**1. Introduction**

A **social media sentiment analysis** tells you how people feel about your brand online. Rather than a simple count of mentions or comments, **sentiment analysis** considers emotions and opinions. It involves collecting and analyzing information in the posts people share about your brand on **social media**.

Sentiment analysis involves natural language processing because it deals with human-written text. The ability to categorize opinions expressed in the text —and especially to determine whether the writer's attitude is positive, negative, or neutral—is highly valuable. In this project, we will use the process known as sentiment analysis to categorize the opinions of people.

**Motivation for the project:**

I had a little background in Machine Learning, Data Science and therefore I decided to polish my skills by building this project. This Project gave me the insights of the science behind and how a model actually “learns” to work.

**2. SOFTWARE REQUIREMENTS :**

Language : python3.8

Operating system : Window10

**3. HARDWARE REQUIREMENTS :**

Processor : Intel core i3 7 th Gen

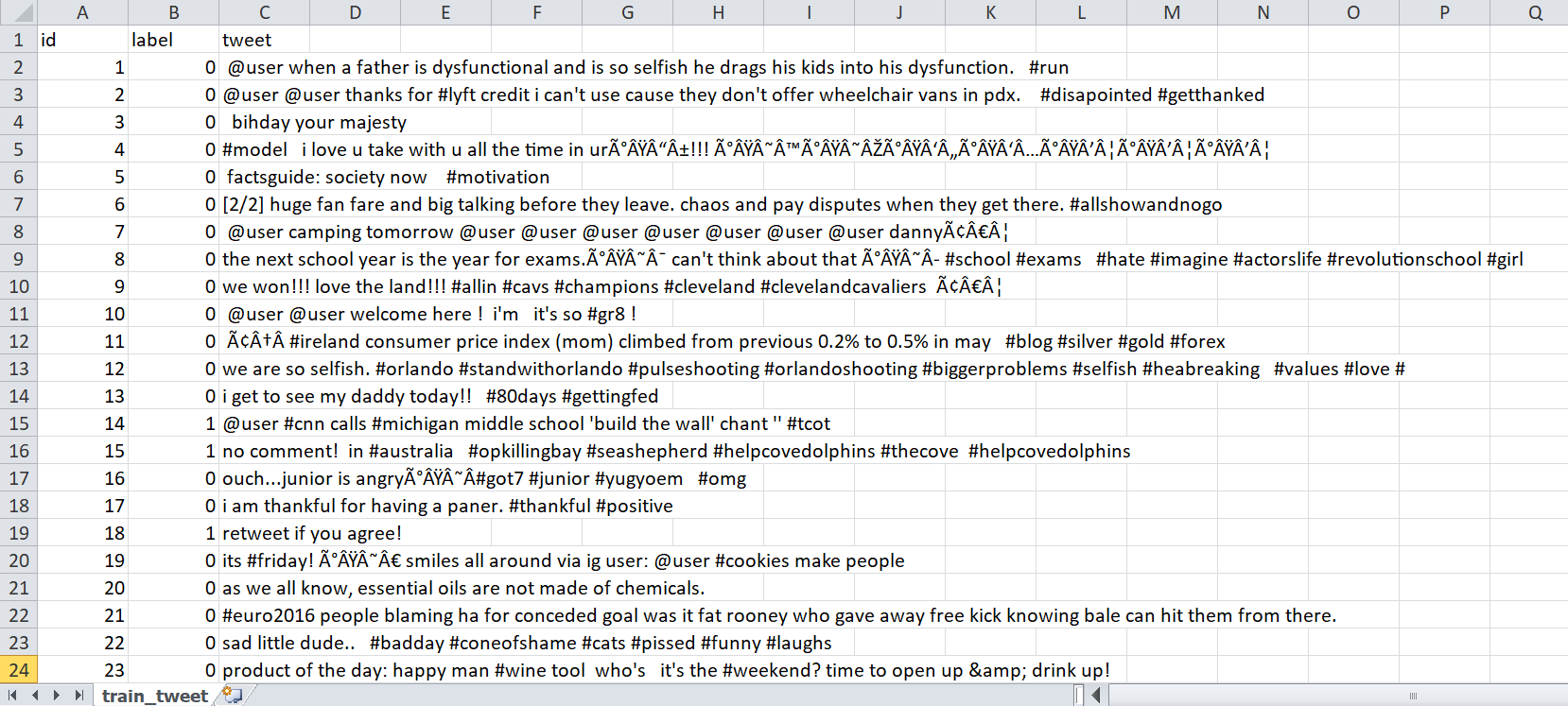
Random Memory : 128MB

Hard Disk : 20GB

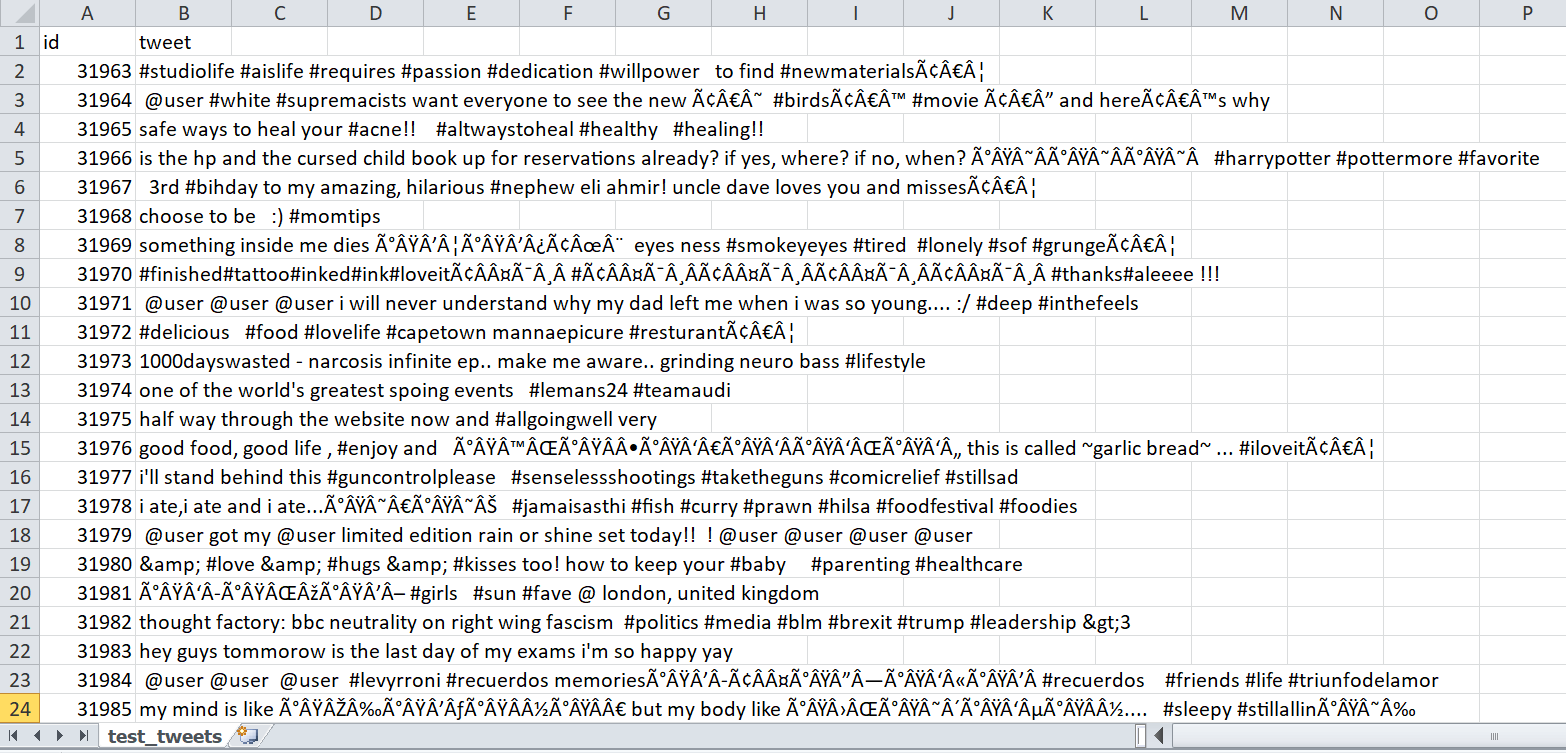
Processor Speed : 300 min

**4. DATASET**

My dataset is divided into two categories training dataset and the testing dataset.



Train dataset (csv file)[with positive label ‘1’ & negative with ‘0’]



Test dataset (csv file)[without lables]

**Creating Twitter Sentiment Analysis Python Program**

1. Importing the libraries: Seaborn, pandas, NumPy, nltk, sklearn.
2. Loading the dataset and fetching the tweets

**Tweets Preprocessing and Cleaning :**

This is one of the essential steps in any natural language processing (NLP) task. We never get filtered, ready-to-use data. To make it workable, there is a lot of processing that needs to happen. The preprocessing of the text data is an essential step as it makes the raw text ready for mining.

**5. Data Exploration**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

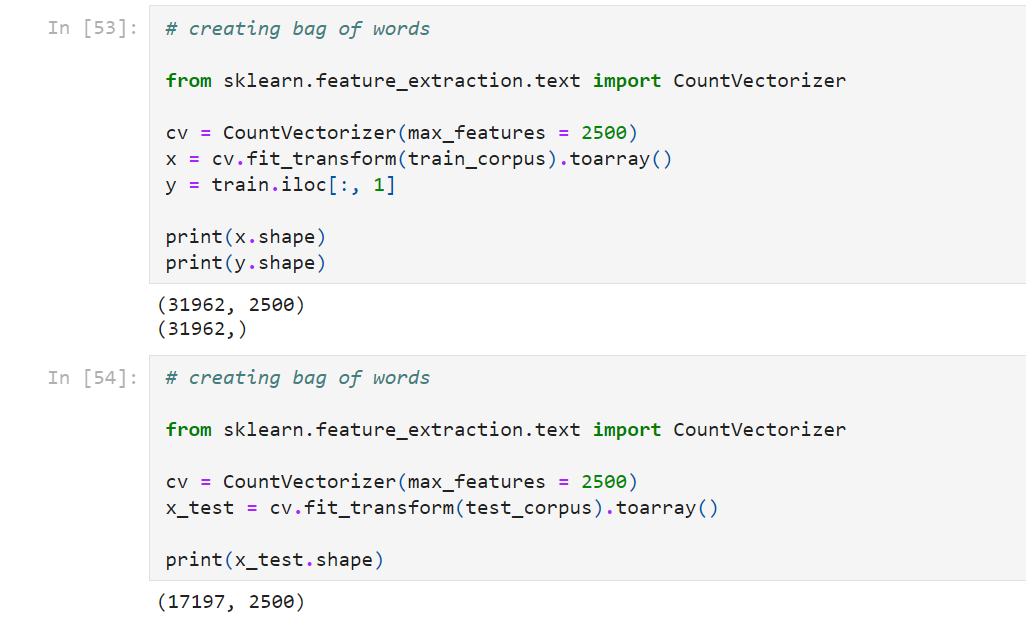
**6. Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses).The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

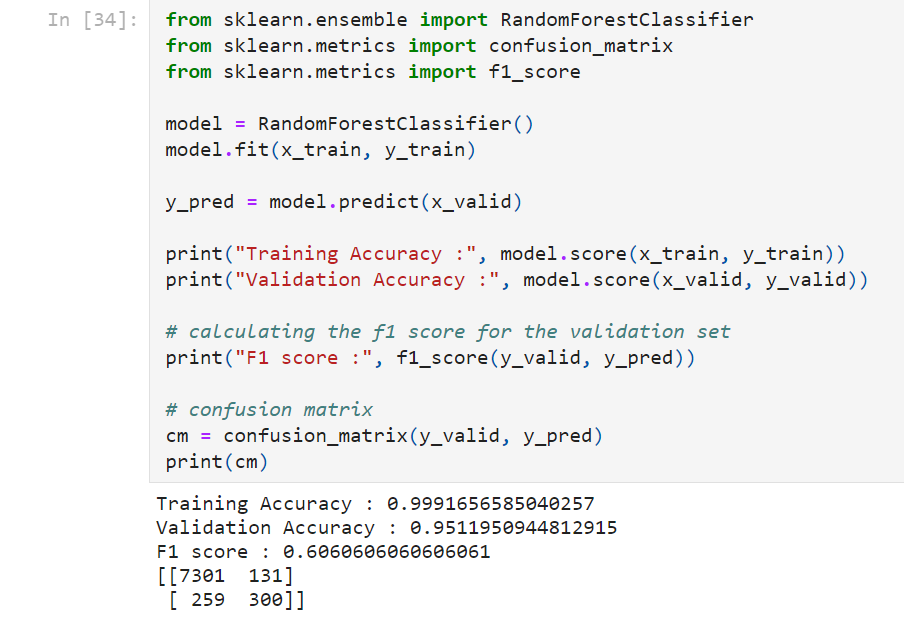
****

**7. Data Transformation**

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

****

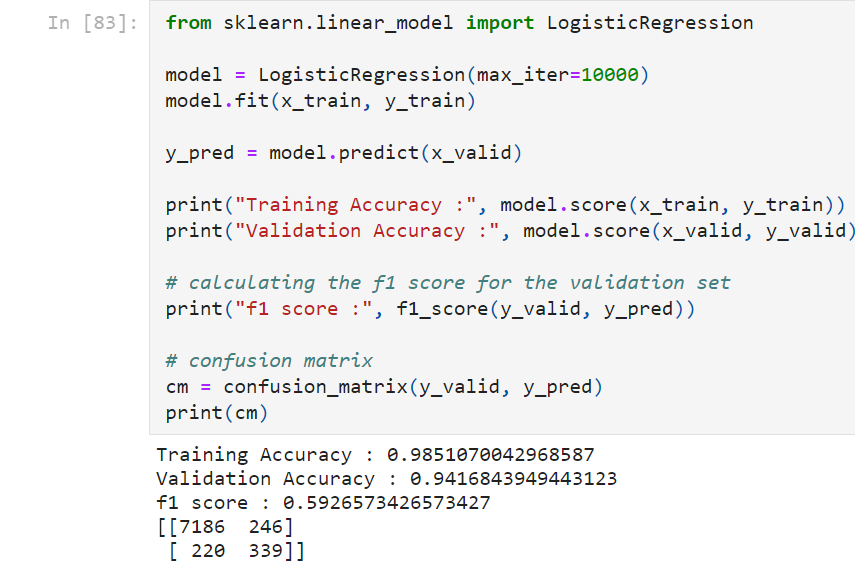
**8. FINAL MODEL SCORE**

****

**For RandomForestClassifier:**

Training Accuracy : 0.9991656585040257

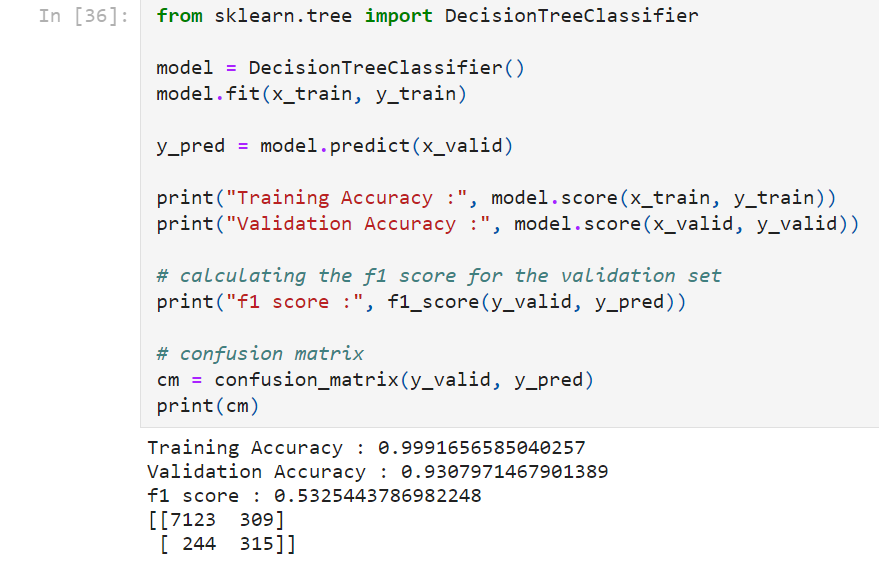
Validation Accuracy : 0.9511950944812915

****

**For LogisticRegression:**

Training Accuracy : 0.9851070042968587

Validation Accuracy : 0.9416843949443123

****

**For DecisionTreeClassifier:**

Training Accuracy : 0.9991656585040257

Validation Accuracy : 0.9307971467901389

Concluding the final results, the highest accuracy of our model comes out to be:

**0.9991656585040257** or **99.91%**

**9. LANGUAGE USED**

● Python

Python is widely used in scientific and numeric computing:

* SciPy is a collection of packages for mathematics, science and engineering.
* Pandas is a data analysis and modeling library.
* IPython is powerful interactive shell that features easy editing and recording of work session, and supports visualizations and parallel computing.
* The Software Carpentry Course teaches basic skills for scientific computing, running bootcamps and providing open-access teaching materials.