PERSONALITY DETECTION FROM TEXT

A PROJECT REPORT



**Submitted By:**

**Mayukh Mitra (1851005)**

**Pushan Pore (1851012)**

**Manjari Nandi Majumdar (1851039)**

**Srinjoy Paul (1851062)**

**Under the Supervision and Guidance of**

**Prof. Amitabha Acharya**

**Assistant Professor,**

**Department of Computer Science and Engineering**

**In partial fulfilment for the award of the degree**

**of**

**Bachelor in Technology**

**In**

**Computer Science and Engineering**

**HERITAGE INSTITUTE OF TECHNOLOGY**

**MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY**

**March, 2022**

**HERITAGE INSTITUTE OF TECHNOLOGY**

**MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY**



**BONAFIDE CERTIFICATE**

**It is hereby certified that this project report “PERSONALITY DETECTION FROM TEXT” is the bonafide work of Mr. Mayukh Mitra, Mr. Pushan Pore, Ms. Manjari Nandi Majumdar and Mr. Srinjoy Paul, who carried out the project under my supervision.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Prof. (Dr.) Subhashis Majumder Prof. Amitabha Acharya**

**Head of the Department Assistant Professor**

**Computer Science and Engineering Computer Science and Engineering**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Prof.**

**Examiner**

**March 2022**

**ACKNOWLEDGEMENT**

**We would like to extend our token of gratitude to the people who enlightened this project with their constant guidance, support and encouragement.**

**We would take this opportunity to thank Prof. Dr. Basab Chaudhuri, Principal of Heritage Institute of Technology Kolkata, for giving us the responsibility, opportunity and resources for the proceedings of this project.**

**We are thankful to Prof. Dr. Subhashis Majumder, Head of the Department of Computer Science and Engineering for his inspiring insights and support.**

**We would like to express our gratefulness to our Project Guide, Prof. Amitabha Acharya for his constructive suggestions, invaluable time and help in the partial fulfilment for the award of the degree.**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Mayukh Mitra**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Pushan Pore**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Manjari Nandi Majumdar  
  
  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Srinjoy Paul**

Table of Contents

Sl. No. Topic Pg. No.

1. Introduction 5

2. Motivation 5

3. Objective 6

4. Methodology 7

5. Algorithm and Approach 7

6. Description of Dataset 8

7. Application Code

* Machine Learning Approach 9
* Deep Learning approach 12
* Support Vector Machine Approach 16

8. Result 18

9. Conclusion 19

# Introduction

Written documents often capture the heart and soul of our thought process. Hence, written texts are preferred over oral contracts. Text messaging, like email and instant messaging, is an emerging means of electronic communication with important implications for our understanding of communication processes. With the developing social media, people are now pouring their hearts out through blogs and write ups.

The spark that ignites here lies in the basic domain of linguistics and vocabulary.

It has been observed through various research domains that the linguistic style, textual tone, vocabulary and grammar of a person is consistent with his or her personality.

The purpose of this project was to undertake some analyses of how the language used in text varies as a function of personality traits and the interpersonal context.

We have opted for multiple models for taking the average of the outcomes for a better accuracy. We use both supervised machine learning, as well as deep learning approach.

# Motivation

**Personality** is a combination of an individual’s behaviour, emotion, motivation, and thought pattern. Our personality has a great impact on our mind, our behaviour our course of. Knowing one’s personality can be very helpful in order to analyse one’s mental health.

Sentiment analysis plays a key role here.

This project has far-reaching real-world applications like:

* Product Recommendations by e-stores
* Mental health check-ups
* Forensic application to understand criminal psychology
* To detect criminal tendency
* To detect depression and suicidal tendencies.

# Objective

In the 1970s two research teams led by Paul Costa and Robert R. McCrae of the National Institutes of Health and Warren Norman and Lewis Goldberg of the University of Michigan at Ann Arbor and the University of Oregon, respectively, discovered that most human character traits can be described using five dimensions. Surveys of thousands of people yielded these largely independent traits. The Big – five personality test is one of the most accepted model for personality definition.

The objective of this model is to detect the absence/presence of five different personality attributes based on texts gathered from various personalities.

These personality attributes are:

* **Extroversion (EXT)**: - Is the person outgoing, talkative, and energetic versus reserved and solitary?
* **Neuroticism (NEU)**: - Is the person sensitive and nervous versus secure and confident?
* **Agreeableness (AGR**): - Is the person trustworthy, straightforward, generous, and modest versus unreliable, complicated, meagre, and boastful?
* **Conscientiousness (CON):** - Is the person efficient and organized versus sloppy and careless?
* **Openness (OPN)**: - Is the person inventive and curious versus dogmatic and

cautious.

# Methodology

* Data Preprocessing
* Date Cleaning.
* Text Preprocessing – coversion to lowercase, removing special characters.
* Document Level Feature Extraction.
* Data Filtering – removing stop words.
* Feature Extraction with Count Vectorizer.
* Classification Model Creation – Using Naïve, Deep Learning and Support Vector Approach.
* Training
* Testing

# Algorithm and Approach

We took **three different approaches** to implement our project, the first one is naïve bayes based on Machine Learning, the next approach deals with Neural Networks, which is based on Deep Learning, while the third approach deals with SVM(support vector machines) shows a high generalization capability due to their structural risk minimization oriented training.

**I. Machine Learning Approach:** In this approach we first clean the gathered

data and store it in a dataframe, after that we remove the unwanted characters and

words and pick-up the words which are then used to create the bag of words, then

we convert the text in the form of vectors and then we build Bayes classifier

model to train and test on our data and then finally we predict the output using the

model built. The multinomial naïve Bayes is widely used for assigning documents to classes based on the statistical analysis of their contents. It provides an alternative to the "heavy" AI-based semantic analysis and drastically simplifies textual data classification.

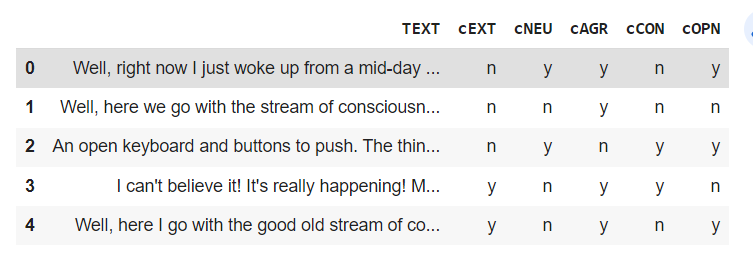
**II. Deep Learning Approach:** In this approach we first clean the gathered data

and store it in a dataframe. We feed sentences from the essays to convolution filters to obtain the sentence. We represented each individual essay by aggregating the vectors of its sentences. We concatenated the obtained vectors with the Mairesse features,4 which were extracted from the texts directly at the preprocessing stage; this improved the method’s performance. then we merge five target attributes into a single target attribute, after that we create the deep learning model of 4 layers using ANN, then we split the data into train and test dataset. After that we measure the accuracy based on the test data of the model built.

**III. Support Vector Machine Approach:** SVM classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points. In this approach we first clean the gathered data and store it in a dataframe, after that we remove the unwanted characters and words and pick-up the words which are then used to create the bag of words, then we convert the text in the form of vectors. We then find support vectors for each personality trait. This gives us the linear discriminant function for mapping the output.

# Description of Dataset

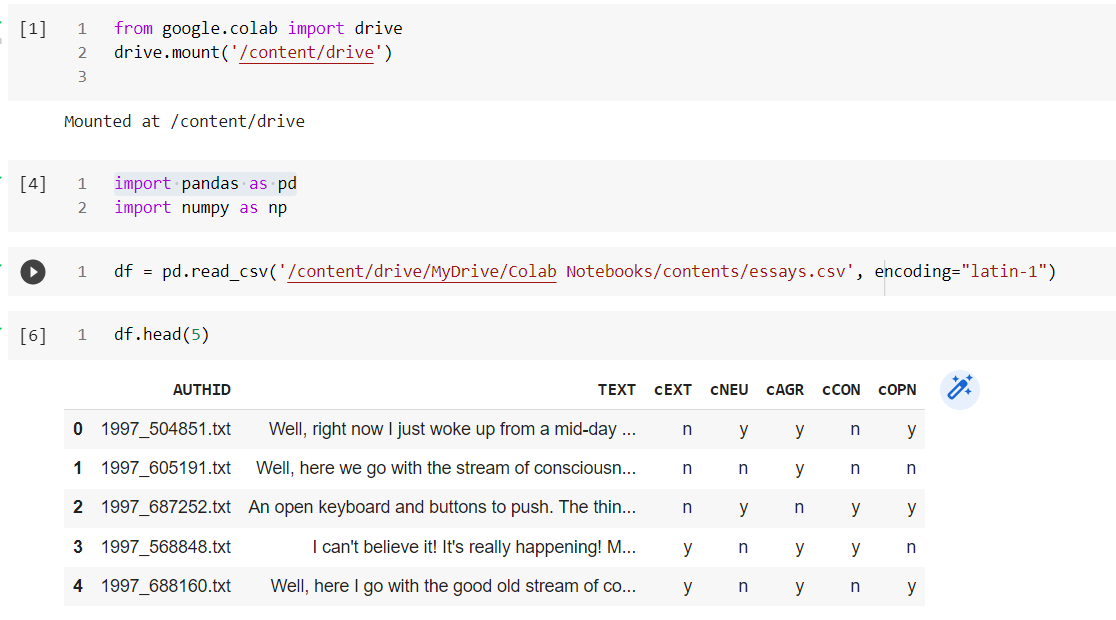
* User Id (unique id of the is user who has typed the specimen text)
* Text (the domain we will work on)
* 5 columns for 5 personality traits – Extroversion, Neuroticism, Agreeableness, Conscientiousness and Openness. These form the five class labels.



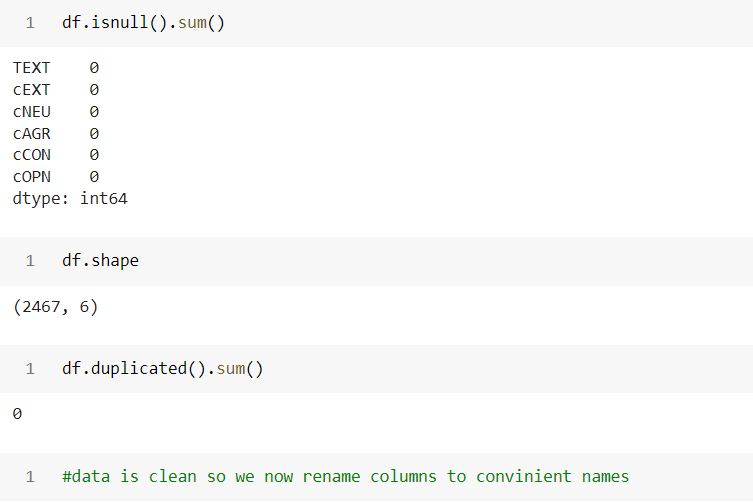
# Application Code

1. **Naïve Approach – Machine Learning Approach**

We first import the data set that has 7 columns:



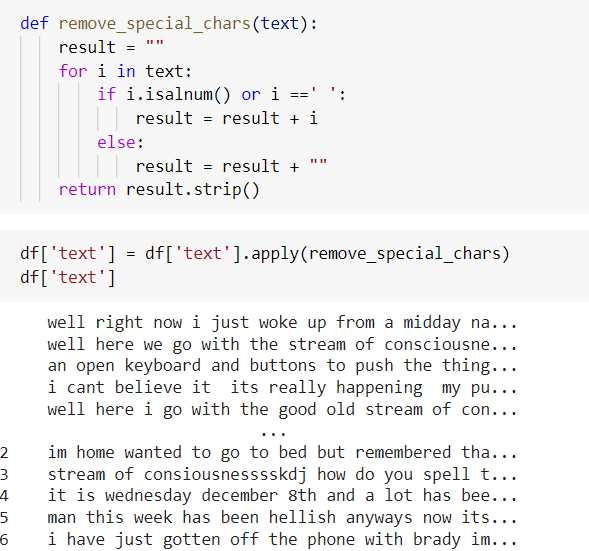
We next clean the data given.



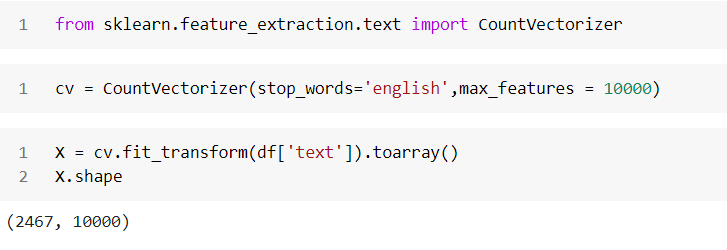
We next pre-process the data- renaming columns and changing the data type to integer.



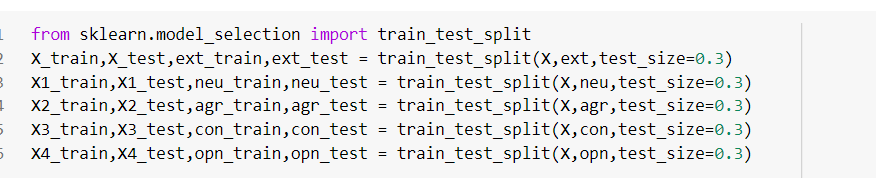
We next hop to text preprocessing – like changing text to small letters entirely, removing special characters, removing stop words.



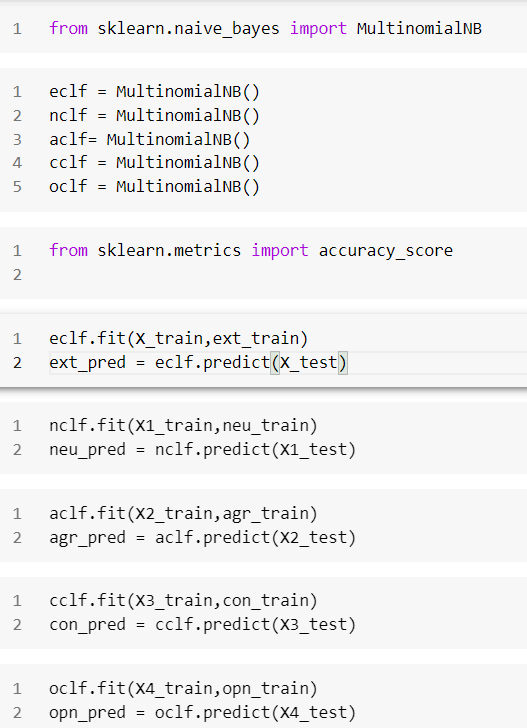
We next convert the text to sentence vectors. This set of vectors is known as Bag of Words. (We have chosen 10000 words here, absence and presence of which will determine the personality trait presence).



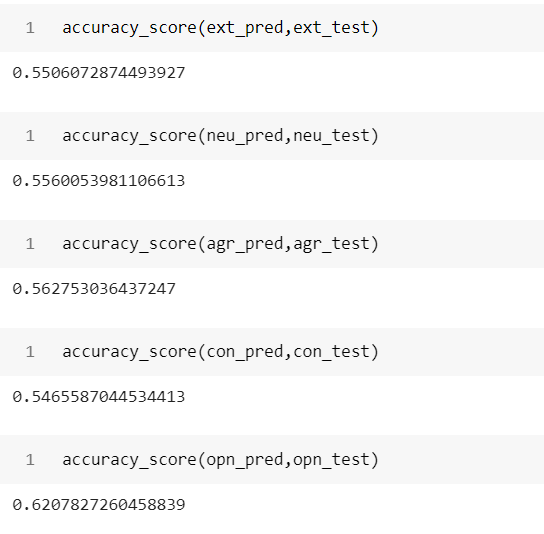
We split the Array into training and testing set.



We create five Multinomial Naïve Bayes Model for the five traits.



We finally determine the accuracy score.



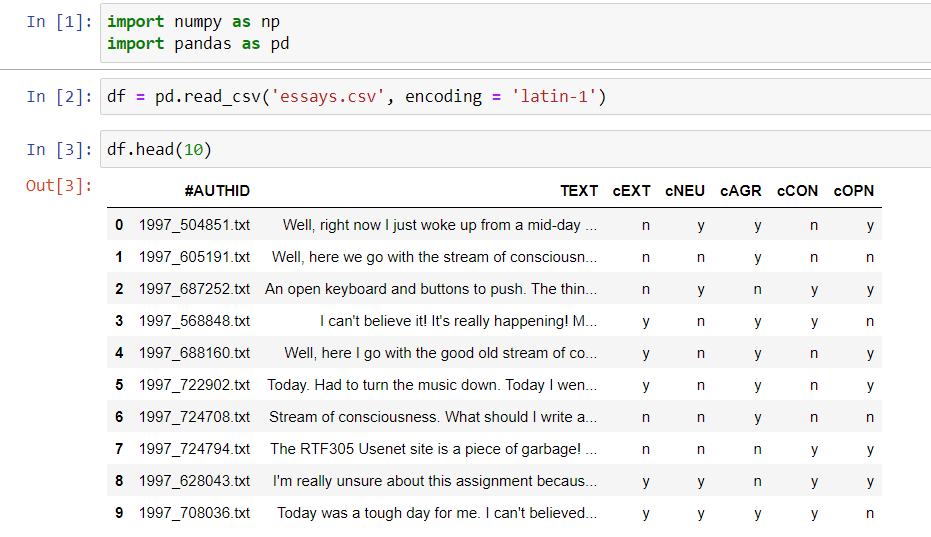
1. **Deep Learning Approach – Artificial Neural Networks**

We first import the data set that has 7 columns:

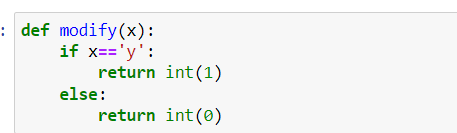
User Id (unique id of the is user who has typed the specimen text)

Text (the domain we will work on)

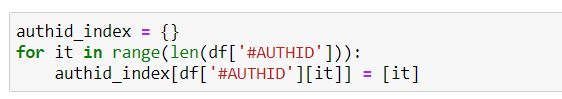
5 columns for 5 personality traits – Extroversion, Neuroticism, Agreeableness, Conscientiousness and Openness.



We preprocess the data.



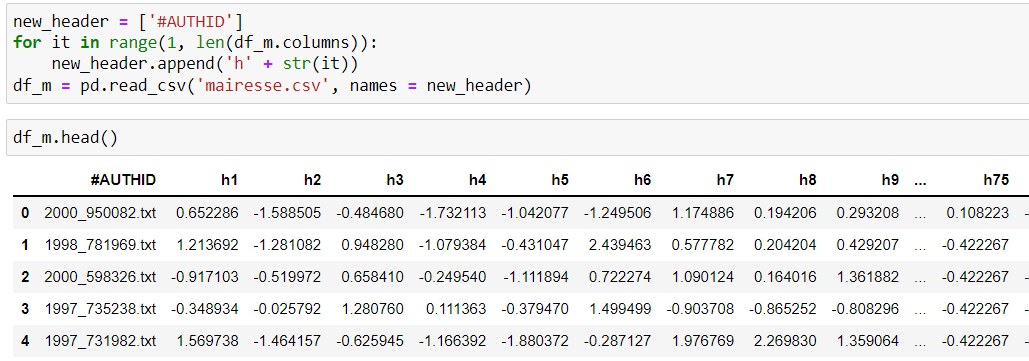
We then extract the author id column.

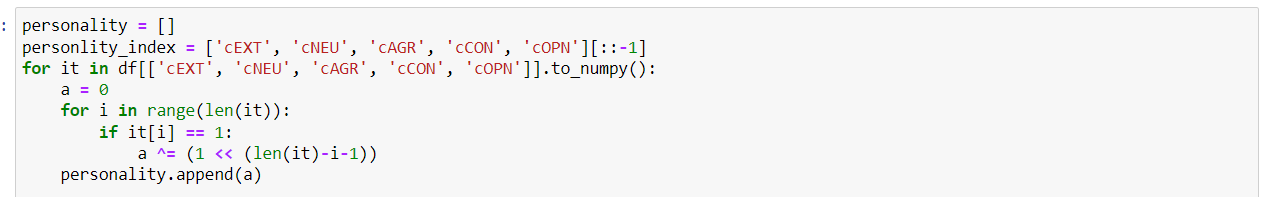


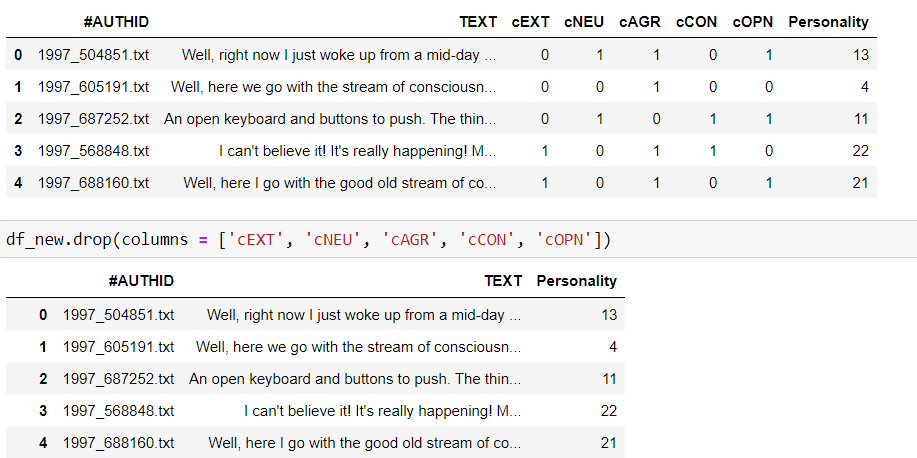
Next, we move on to Document Level Feature Extraction.

We used the Mairesse baseline feature set (10\*85) , which includes such global

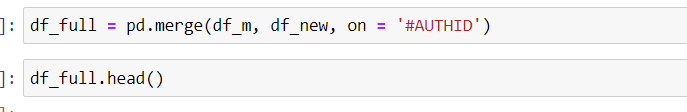
features as the word count and average sentence length. We append the author id with this set to uniquely identify the columns. We rename the columns for better understanding.



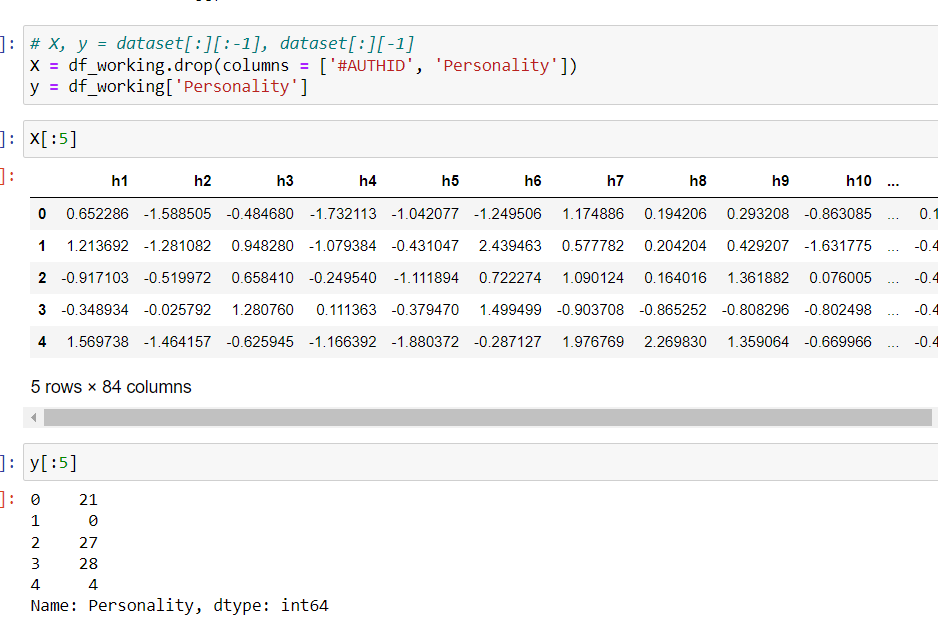
We next combine the five personality traits as a single trait. Here we view the yes(1) and no(0) as a 5 bit number. Hence we can identify this as a decimal number from 0 to 31( 2^5 -1).  




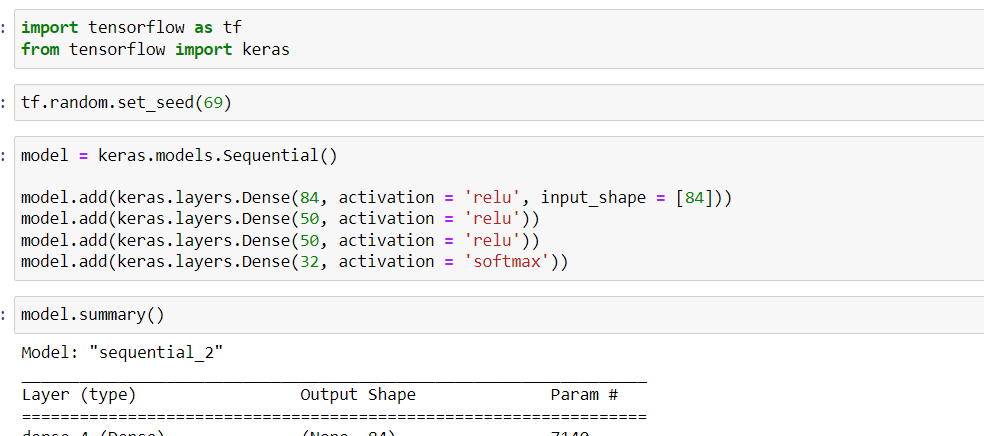
We combine the two datasets.

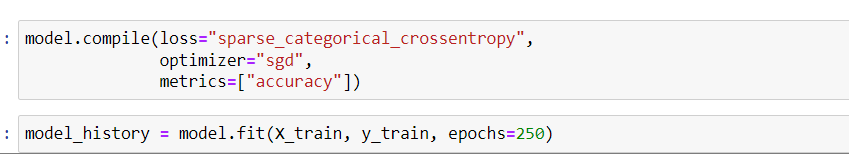


We split the set into training and testing.

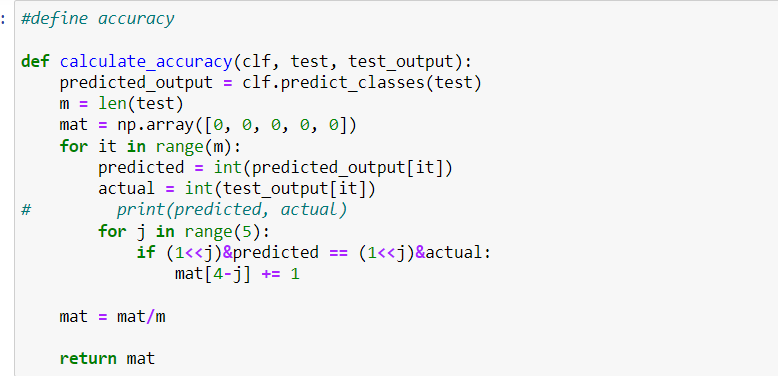


We use ANN model with Keras from Tensorflow. Here we create a 84-50-50-32 Model.

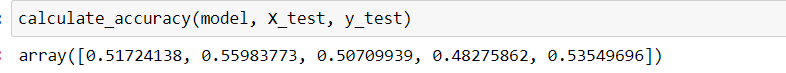




We finally define the accuracy.

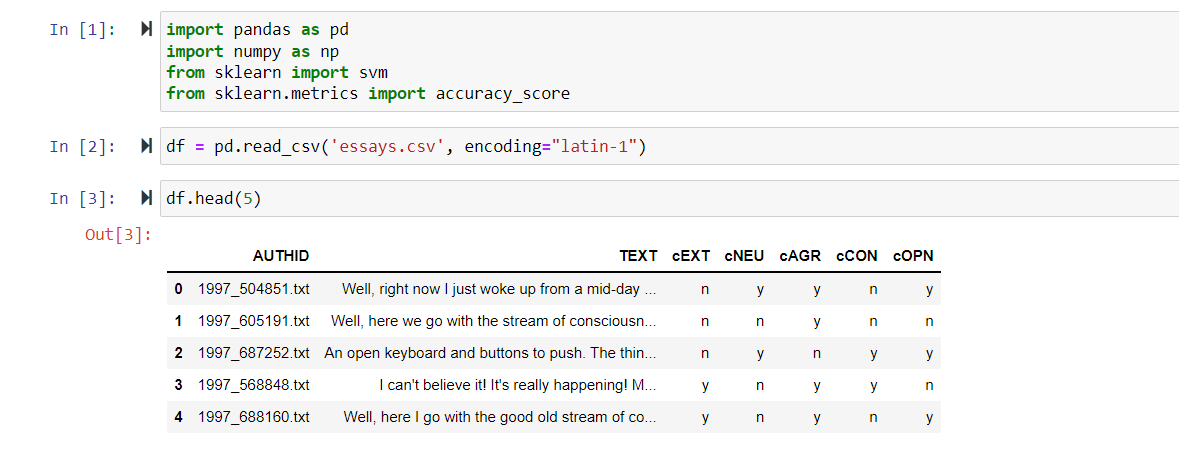


We find the accuracy Score.



**3. SVM – Support Vector Machines :-**

We first import the data set that has 7 columns:



We next clean the data given.



We next pre-process the data- renaming columns and changing the data type to integer.



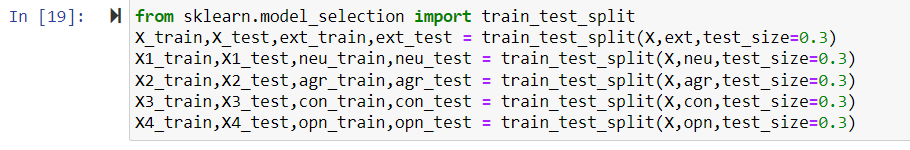
We next hop to text preprocessing – like changing text to small letters entirely, removing special characters, removing stop words.



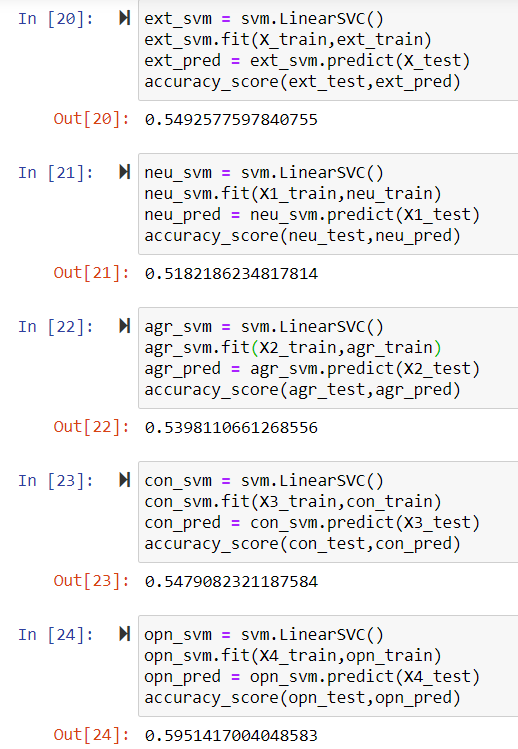
We next convert the text to sentence vectors. This set of vectors is known as Bag of Words. (We have chosen 10000 words here, absence and presence of which will determine the personality trait presence).

+

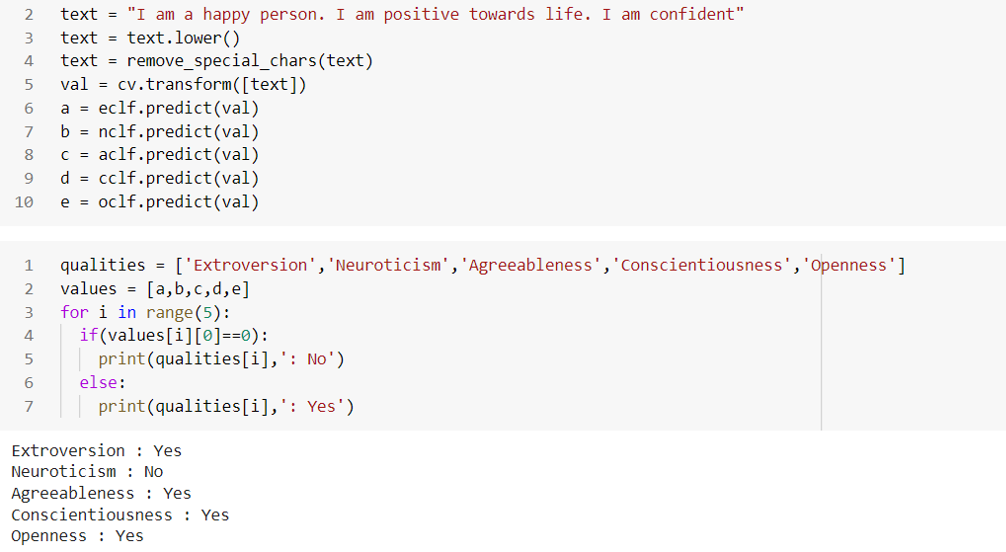
We split the Array into training and testing set.



We create five SVM Models for the five traits. We finally determine the accuracy score.



# Result



Here, the text “**I am a happy person. I am positive towards life. I am confident**.” and the output are consistent. As the person is an **extrovert** enough to boldly state that “he is happy”, **not a nervous** man, **agreeable** and positive, has a **conscious** and is positive towards life, he seems **open hearted**. We have used an example which was not present in the given dataset.

# Conclusion

So here we are using both deep learning as well as machine-learning approach where we found almost similar accuracy for both scenarios.

Naïve Bayes is a very simple approach. ANN is the complicated approach. SVM has the simplicity of Naïve Bayes as well as the technicality of ANN.

|  |  |  |  |
| --- | --- | --- | --- |
| Traits | Naïve Bayes | ANN | SVM |
| Ext | **0.59061** | 0.51724 | 0.54925 |
| Neu | 0.55601 | **0.55983** | 0.51821 |
| Agr | **0.58275** | 0.50709 | 0.53981 |
| Con | **0.56655** | 0.48275 | 0.53549 |
| Opn | **0.64078** | 0.53549 | 0.59514 |

Sometimes, easy paths are the correct ones and it is always not about taking the tough road. This is exactly what we see here. The simplest Naïve Bayes approach yields a comparatively better accuracy.

Purpose of this whole project was to find an effective solution for prediction of personality and mental health behaviour of human beings. And the strategies used, turned out to be quite effective in doing so.