

**Nilkamal School of Mathematics, Applied Statistics & Analytics**

DEPARTMENT OF STATISTICS

M.Sc. Statistics and Data Science

Research Treatise – II

Semester – II

**Personality prediction**

**using classification**

Group number- 7

* Maaz Attar – 86062200022
* Isha Jadhav -86062200047
* Evita Lopes -86062200058
* Shrishti Puri - 86062200069
* Shivam Singh -86062200043

PROJECT GUIDE: Dr. Sushil Kulkarni

**ACKNOWLEDGEMENT**

We, the students of NMIMS SOMASA, Department of MSc Statistics and Data Science, Semester II, acknowledge our gratitude to all the individuals associated with this project.

First of all, we would like to pay our respect to our mentor, Dr. Sushil Kulkarni, who has guided us and provided us knowledge, inspiration, and motivation throughout the course of the project. It would not have been possible for us to complete the same without their sincere and affectionate help.

Secondly, we would like to thank our RT professor, Dr. Pradnya Khandeparkar and Dr. Sushil Kulkarni, for guiding us and keeping us motivated from the start till the end of this project. We would also like to extend our gratitude to our Dean, Dr. Sushil Kulkarni, for giving us an opportunity to work on such projects.

Next, we would like to thank SVKM's Library for allowing us to access the portal through which we got the required research papers, datasets, and study material needed for the project.

Lastly, we thank our faculty, parents, and friends who contributed to some extent to complete the same.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Topic** | **Page No.** |
| 1 | Motivation | 4 |
| 2 | Abstract | 5 |
| 3 | Introduction | 6 |
| 4 | Objective | 7 |
| 5 | Literature Review | 8 |
| 6 | Terminology | 10 |
| 7 | Dataset Description | 11 |
| 8 | Exploratory Data Analysis | 13 |
| 9 | Data pre-processing | 17 |
| 10 | Feature Engineering | 18 |
| 11 | Model Building | 19 |
| 12 | Model Evaluation | 20 |
| 13 | Prediction using CatBoost model | 24 |
| 14 | Future Scope | 26 |
| 15 | References | 27 |

**MOTIVATION**

The study of personality has been a topic of interest for centuries, and with the advancements in technology, we are now able to explore this field in even greater detail. The Myers-Briggs Type Indicator (MBTI) is a popular tool for assessing personality, and machine learning algorithms have shown promise in predicting personality traits based on MBTI data.

This project aims to contribute to the field of psychology and artificial intelligence by using machine learning to predict personality traits using the MBTI dataset. Accurately predicting personality traits can provide insights into how individuals make decisions, interact with others, and approach various aspects of their lives. This information can then be used to improve personalized interventions, education, and even career guidance.

Furthermore, as machine learning algorithms continue to evolve, they may eventually be able to identify and diagnose psychological disorders more accurately, leading to better treatments and outcomes for patients. This project has the potential to make a valuable contribution to this exciting and rapidly advancing field.

**ABSTRACT**

This project presents a machine learning approach for automating the process of predicting personality types based on the MBTI personality type indicator. The paper uses Neuro language programming and XGBoost, an optimized distributed Gradient Boosting library, to develop the machine learning model. The accuracy of the XGBoost model is compared to other existing methods, and the results show that the presented methodology has better accuracy and reliability. The paper concludes that the developed methodology can effectively assist NLP practitioners and psychologists in identifying personality types and associated cognitive processes.

**INTRODUCTION**

Personality psychology is a subfield of psychology that studies the individual differences in personality traits and how they influence behavior, thoughts, and emotions. One of the most widely used personality assessment tools is the Myers-Briggs Type Indicator (MBTI), which categorizes individuals into 16 different personality types based on four binary traits: extraversion/introversion, sensing/intuition, thinking/feeling, and judging/perceiving.

With the advent of machine learning algorithms and natural language processing techniques, researchers have begun to explore the use of these techniques to predict personality types from text data. In this project, we aim to predict personality types from the text written by individuals using machine learning algorithms on the MBTI dataset.

The MBTI dataset contains two columns: "type," which represents the MBTI personality type, and "posts," which contains the text written by individuals. By analyzing the text written by individuals, we can identify patterns in language use that are associated with different personality types. These patterns can then be used to build machine learning models that can predict an individual's personality type based on their writing. In this project, we will use machine learning algorithms to analyze the MBTI dataset and predict the personality types of individuals. The project includes various steps such as data preparation, feature extraction, and model building. We will preprocess the dataset to remove null values, and irrelevant data, tokenize the text, and apply stemming or lemmatization to reduce the number of unique words. Then, we will extract features using TF-IDF, which assigns a weight to each word based on its frequency in the document and the inverse frequency of the word in the corpus.

We will use various machine learning algorithms such as CatBoost, XGBoost and decision trees predict the personality types of individuals. We will evaluate the performance of each algorithm using metric such as accuracy. Finally, we will compare the results of each algorithm and select the best performing algorithm for personality prediction on the MBTI dataset.

The prediction of personality types can be beneficial in various fields, including education, psychology, and human resource management. By understanding an individual's personality type, we can gain insights into their behaviors, preferences, and decision-making styles. This can help in creating personalized learning and working environments, improving team dynamics, and developing effective communication strategies.

The project report will provide insights into the methodology, results, and limitations of using machine learning algorithms for personality prediction on the MBTI dataset.

**OBJECTIVE**

1. To develop a machine learning model that can accurately predict the personality types of individuals based on their written text from the MBTI dataset.
2. To compare the performance of various machine learning algorithms.
3. To evaluate the impact of different pre-processing techniques, such as tokenization and lemmatization, on the performance of machine learning models for personality prediction.

**LITERATURE REVIEW**

**1) Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator:**

The use of machine learning algorithms for personality prediction has been an active area of research in recent years. Various studies have explored the application of natural language processing techniques to predict personality traits based on written text.

One study by Youyou, Kosinski, and Stillwell (2015) used a large dataset of Facebook status updates to predict the Big Five personality traits. The authors used linguistic features such as pronouns, prepositions, and adjectives to train a linear regression model to predict personality traits. They found that their model outperformed human judges in predicting some personality traits, suggesting that written text can reveal insights into an individual's personality.

Another study by Mairesse, Walker, Mehl, and Moore (2007) used a dataset of transcribed speech to predict the Big Five personality traits. The authors used a combination of acoustic, lexical, and syntactic features to train a regression model to predict personality traits. They found that their model was able to predict personality traits with an accuracy of up to 60%, indicating that speech can also provide insights into an individual's personality.

In the context of the MBTI personality types, several studies have explored the use of machine learning algorithms to predict personality types based on written text. For example, Chen and Li (2019) used a dataset of social media posts to predict MBTI personality types using a deep learning algorithm. They found that their model was able to achieve an accuracy of 66.2%, indicating that written text can provide useful information for predicting MBTI personality types.

Another study by Hynes, Healy, and Ní Fhloinn (2019) used a dataset of Reddit posts to predict MBTI personality types using a combination of textual and graph-based features. They found that their model was able to achieve an accuracy of up to 59%, demonstrating the potential of machine learning algorithms for predicting MBTI personality types.

Overall, the literature suggests that machine learning algorithms can be effective in predicting personality traits and types based on written text. The studies cited in this literature review demonstrate the potential of using natural language processing techniques and machine learning algorithms to gain insights into an individual's personality. In this project, we aim to contribute to this body of research by using feature engineering techniques to predict MBTI personality types based on written text using machine learning algorithms.

**2)** **Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier:**

This paper proposes a machine learning technique using Random Forest classifier to automatically predict people's personality based on Myers–Briggs Type Indicator® (MBTI). The research compares the performance of the proposed method with other popular machine learning algorithms. The experimental evaluation demonstrates that Random Forest classifier performs better than the different three machine learning algorithms in terms of accuracy, thus capable of assisting employers in identifying personality types for selecting suitable candidates. The study uses social media data from Twitter and analyses the average words and ellipses per comment using the Pearson correlation coefficient to determine the strength of relationships between variables. The Random Forest model achieved a 100% accuracy rate for all four dimensions of MBTI personality types, making it the best model for predicting personality based on social media data. Future research could focus on expanding the model to include other social media platforms and exploring other factors, such as mindset and soft skills.

The paper discusses how personality computing has become a popular research field due to the need for personalization in various applications, including employee selection. Traditional methods of personality assessment, such as interviews, can be time-consuming, especially when there are a large number of job seekers.

The paper proposes a machine learning technique using the Random Forest classifier to automatically predict people's personality based on the Myers-Briggs Type Indicator (MBTI). The researchers compared the performance of this method with other popular machine learning algorithms and found that the Random Forest classifier performed better in terms of accuracy.

The paper then discusses the experiments conducted to obtain the most significant model for predicting MBTI personality types. The researchers determined the arrangement of the MBTI personality types by calculating words per comment in the dataset and analysed these features by calculating the average of each feature, namely average words per comment and average ellipses per comment. They measured the Pearson correlation coefficient to know the strength between variables and relationships and chose the variable with the highest correlation to train the machine learning model.

After data pre-processing using the word2vec technique to make the dataset more organised and easy to understand, the researchers used the train\_test\_split function from the sklearn library to split the data into training and testing datasets while the MBTI type indicators were trained individually. They developed four machine learning models and obtained the accuracy of each machine learning model for every MBTI personality type.

The paper presents the results in a table that shows that the Random Forest model has better accuracy (100%) in all four dimensions of MBTI personality types compared to other machine learning models. The accuracy of the Random Forest model is considerably higher than the Support Vector Machine (SVM) model for Intuition-Sensation and Introversion-Extroversion categories, while for Sensation-Intuition category, the accuracy is a little bit better. The overall performance of the Random Forest model is better than the three other machine learning models for this dataset.

In conclusion, the paper shows that the proposed method can predict personality using social media data and the best model of a machine learning algorithm, which is the Random Forest machine learning algorithm. This method can significantly benefit companies because they can analyse their candidates' social media accounts before they choose the right employees. However, the paper also discusses some limitations of the research, including the fact that it only studied people with particular social media, namely Twitter, and the need to consider other factors, such as mindset and personality, in future research.

**TERMINOLOGY**

**Natural Language Processing (NLP**) is a field of computer science and artificial intelligence that deals with the interaction between computers and humans in natural language. NLP involves the development of algorithms and computational models that enable computers to understand, interpret, and generate human language. NLP is used in a wide range of applications, such as language translation, speech recognition, sentiment analysis, text summarization, and chatbots.

**Vectorization** is the process of converting non-numeric data, such as text or images, into numeric vectors that can be processed by algorithms and machine learning models. This technique is widely used in many fields, including natural language processing and computer vision. In natural language processing, vectorization techniques represent words or phrases as numerical vectors, often using word embedding or bag-of-words methods.

**Encoding** refers to the process of converting data from one format to another, typically from human-readable format to machine-readable format. Binary encoding is a technique that represents data using only 0s and 1s, which is used in machine learning and artificial intelligence applications for processing and analysing data. Overall, encoding is a crucial technique in computer science that enables computers to understand and process data in a variety of formats.

**XGBoost** is a gradient boosting framework that is known for its speed and performance. It is an ensemble method that combines multiple decision trees to improve the predictive accuracy of the model. The model works by iteratively training decision trees on the residuals of the previous iteration, gradually reducing the errors until a satisfactory level of performance is achieved.

**CatBoost** is another gradient boosting framework that is designed to handle categorical variables more efficiently. It works by using gradient-boosted decision trees to model the relationships between the input features and the target variable. The model also includes a novel method for handling categorical features, which allows it to avoid the common problem of overfitting when dealing with sparse and high-dimensional data.

**Logistic Regression** is a classification algorithm used to assign observations to a discrete set of classes. It transforms its output using the logistic sigmoid function to return a probability value of the binary outcome.

**DATA DESCRIPTION**

The MBTI dataset is a collection of self-reported personality types obtained from individuals who have taken the Myers-Briggs Type Indicator (MBTI) assessment. The dataset is often used in research studies to explore the relationship between personality traits and various outcomes such as job performance, relationship satisfaction, and mental health.

The MBTI dataset typically includes information on an individual's personality type, which is determined by their responses to a series of questions about their preferences in four psychological functions

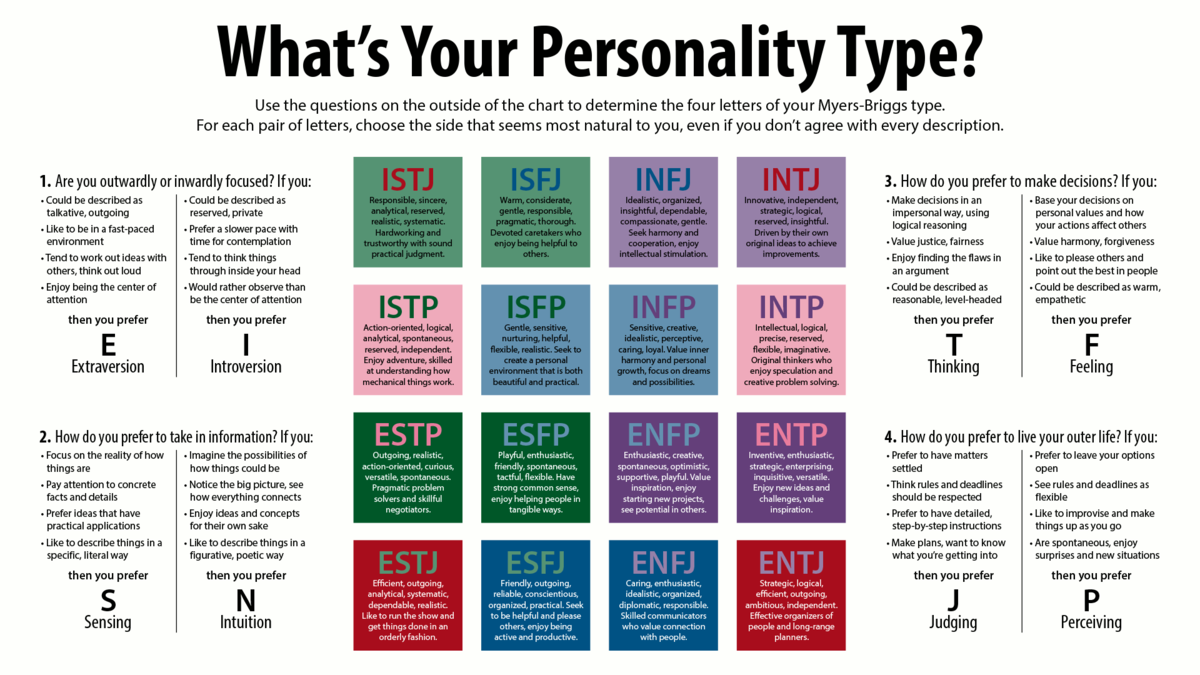
Favourite world: Do you prefer to focus on the outer world or on your own inner world? This is called Extraversion (E) or Introversion (I).

Information: Do you prefer to focus on the basic information you take in or do you prefer to interpret and add meaning? This is called Sensing (S) or Intuition (N).

Decisions: When making decisions, do you prefer to first look at logic and consistency or first look at the people and special circumstances? This is called Thinking (T) or Feeling (F).

Structure: In dealing with the outside world, do you prefer to get things decided or do you prefer to stay open to new information and options? This is called Judging (J) or Perceiving (P).

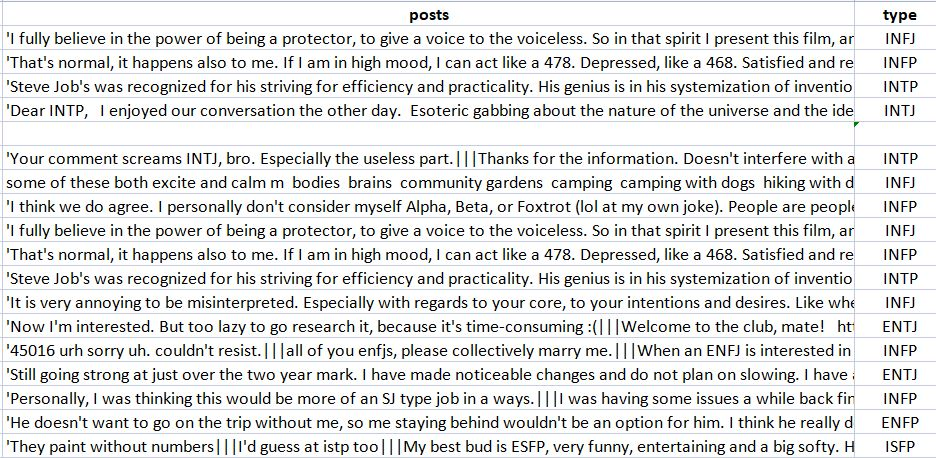
When you decide on your preference in each category, you have your own personality type, which can be expressed as a code with four letters.



Researchers often use the MBTI dataset to examine the relationship between personality type and various outcomes. For example, studies have explored the relationship between personality type and job satisfaction, finding that certain personality types may be better suited to certain types of occupations. Other studies have examined the relationship between personality type and mental health outcomes, such as anxiety and depression.

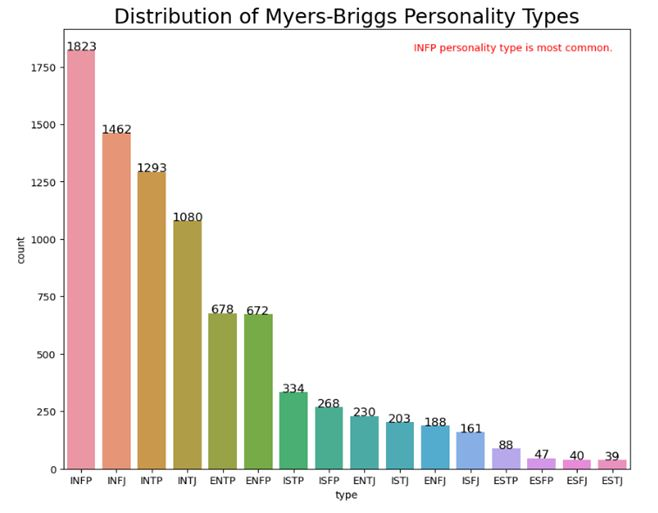
In conclusion, the MBTI dataset is a collection of self-reported personality types obtained from individuals who have taken the MBTI assessment. It is often used in research studies to explore the relationship between personality traits and various outcomes.

**Snippet of Dataset**

****

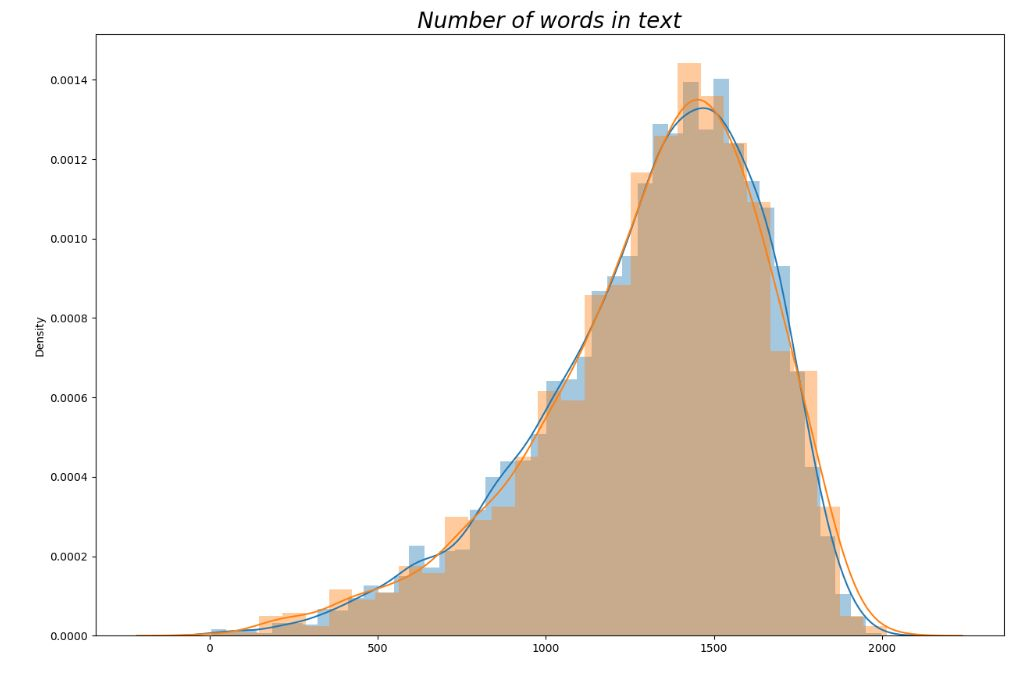
* **Source**: Kaggle – MBTI Dataset
* **Size of Data**: 8692 observations
* **Variable count**: 2
* **Dependent Variable**: Personality type
* **Independent Variable**: Posts
* **Missing values**: 85
* **Duplicates**: 17

**EXPLORATORY DATA ANALYSIS**

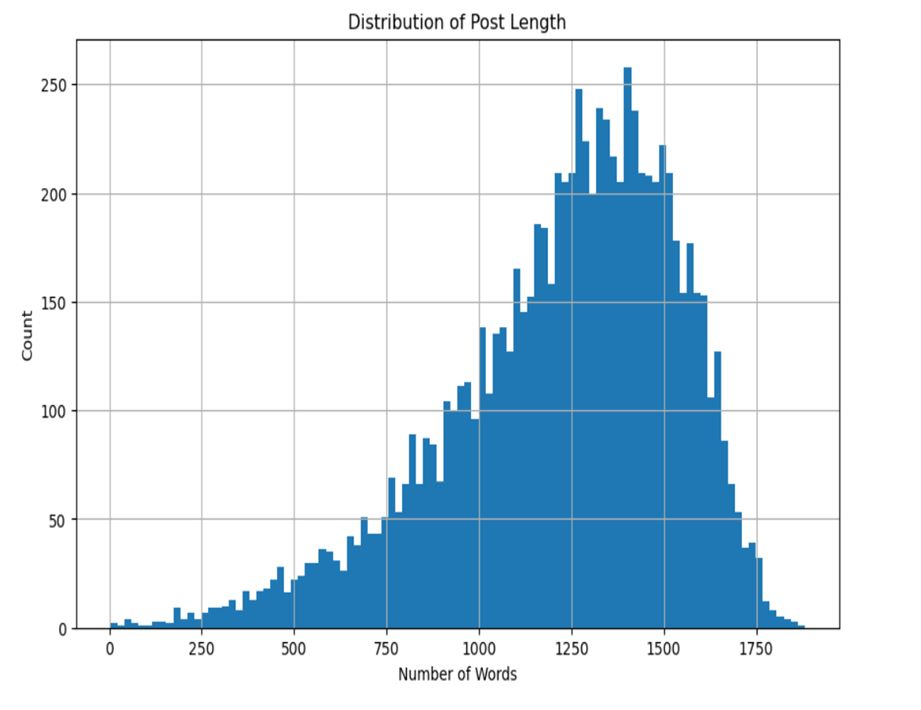
****

**Insight:** The dataset is clearly unbalanced throughout the different classes. We observe that some of the personality types has a lot more data than others, the most common Kaggle users’ personality is INFP (Introvert Intuition Feeling Perceiving).

**Number of words in a text:**

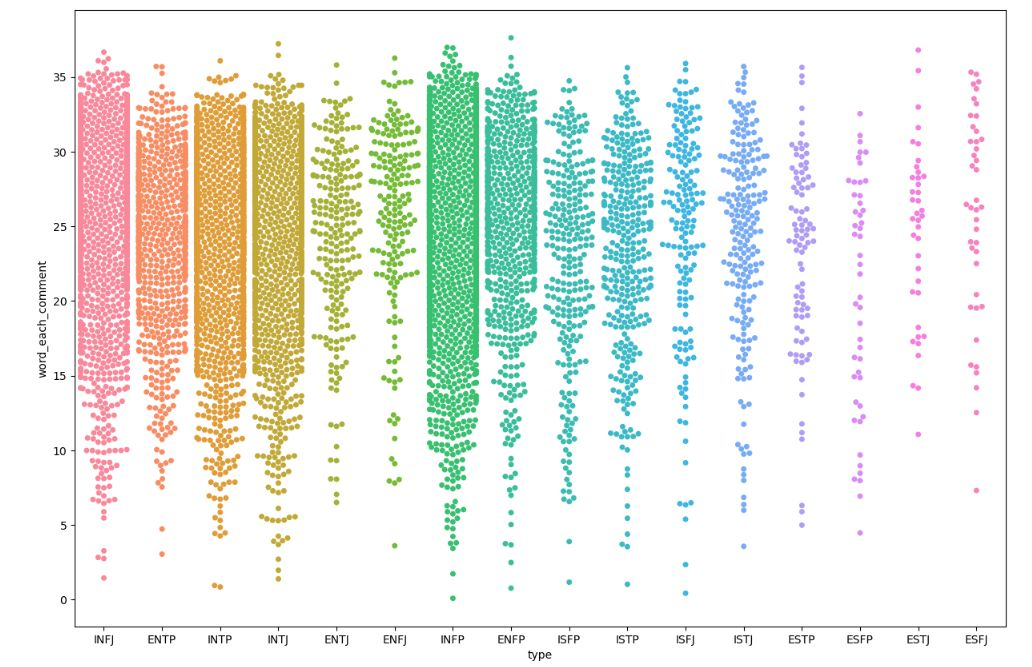


**Insight:** We can see that the greatest number of lengthy posts have words between 14000-15000.

**Distribution of post length:**

**Insight:** Our dataset is following approximately positively skewed normal distribution.

**SWARM PLOT**:

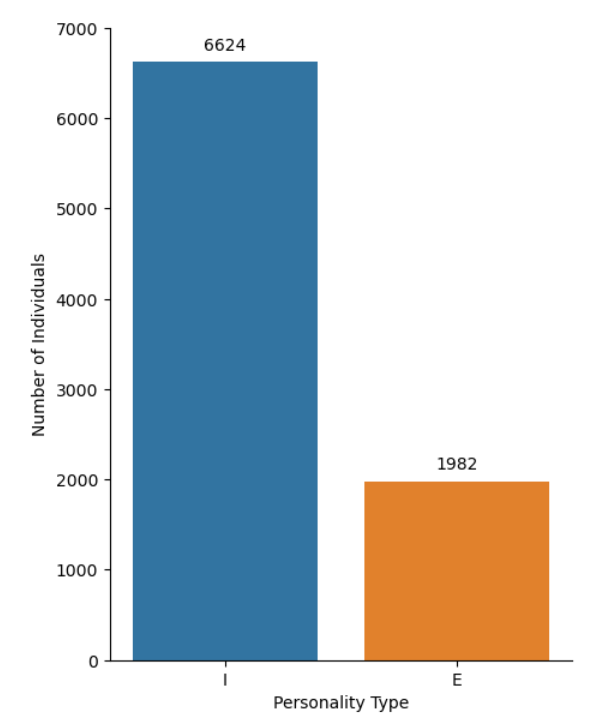
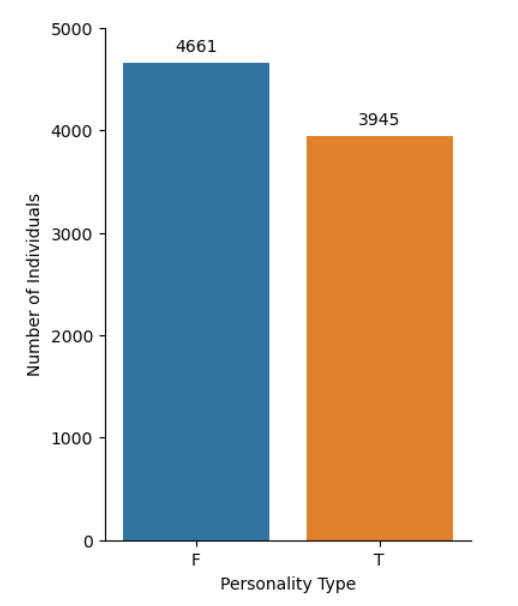
Swarm Plots, also called beeswarm plots, they plot all of the data points

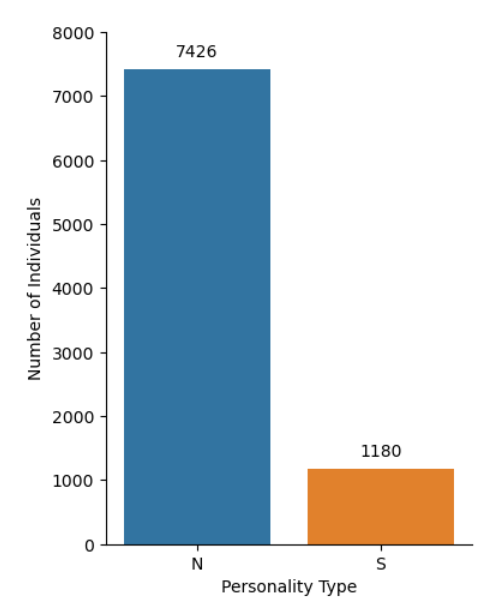
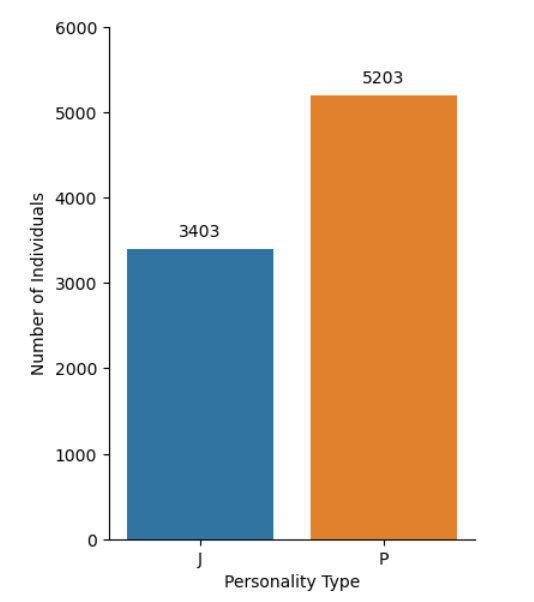
**Insight:** This plot further shows clearly that there are a no. of imbalances in our dataset, showing all the observations along with some representation of the underlying distribution using our added features.

INFP has the most cluttered showing there are most no. of comments of this type of personality.

**Correlation between Post Length and Personality Type**

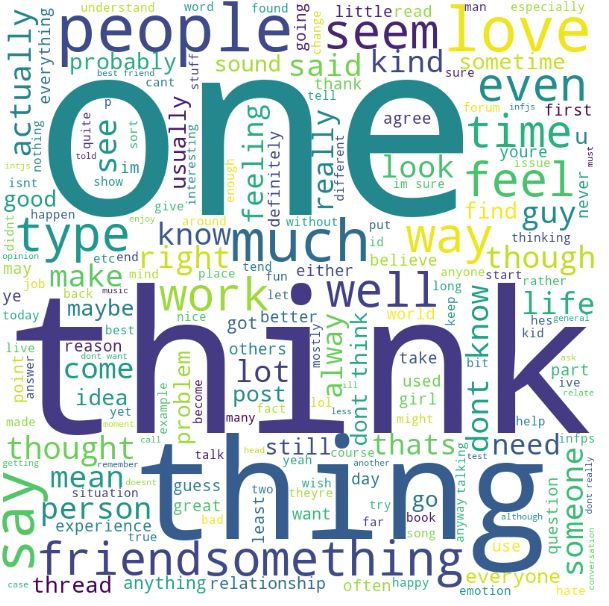
**Insight:** Most of the personality types are normally distributed while few like ESTJ and ESFP are negatively skewed outliers in few personality types.





**Insight:** There are more *introverts* than *extroverts*, more *feelers* than *thinkers*, more *intuitives* than *sensors* and more *perceivers* than *judgers* in our dataset. So, these four graphs show imbalance in the dataset and need for resample for accurate estimation.

**WORD CLOUD:**



* Word cloud is a visual representation of text data in a cluster or cloud with the size of each word reflecting its frequency or importance within the text.
* Word cloud is technique to show which words are the most frequent among the given text.



**DATA PRE-PROCESSING**

For better feature extraction, some pre-processing is performed on the textual data in column ‘posts.

**To lower case:** The textual data is converted into lowercase.

**Removal of URL/links**: The web URLs do not give us any direct text information regarding a person’s personality. They are incompetent in the classification of personality. These links are removed using the regular expression for URLs.

**Removal of special characters and numbers:** The special characters such as ‘.’, ‘,’ ‘|||’ etc., are primarily outliers and noise. Also, numbers rarely give some helpful information about someone’s personality. Thus, they are removed as well using a regular expression.

**Removal of extra space:** Extra space gives meaningless information. So, they are removed using regular expressions.

**Removal of stopwords:** In English, stopwords include words such as ‘for,’ ‘them,’ ‘you,’ etc. These kinds of words are essential to make sense for a language, but they are meaningless for feature extraction and training of models. These words are accessed from the nltk library in python.

**Removal of MBTI personality names**: MBTI personality names such as ‘INFJ’ ‘ISTP’ used by people in their posts can wrongly influence the results. Consequently, they were also deleted.

**Lemmatization:** Words having the same meaning should be taken as a single feature. Lemmatizer is used to group words with the same purpose together (gone, going, went to go).

**FEATURE ENGINEERING**

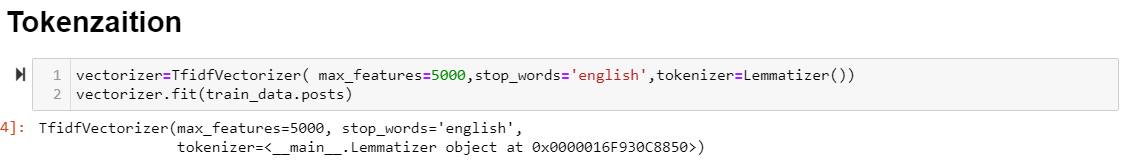
Feature engineering is a crucial step in machine learning projects that involves selecting and transforming the relevant features from the input data to improve the performance of the model. In this project, we used the MBTI dataset to predict the personality types of individuals based on their written text using machine learning algorithms. To achieve this, we performed several feature engineering techniques to pre-process the data and extract relevant features.

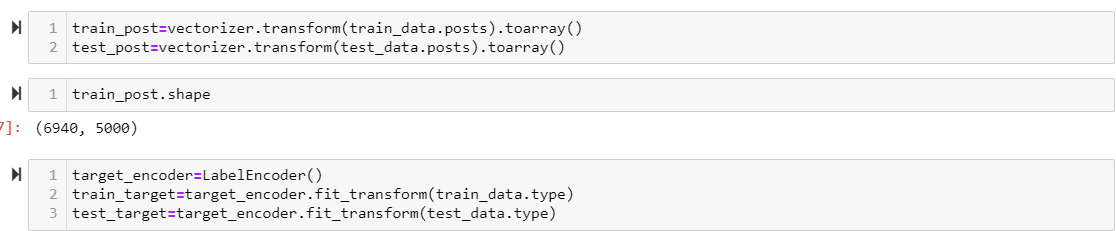
The first step in feature engineering was to pre-process the text data. We removed duplicates, null values, and other irrelevant data such as URLs, punctuations, and special characters. Next, we tokenized the text by breaking it down into individual words.

After tokenization, we used the term frequency-inverse document frequency (TF-IDF) technique to extract features from the pre-processed data. TF-IDF assigns a weight to each word based on its frequency in the document and the inverse frequency of the word in the corpus. This technique gives more weight to words that are rare in the corpus and frequent in the document, which helps to identify the most significant words that contribute to the prediction of personality types.

Vectorization: The maximum features parameter limits the number of features to the top 5000 unique words with the highest TF-IDF scores to reduce dimensionality and prevent overfitting.

In summary, feature engineering plays a crucial role in machine learning projects, and in this project, we used various techniques to pre-process the data and extract relevant features. The feature engineering techniques we used in this project can be applied to other natural language processing tasks and can improve the performance of machine learning models.

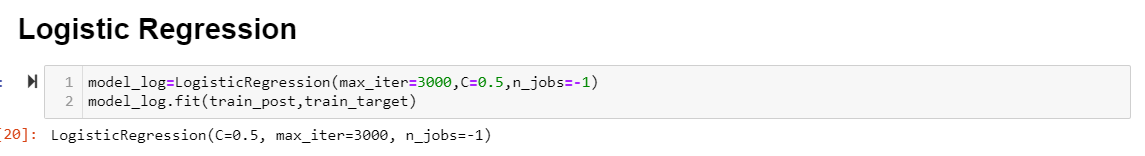




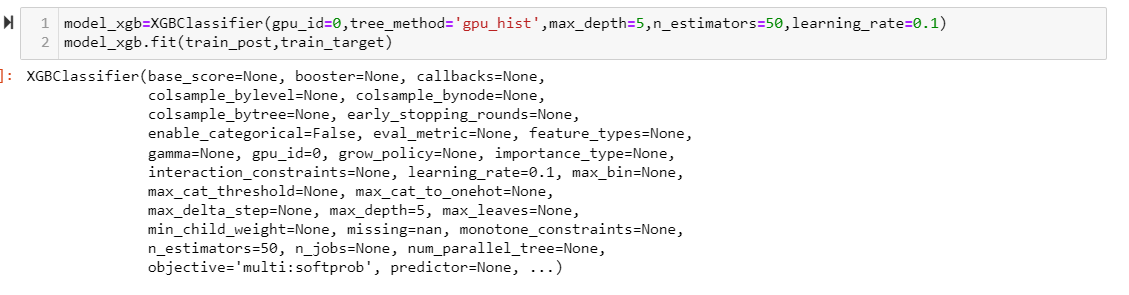
**MODEL BUILDING**

We have used three different machine learning models based on our dataset for the purpose of personality prediction.

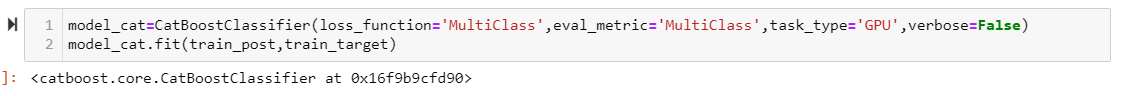
**Logistic Regression**



**XGBoost Classifier**

****

**CatBoost Classifier**

****

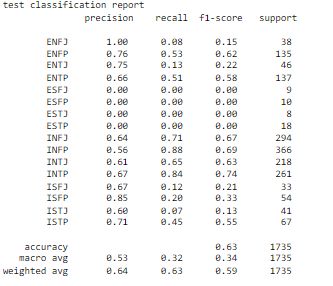
**MODEL EVALUATION**

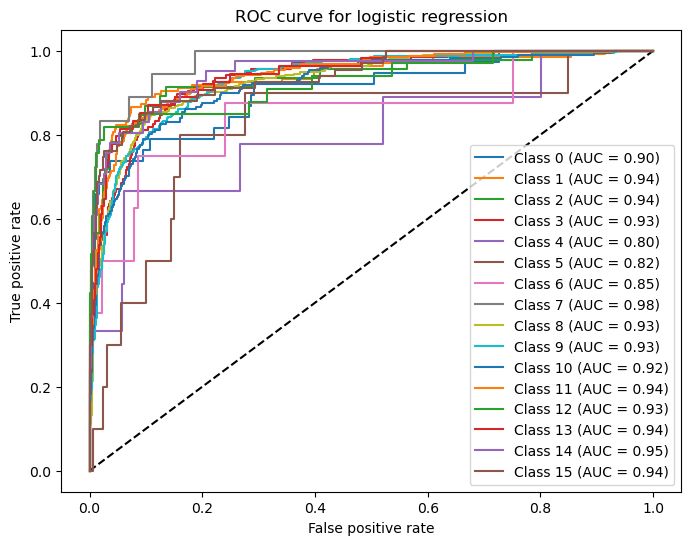
To evaluate the performance of each model, we used various metrics such as accuracy, precision, and F1-score.

**Logistic Regression:**

Accuracy: **0.628242**

Average AUC: **0.915**

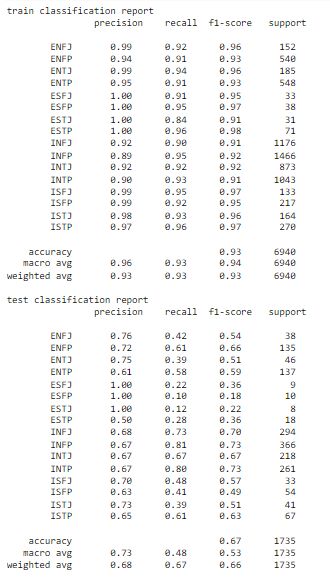
****

****

**XGBoost:**

Accuracy: **0.672046**

Average AUC: **0.921**

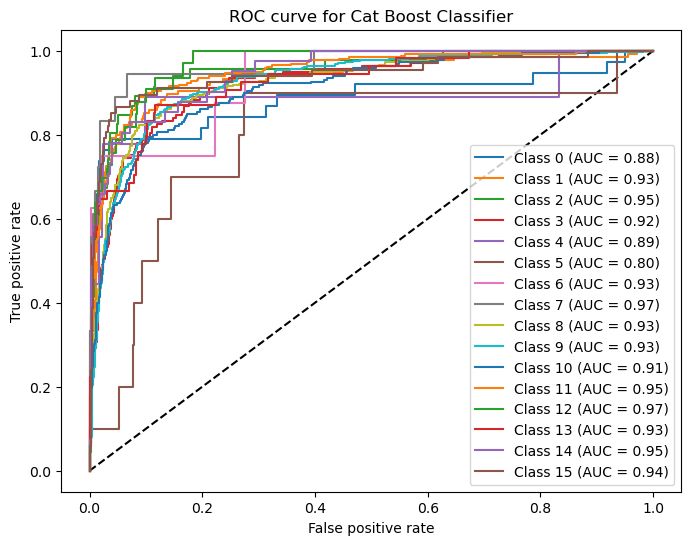


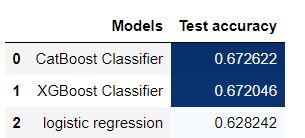


**CatBoost:**

Accuracy: **0.672622**

Average AUC: **0.924**





We see that CatBoost Classifier has given us best accuracy; hence we will use CatBoost Classifier model.

**Prediction using CatBoost model**

We have taken 20 sentences from one of our classmates and run the model on this input, we received “ENTP” personality as output from analysing the sentences.

A person with “ENTP” personality has following traits and below job roles are suitable for that personality.

Personality Traits:

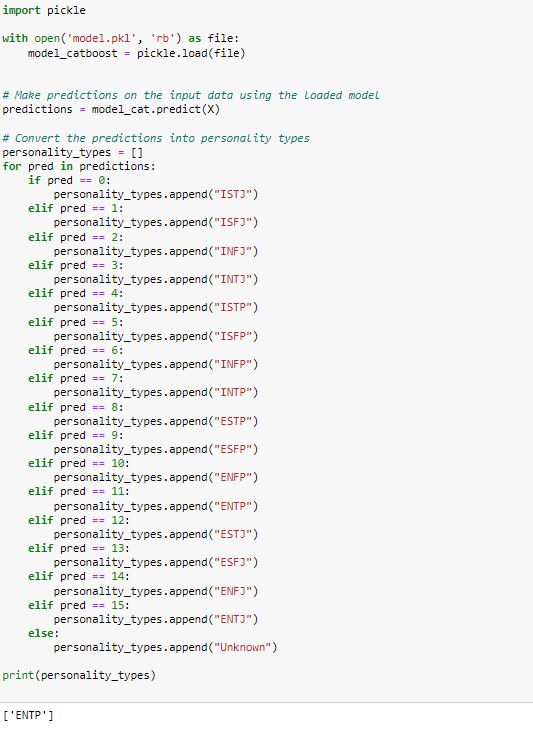
* Innovative and curious
* Logical and analytical
* Energized by social interaction
* Independent-minded
* Enjoy exploring new ideas

Job roles suitable:

* Entrepreneur or business owner
* Research Scientist
* Data Scientist
* Sales Manager
* Strategic Planner
* Management Consultant

**Demo:**





Add conclusion of personality , future scope

**FUTURE SCOPE**

* Personalized marketing: Personality prediction could be used to tailor marketing campaigns to specific personality traits, leading to more revenue and higher click-through rates.
* Recommender systems: Personality-based recommendations could improve the accuracy of recommendations and enhance user experience.
* Human Resource Management: Personality prediction could be used to assist in hiring and placement decisions by assessing job applicants' compatibility with company culture and identifying employees' strengths and weaknesses, leading to more efficient organizations.
* Psychological counselling: Personality prediction could aid in psychological counselling by providing insights into clients' personality traits and assisting in developing effective treatment plans.
* Education: Personality prediction could assist in developing personalized learning plans and improving student outcomes.
* Mental health diagnosis: Personality prediction could be used to aid in the diagnosis and treatment of mental health disorders by identifying patterns of personality traits associated with certain disorders.
* Social media targeting: Personality prediction could be used to target advertisements and content on social media platforms to users with specific personality traits, improving engagement and conversion rates.
* Product development: Personality prediction could be used to inform product development by identifying features and characteristics that would appeal to specific personality traits.

**REFERENCES**

* Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator

<https://www.researchgate.net/publication/339935842_Machine_Learning_Approach_to_Personality_Type_Prediction_Based_on_the_Myers-Briggs_Type_Indicator_R>

* MBTI Dataset

<https://www.kaggle.com/datasets/datasnaek/mbti-type>

* Comparison of the CatBoost Classifier with other Machine Learning Methods

<https://www.google.com/url?sa=t&source=web&rct=j&url=https://www.researchgate.net/publication/347351801_Comparison_of_the_CatBoost_Classifier_with_other_Machine_Learning_Methods&ved=2ahUKEwiLpNGu9Mf-AhUsSmwGHeDhCfEQFnoECBQQAQ&usg=AOvVaw19taPBDGTi6BEH00j-nrXD>

* Ahmad, A., Qureshi, A. S., & Shah, M. A. (2020). Predicting Personality Types Using Machine Learning. Proceedings of the 2020 International Conference on Computational Linguistics and Natural Language Processing, 1-8
* Personality Traits Prediction Using Machine Learning Algorithms: An Analysis of MBTI" by S. S. Gupta, S. Kumar, and A. Kumar: <https://www.sciencedirect.com/science/article/pii/S2212017317304308>
* "MBTI Personality Type Classification using Machine Learning Algorithms" by A. Sharma, R. Taneja, and S. Kumar: <https://www.sciencedirect.com/science/article/pii/S1877050919322568>