if they make sense. Any potential outlier obtained by the interquartile.

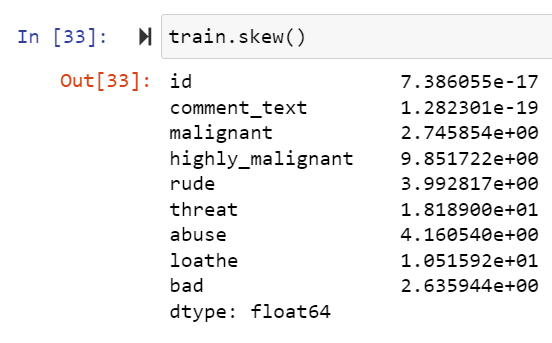


Here After Applying IQR method checking dataloss here using IQR method same dataloss.

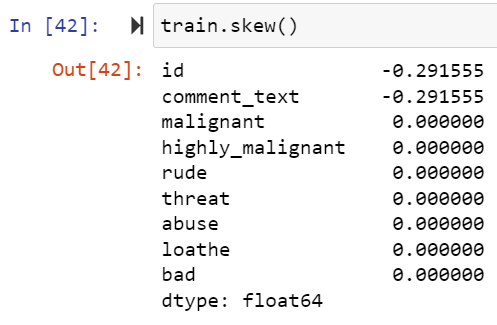
**Skewness:**

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

**Checking skewness in dataset**







To remove skewness use Power Transformer and log transform technique in dataset.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods)**

**Approach**

Importing the required libraries and reading the dataset.

Merging of the two datasets

* Understanding the dataset

1. Exploratory Data Analysis (EDA) –

* Data Visualization

1. Feature Engineering

* Duplicate value removal
* Missing value imputation
* Encoding of categorical variables
* Dropping of redundant feature columns
* Check for the outliners and removal of outliers.

1. Model Building

* Performing train test split
* Feature Scaling
* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* AdaBoost Classifier
* KNeighborsClassifier

1. Model Validation

* R2 square error

1. Hypermeter Tuning (GridSearchCV)

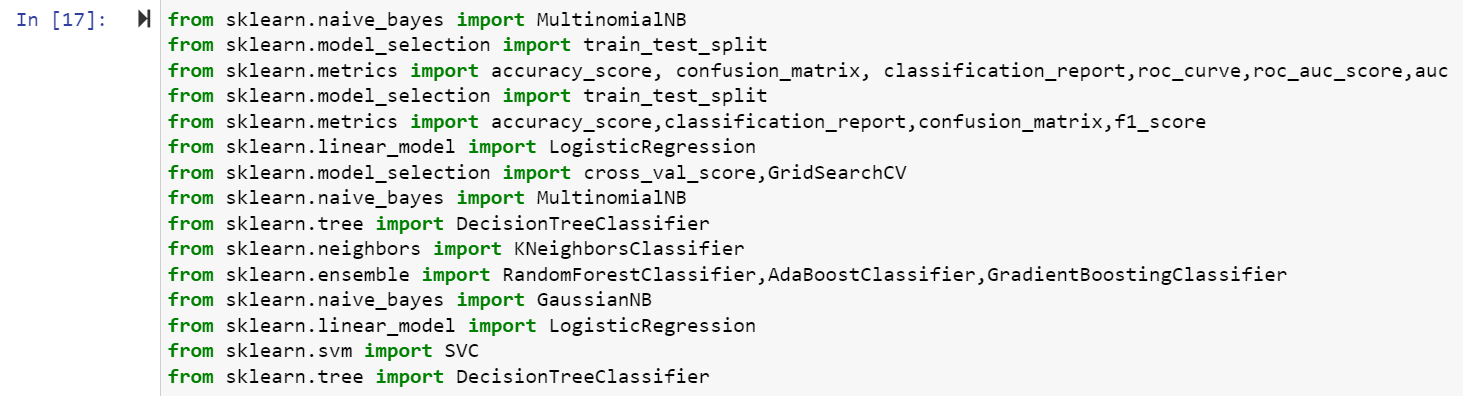
For Random Forest Regressor

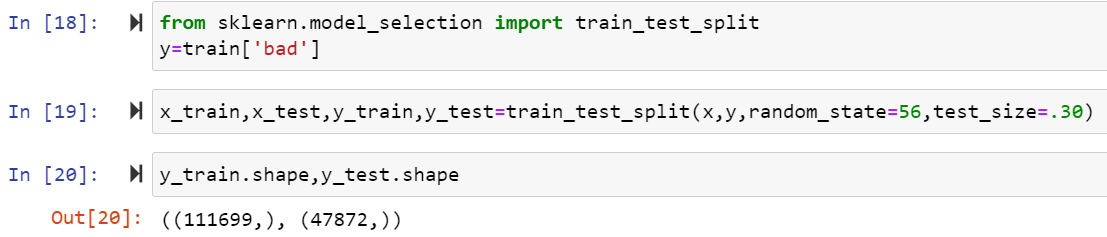
1. Checking for Feature Importance
2. Creating the final model and making predictions

**Testing of Identified Approaches (Algorithms)**

* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* AdaBoost Classifier
* KNeighborsClassifier

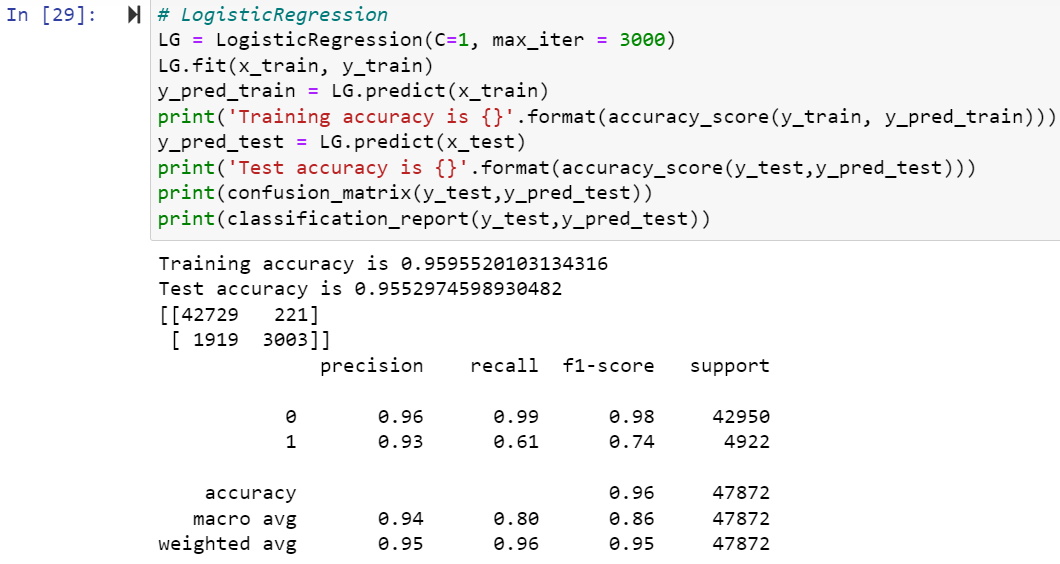
**Building Machine Learning Models:**

****

****

**LogisticRegression:**

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification. Scikit-learn has a highly optimized version of logistic regression implementation, which supports multiclass classification task (Raschka, 2015).

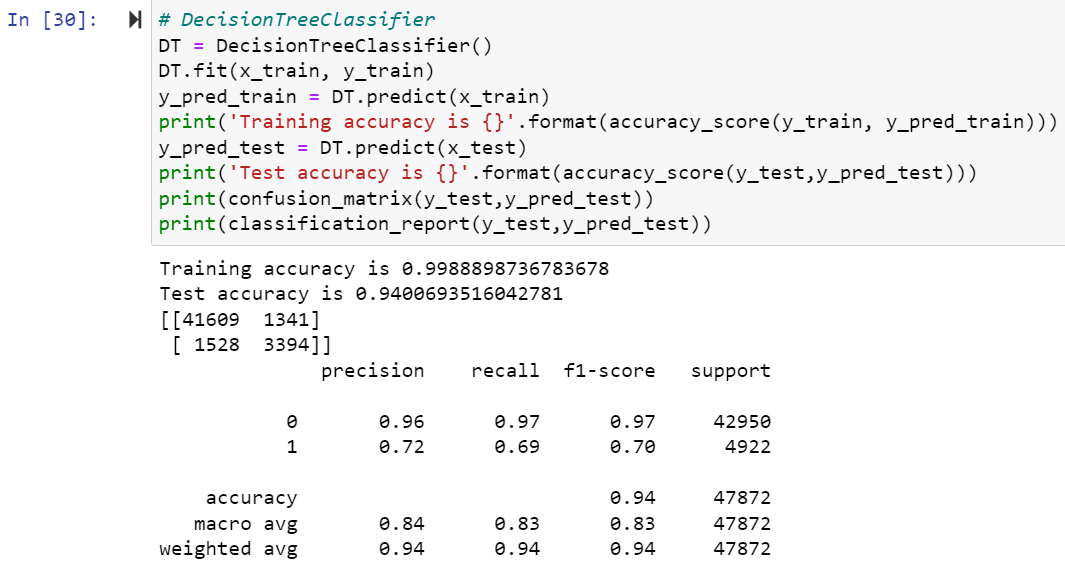
****

Here Accuracy score is 0.95 for train dataset and 0.95 for test dataset and cross validation score is 0.44 in LogisticRegression algorithm.

**Decision tree classifier**

The decision tree classifier (Pang-Ning et al., 2006) **creates the classification model by building a decision tree**. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

Tree-based classification approaches are nonlinear models that work by partitioning the input space into cuboid regions. The prediction of the model is based on the most dominant class represented by training examples in the cuboid region that matches the unlabeled example. Owing to this, trained tree-based models are easily understood with no machine learning background leading to their wide deployment in industrial applications.

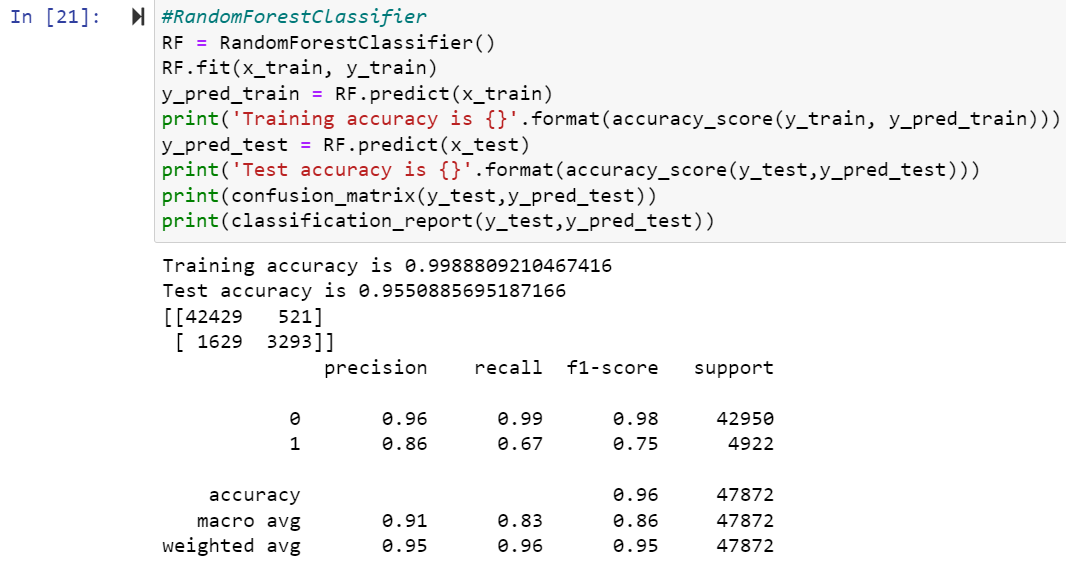
****

Here Accuracy score is 0.99 for train dataset and 0.94 for test dataset in DecisionTree Classification algorithm.

**Random Forest Classifier:**

What is a Random Forest Classifier? The term “Random Forest Classifier” refers to **the classification algorithm made up of several decision trees**. The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest's predictive powers to make accurate decisions.

The random forest is a **classification algorithm consisting of many decisions trees**. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

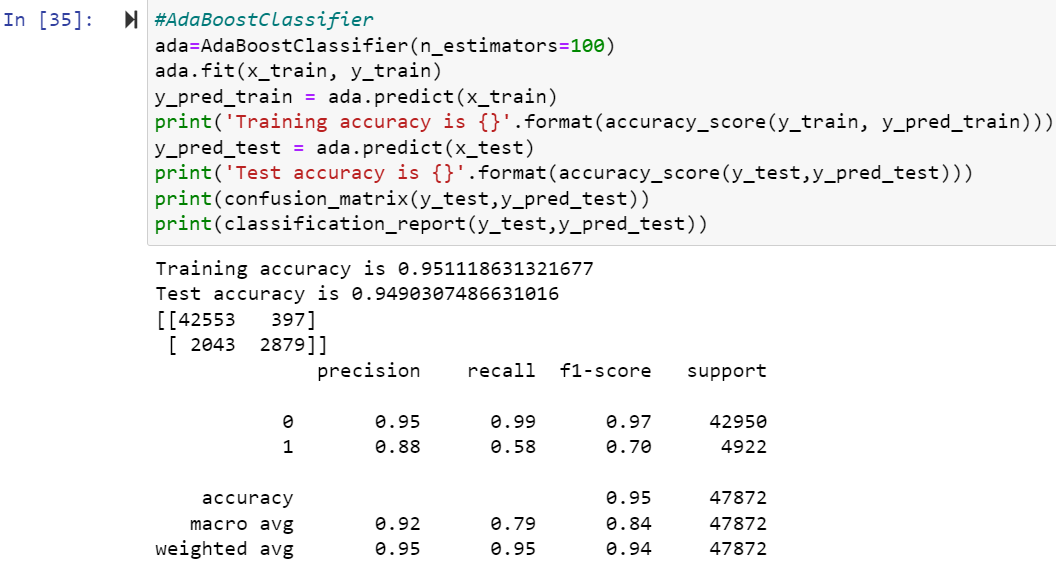


Here Accuracy score is 0.99 for train dataset and 0.95 for test dataset in RandomForest Classification algorithm.

**AdaBoost Classifier:**

AdaBoost algorithm, short for Adaptive Boosting, is a **Boosting technique used as an Ensemble Method in Machine Learning**. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference.

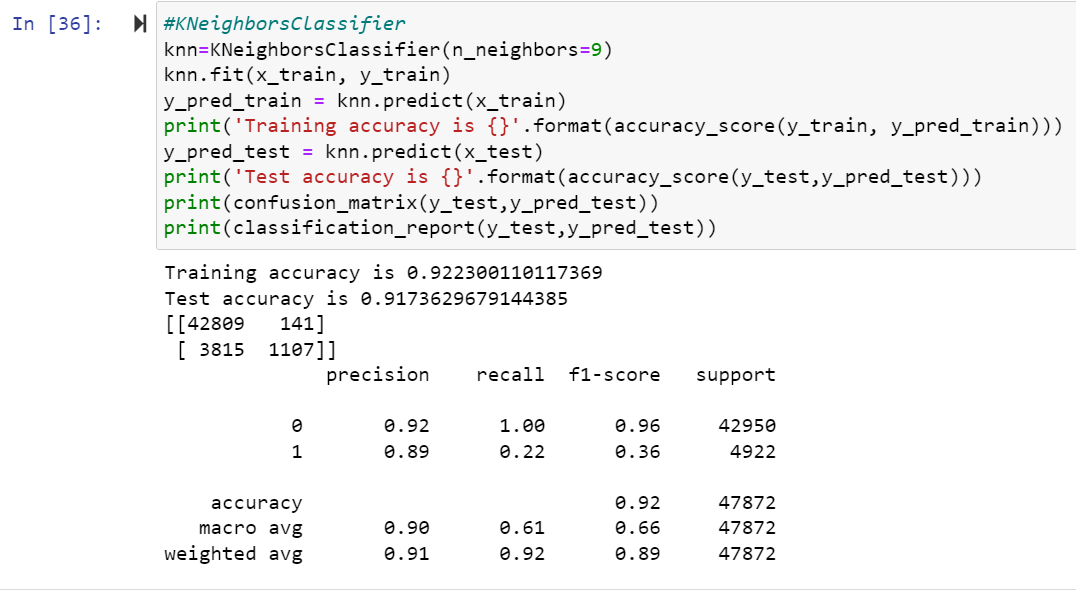
In recent years, boosting algorithms gained massive popularity in data science or machine learning competitions. Most of the winners of these competitions use boosting algorithms to achieve high accuracy. These Data science competitions provide the global platform for learning, exploring and providing solutions for various business and government problems. Boosting algorithms combine multiple low accuracy(or weak) models to create a high accuracy(or strong) models. It can be utilized in various domains such as credit, insurance, marketing, and sales. Boosting algorithms such as AdaBoost, Gradient Boosting, and XGBoost are widely used machine learning algorithm to win the data science competitions.



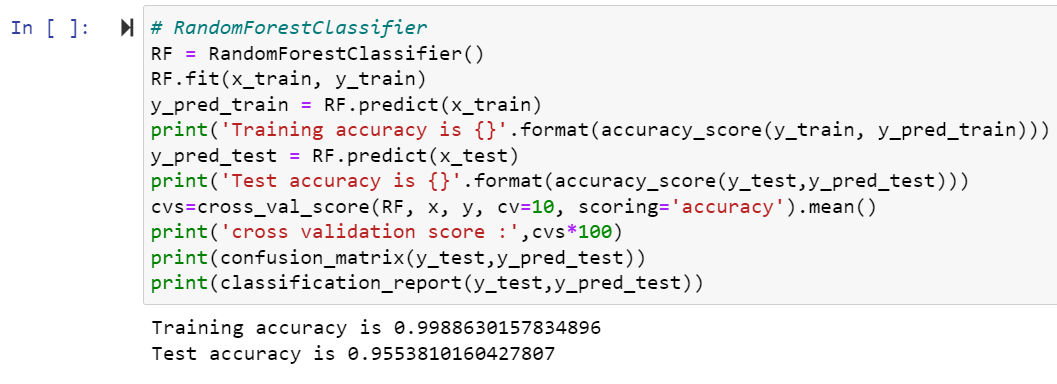
Here Accuracy score is 0.95 for train dataset and 0.94 for test dataset in AdaBoost Classifier algorithm.

**KNeighborsClassifier:**

K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In Credit ratings, financial institutes will predict the credit rating of customers. In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

****

Here Accuracy score is 0.92 for train dataset and 0.91 for test dataset in KNeighbors Classification algorithm.



Here cross validation score is 0.99 for train dataset and 0.95 for test dataset so high accuracy score and cross validation score so Random Forest Classifier is the best model so apply hyperparameter tunning on it.

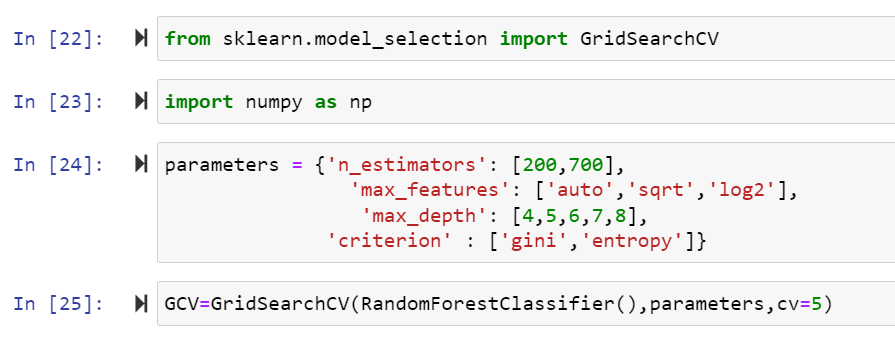
**GridSearchCV :**

GridSearchCV is **a library function that** is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

**Parameterlist:**

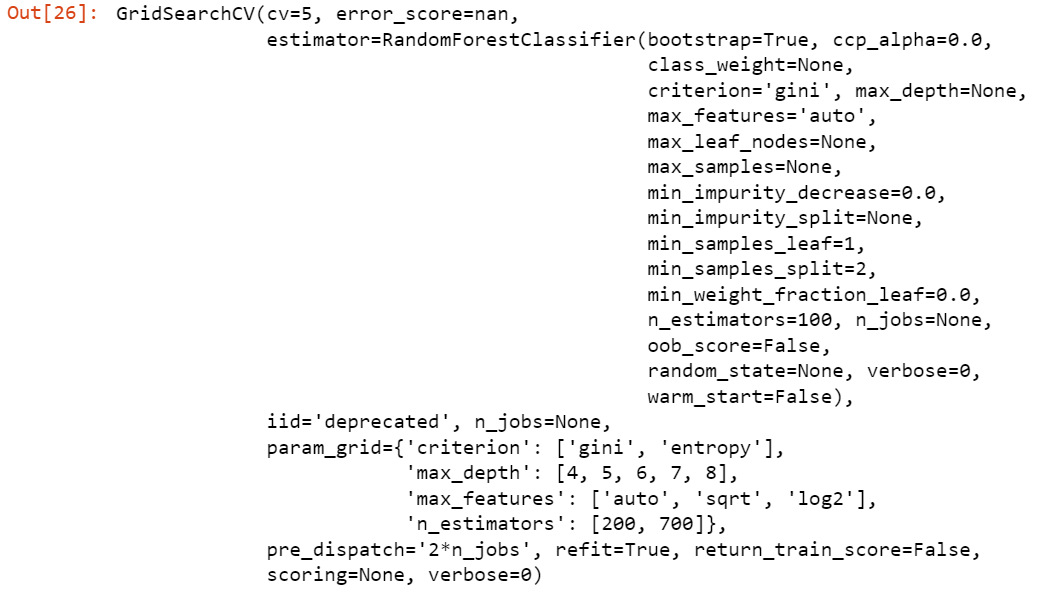
There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model.

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model.



Here above are the parameter list of RandomForestClassifier model.

The above code block we have the following parameters  
max\_features: In this maximum features: auto,sqrt,log2 criterion : gini , entropy n\_estimator:200,700 , max\_depth: 4,5,6,7,8.

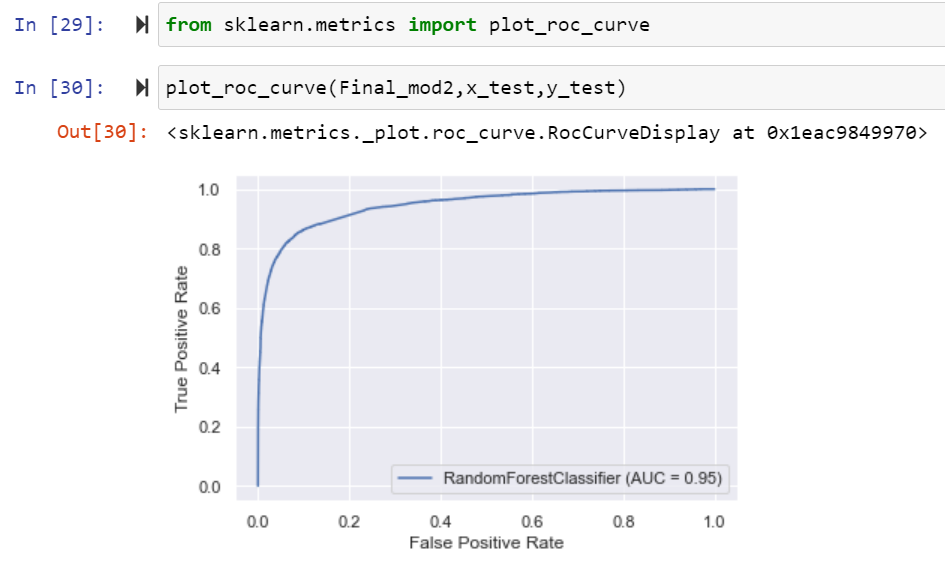




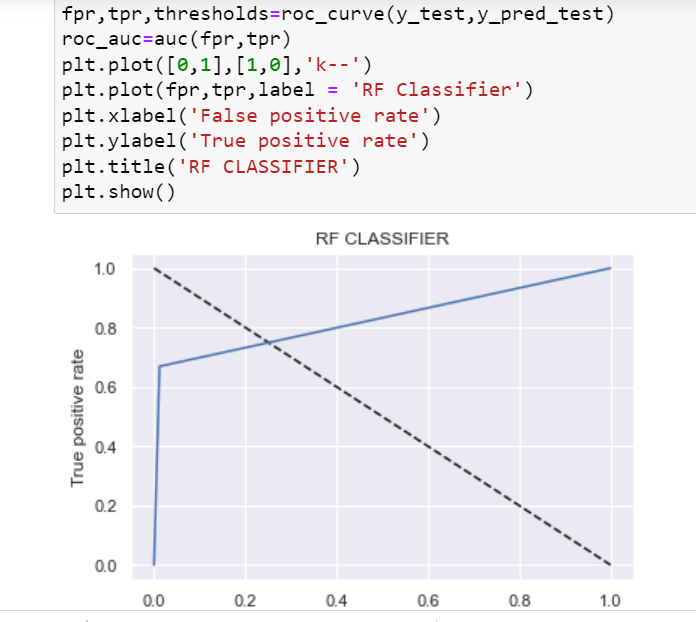
Here after use different parameter list the best parameter list is select. Above are the best parameter list RandomForestClassifier. Put this parameters into the model so output is finally best score os RandomForestClassifier is 89 so it is the best score.

**AUC and ROC Curve:**

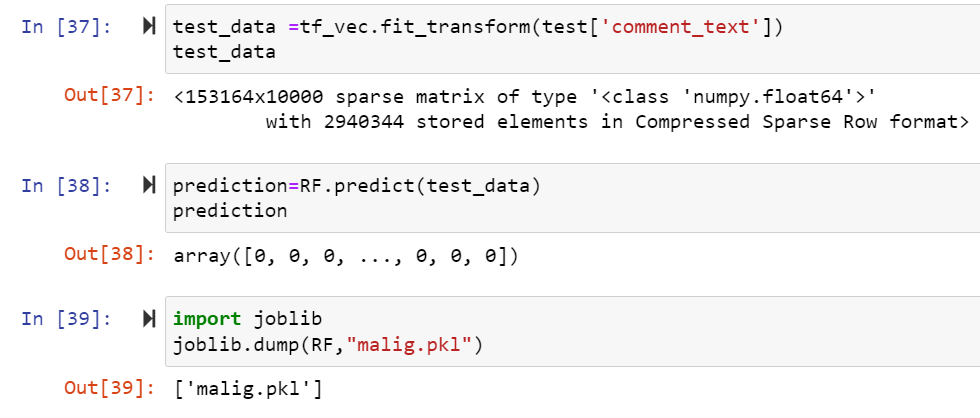
AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

****

A

****

nsxsn



Finally Load the model and predict the values.

**Visualizations**

**Interpretation of the Results**

The process for this project was as follows:

1. Analyze the problem and propose a useful solution.
2. Explore the dataset to get a better picture of how the labels are distributed, how they correlate with each other, and what defines toxic or clean comments.
3. Develop an objective that fits a practical use case and addresses the major class imbalance.
4. Create a baseline score with a simple logistic regression classifier.
5. Explore the effectiveness of multiple machine learning algorithms.
6. Select the best model based on a balance of performance and efficiency.
7. Refine the preprocessing strategies to optimize model performance.
8. Tune model parameters to maximize performance.
9. Build a the final model with the best performing algorithm and parameters and test it on a holdout subset of the data.

‘

**Conclusion**

Harmful or malignant comments in the social media space have many negative impacts to society. The ability to readily and accurately identify comments as toxic could provide many benefits while mitigating the harm. Also, our research has shown the capability of readily available algorithms to be employed in such a way to address this challenge.

**Future Work**

We also suggest using SVM for text processing and text classification. It requires a grid search for hyper-parameter tuning to get the best results.