

**Malignant Comments Classifier Project**

Submitted by:

Vinita Shinkar

**ACKNOWLEDGMENT**

The goal is to create a classifier model that can predict if input text is Malignant. Explore the dataset to get a better picture of how the labels are distributed, how they correlate with each other, and which comments are having inappropriate content. Explore the effectiveness of multiple machine learning approaches and select the best for this problem. Select the best model and tune the parameters to maximize performance. Build the final model with the best performing algorithm and parameters and test it on a holdout subset of the data.

**INTRODUCTION:**

The advances in IT technologies and generalizing virtualization all over the world has led to an unprecedented participation in social media; and there is no doubt that social media is one of the biggest hallmarks of the 21st century. Meanwhile, social media is a place to express individual opinions and share thoughts in line with a constructive contribution to develop a safe place for everybody practicing their rights accordingly. However, behind the shield of computers as virtual walls, some individuals also think they can abuse and harass other people’s opinions and characters. Accordingly, a jargon word has been coined recently to address such behaviors as “cyberbullying”. Cyberbullying could be introduced as mistreatments that occurs all over digital instruments and devices such as computers, tablets, and mobile phones. Cyberbullying may take place via short massage system (SMS), general apps with the possibility of communication between users, online social media forums, or even online gaming where individuals can virtually participate in. Cyberbullying includes posting, sending, or sharing harmful, negative, mean, or false content about someone else either directly sent to the person or post as a general comment where other could observe it. It also includes sharing private or personal info about someone else in order to humiliation or embarrassment. Such online harassment suppresses so many of our fellow citizens from expressing their opinions. These studies and results related to this kind of online harassment leads to an important area of data science in order to be able to separate and distinguish harassment comments and cyberbullying, we call them malignant comments, from normal comments. A research initiative founded by Jigsaw, and Google are currently working on tools to help improve online conversations. One significant aspect of their efforts is aiming to identify the toxic comments and lunch online toxicity monitoring system on various online social platforms. The need of advanced methods and techniques to improve identification of different types of comment posted online motivated the current study. As mentioned above, distinguishing improper comment that brings harassment to privacy and morality is one of the most crucial subjects related to the current extension of social media. With all the progress and improvement in IT and data science, the world is in requirement of a properly designed technique to find and isolate these kinds of comments that we call it toxic.

**Source of Data:**

The study is based on secondary data collected through various internet web sites.

**Data Analysis:**

Analysis of data and the information collected from the secondary sources were made keeping the objectives of the study in mind.

**PROJECT DESCRIPTION:**

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.

**Idea**

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 Analysis**:

Firstly, we do some EDAs to gain a general understanding of our data, and detecting some important metrics and trends that may come helpful for our further analysis and model building.

This dataset contains 159,571 comments. The data consists of one input feature, the string data for the comments, and six labels for different categories of malignant comments: Malignant’, ‘Highly malignant’, ‘Rude’, ‘Threat’, ‘Abuse’ and ‘Loathe’.

We will now understand about the features in our dataset.

**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.

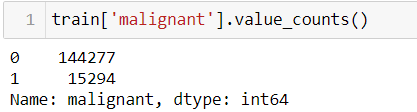
**Abuse:** It is for comments that are abusive in nature.

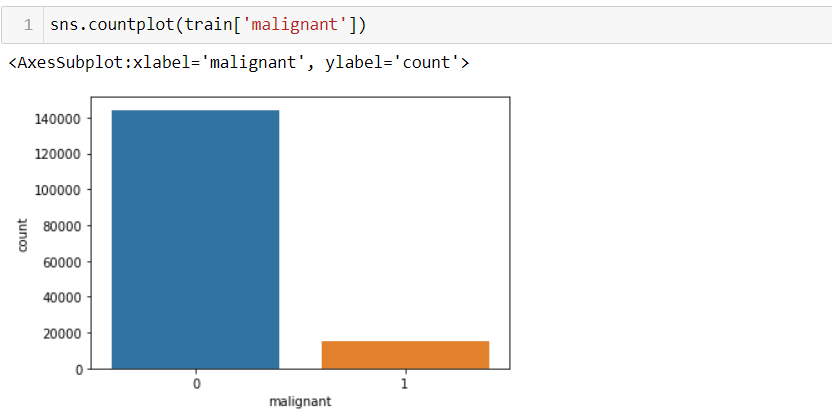
**Loathe:** It describes the comments which are hateful and loathing in nature.

**ID:** It includes unique Ids associated with each comment text given.

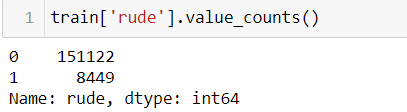
**Comment text:** This column contains the comments extracted from various social media platforms.

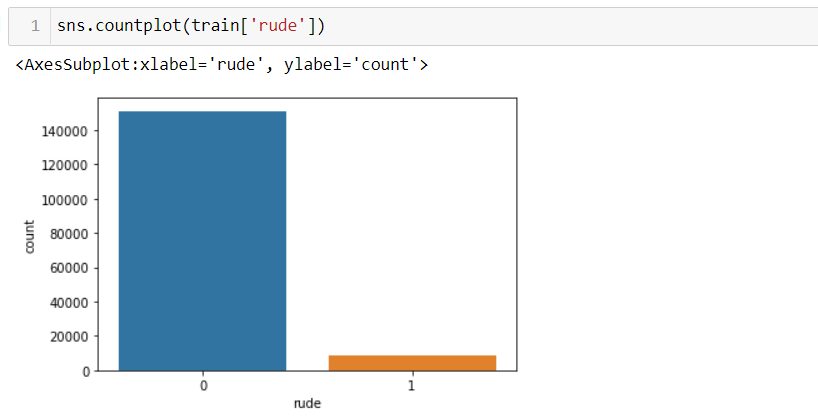
Now, we will go to graphical visualisation.



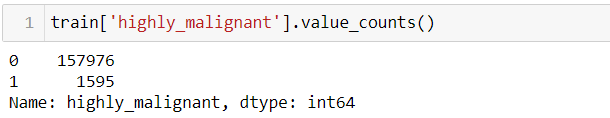
****

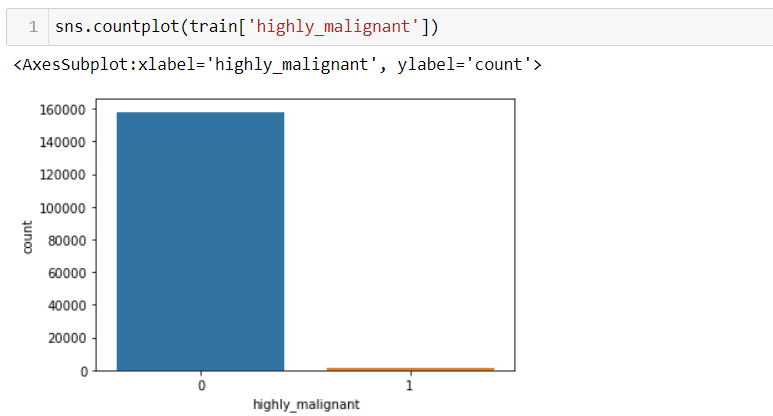
From above images we can observe that there are around 15000 comments that contains inappropriate data.

****

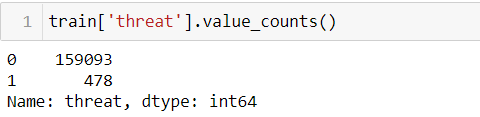
****

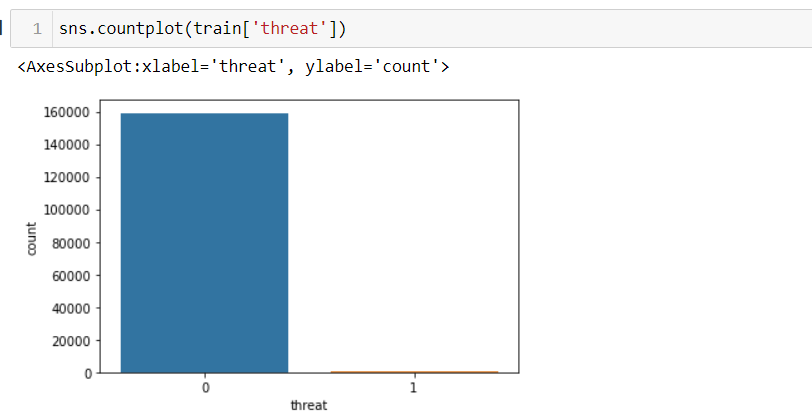
From the above image, we can observe that there are 8449 comments that contains inappropriate data and which fall under rude comments.

****

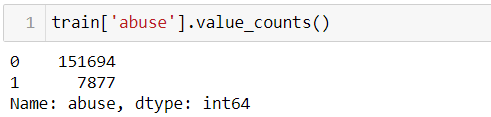
****

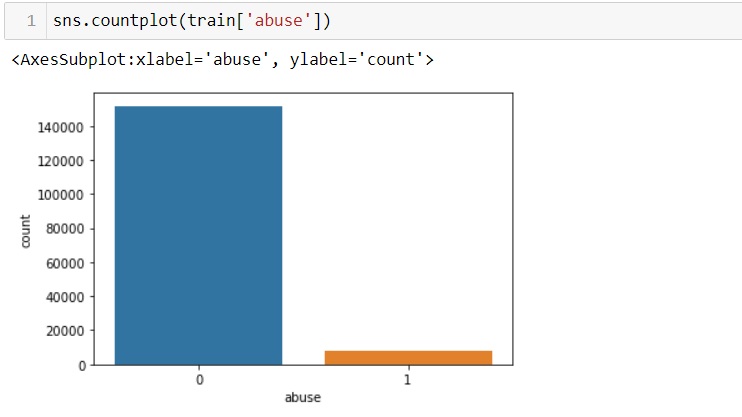
From the above image, we can observe that there are 1595 comments that contains inappropriate data and which fall under highly\_malignant comments.

****

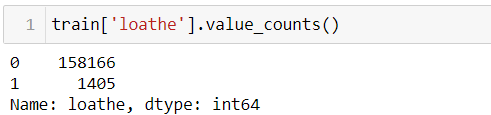
****

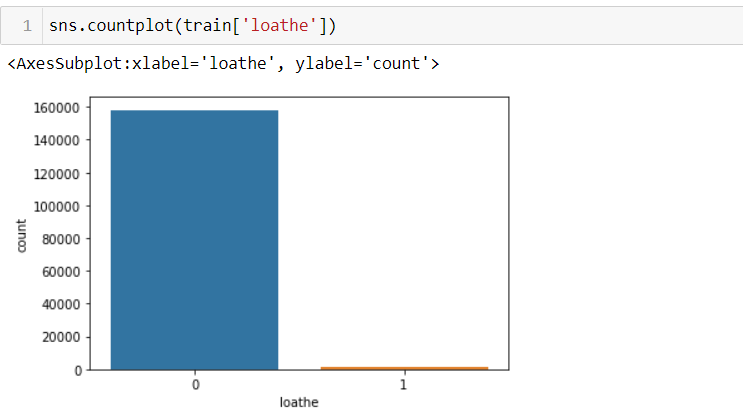
From the above image, we can observe that there are 478 comments that contains inappropriate data and which fall under threat comments.

****

****

From the above image, we can observe that there are 7877 comments that contains inappropriate data and which fall under abuse comments.

****

****

From the above image, we can observe that there are 1405 comments that contains inappropriate data and which fall under loathe comments.

From all the features we can observe that we have few comments that are malignant which cover approximately 10% of the total comments in the dataset.



The correlation matrix below provides more insight into these overlapping categories.

## **Natural Language Processing (NLP)**

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks. Examples include machine translation, summarization, ticket classification, and spell check.

One of the main reasons natural language processing is so critical to businesses is that it can be used to analyze large volumes of text data, like social media comments, customer support tickets, online reviews, news reports, and more.

All this business data contains a wealth of valuable insights, and NLP can quickly help businesses discover what those insights are.

It does this by helping machines make sense of human language in a faster, more accurate, and more consistent way than human agents.

NLP tools process data in real time, 24/7, and apply the same criteria to all your data, so you can ensure the results you receive are accurate – and not riddled with inconsistencies.

Once NLP tools can understand what a piece of text is about, and even measure things like sentiment, businesses can start to prioritize and organize their data in a way that suits their needs.

### **Challenges of NLP**

While there are many [challenges in natural language processing](https://monkeylearn.com/blog/natural-language-processing-challenges/), the [benefits of NLP](https://monkeylearn.com/blog/nlp-benefits/) for businesses are huge making NLP a worthwhile investment.

However, it’s important to know what those challenges are before getting started with NLP.

Human language is complex, ambiguous, disorganized, and diverse. There are more than 6,500 languages in the world, all of them with their own syntactic and semantic rules.

Even humans struggle to make sense of language.

So for machines to [understand natural language](https://monkeylearn.com/blog/natural-language-understanding/), it first needs to be transformed into something that they can interpret.

In NLP, syntax and [semantic analysis](https://monkeylearn.com/blog/semantic-analysis/) are key to understanding the grammatical structure of a text and identifying how words relate to each other in a given context. But, transforming text into something machines can process is complicated.

Data scientists need to teach NLP tools to look beyond definitions and word order, to understand context, word ambiguities, and other complex concepts connected to human language.

## **How Does Natural Language Processing Work?**

In natural language processing, human language is separated into fragments so that the grammatical structure of sentences and the meaning of words can be analyzed and understood in context. This helps computers read and understand spoken or written text in the same way as humans.

Here are a few fundamental NLP [pre-processing tasks](https://monkeylearn.com/blog/data-preprocessing/) data scientists need to perform before NLP tools can make sense of human language:

* **Tokenization:** breaks down text into smaller semantic units or single clauses
* **Part-of-speech-tagging**: marking up words as nouns, verbs, adjectives, adverbs, pronouns, etc
* **Stemming and lemmatization:** standardizing words by reducing them to their root forms
* **Stop word removal**: filtering out common words that add little or no unique information, for example, prepositions and articles (at, to, a, the).

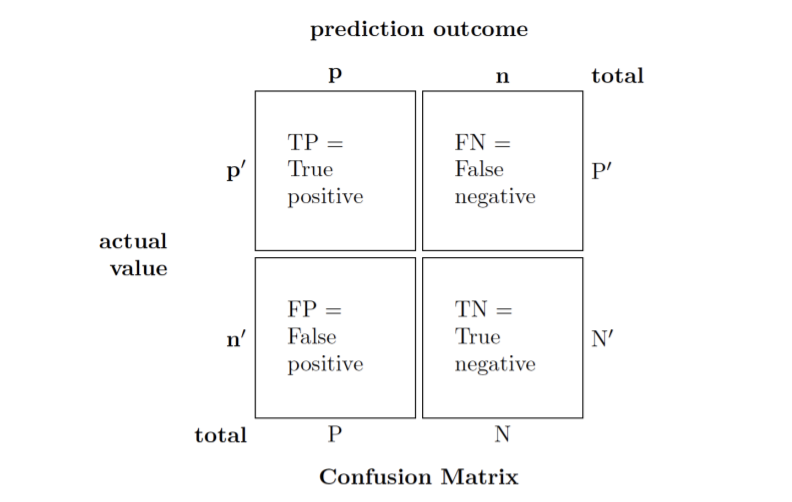
Only then can NLP tools transform text into something a machine can understand.

**Metrics:**

In order to be able to evaluate the performance of each algorithm, several metrics are defined accordingly.

**Confusion Matrix**:

It is very informative performance measures for classification tasks. Ci,j an element of matrix tells how many of items with label i are classified as label j. Ideally we are looking for diagonal Confusion matrix where no item is miss-classified. The matrix in Figure 1 is a good representation for our binary classification. Positive (P) represents toxic label and n (negative) represents non-toxic label.



Elements of confusion matrix; P (positive) represents toxic label and n (negative) represents non-toxic label.

**Accuracy**:

This metric measures how many of the comments are labeled correctly. However, in our data set, where most of comments are not toxic, regardless of performance of model, a high accuracy was achieved.



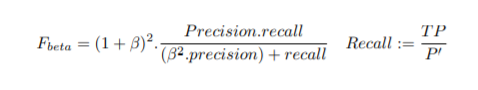
**Precision and Recall**:

Precision and recall in were designed to measure the model performance in its ability to correctly classify the toxic comments. Precision explains what fraction of toxic classified comments are truly toxic, and Recall measures what fraction of toxic comments are labeled correctly.



**F Score**:

Both Precision and Recall are important for checking the performance of the model. However, implementing a more advanced metric that combines both Precision and Recall together is quite informative and applicable (Equation 9). In this equation, setting β = 1 leads equation to return harmonic mean of Precision and Recall.



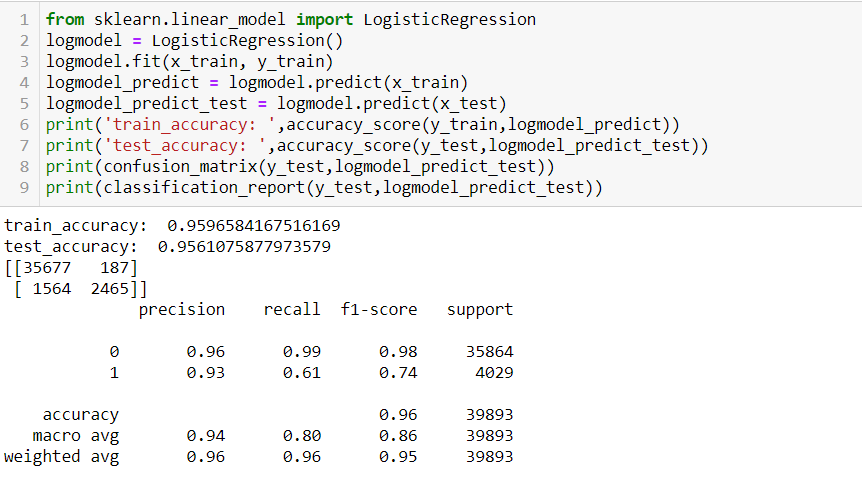
**Model Building :**

**Logistic regression**

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a [data set](https://whatis.techtarget.com/definition/data-set). Logistic regression has become an important tool in the discipline of [machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML). The approach allows an [algorithm](https://whatis.techtarget.com/definition/algorithm) being used in a machine learning application to classify incoming data based on historical data. As more relevant data comes in, the algorithm should get better at predicting classifications within data sets. Logistic regression can also play a role in [data preparation](https://searchbusinessanalytics.techtarget.com/definition/data-preparation) activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load ([ETL](https://searchdatamanagement.techtarget.com/definition/Extract-Load-Transform-ELT)) process in order to stage the information for analysis.

A logistic regression model predicts a [dependent data variable](https://whatis.techtarget.com/definition/dependent-variable) by analysing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school student will be admitted to a particular college.

The resulting analytical model can take into consideration multiple input criteria. In the case of college acceptance, the model could consider factors such as the student’s grade point average, SAT score and number of extracurricular activities. Based on [historical data](https://whatis.techtarget.com/definition/historical-data) about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

****

**Random Forest Classifier**

Random Forest is an example of ensemble learning, in which we combine multiple machine learning algorithms to obtain better predictive performance.

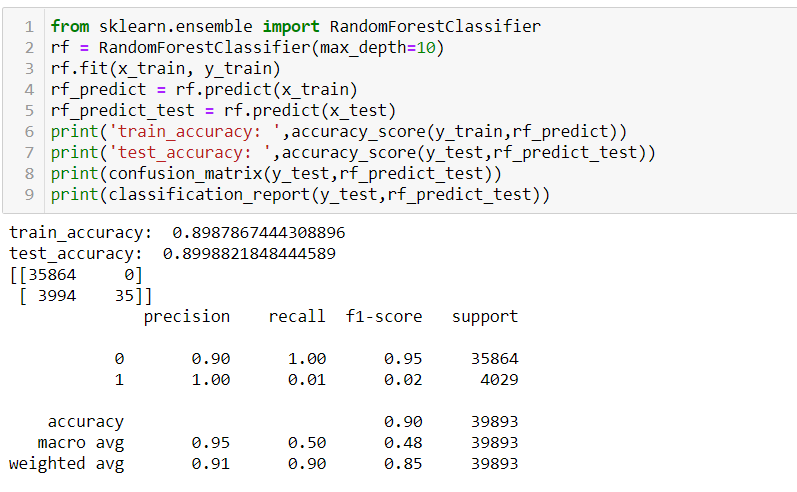
**Why the name “Random”?**

Two key concepts that give it the name random:

1. A random sampling of training data set when building trees.
2. Random subsets of features considered when splitting nodes.

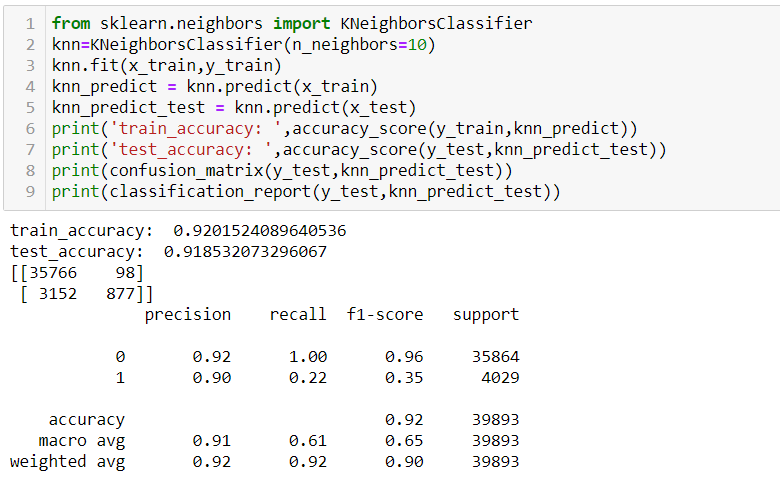
A technique known as bagging is used to create an ensemble of trees where multiple training sets are generated with replacement.

In the bagging technique, a data set is divided into **N** samples using randomized sampling. Then, using a single learning algorithm a model is built on all samples. Later, the resultant predictions are combined using voting or averaging in parallel.



**k-Nearest Neighbors (KNN)**

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

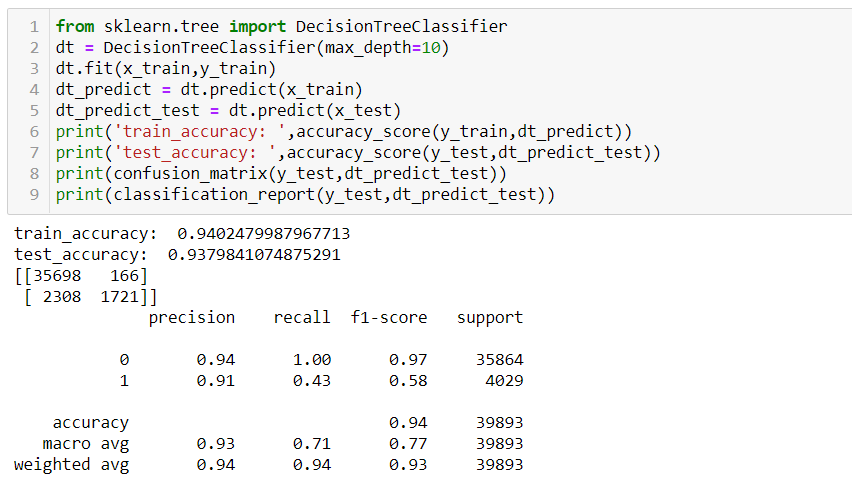


### **Decision Tree Classifier**

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

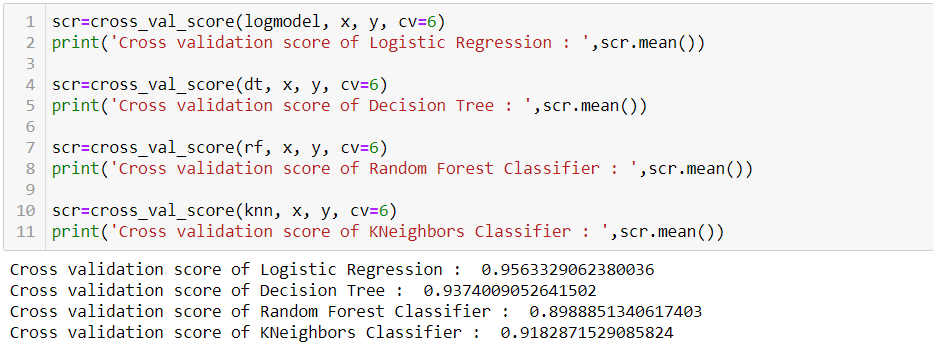
The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.



**Cross Validation**:

The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).



From above cross validation, we can observe that Logistic Regression is having least difference between accuracy score and cross validation.

So, Logistic Regression Model with accuracy score of 95.96% is the best model

**Deployment**

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data.

