

**An Internship Project Report on**

**MBTI Calculator**

Submitted

In partial fulfillment of the requirement for the

Internship in Machine Learning & Artificial Intelligence during the academic year 2022-2023

**Submitted By**

**Nidhi M S 1GA19CS098**

**(GLOBAL ACADEMY OF TECHNOLOGY)**

**Under the guidance of**

Mr. Shashank G

Technical Lead

QuantMasters Pvt. Ltd.

**Quant Masters**

#812, 6th cross 3rd main,

Rajajinagar, Bengaluru - 560021

**CONTENTS**

|  |  |
| --- | --- |
| **Chapters** | **Page No** |
| **ABSTRACT** | **3** |
| **Chapter-1 Introduction**  **1.1 MBTI**  **1.2 Text Classification**  **1.3 Problem Statement** | **4**  **5**  **7** |
| **Chapter-2 Related Work** | **8** |
| **Chapter-3 Proposed Methodology**  **3.1 Dataset**  **3.2 Dataflow**  **3.3 Model Creation** | **9**  **9**  **10** |
| **Chapter-4 Results** | **12** |
| **Chapter-5 Conclusion** | **13** |

**ABSTRACT**

The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axes:

* Introversion (I) — Extroversion (E)
* Intuition (N) — Sensing (S)
* Thinking (T) — Feeling (F)
* Judging (J) — Perceiving (P)

The purpose of this project is to investigate if any patterns exist from which we can detect a person’s specific personality type in connection to their style of writing and to explore the validity of the test in analyzing, predicting, or categorizing behavior.

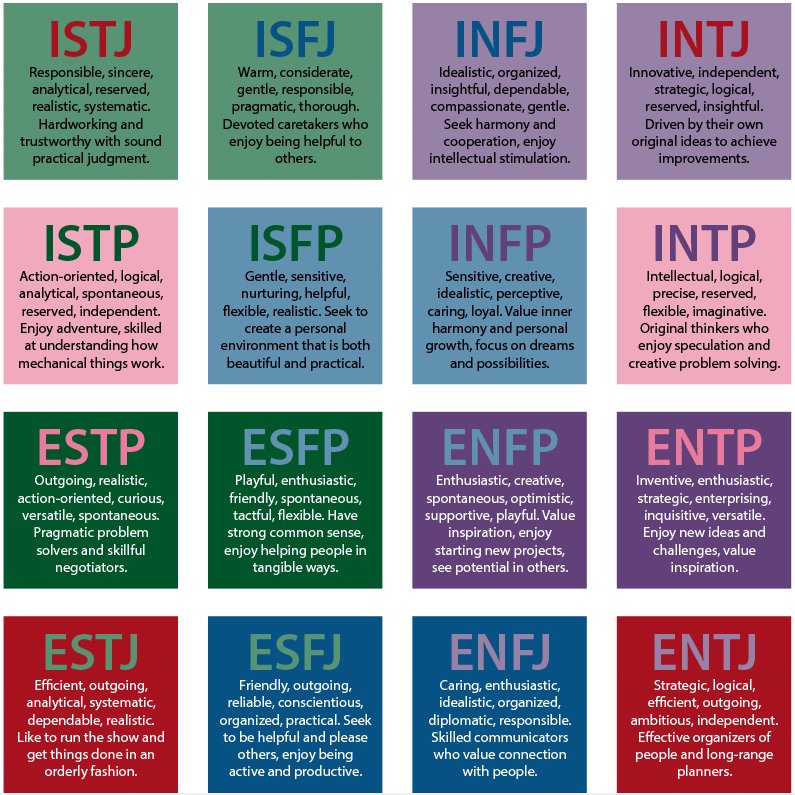
To make our classifications, we will divide 16 personality types into 4 major groups and apply KNN and Random Forests classification method on each column. Since we are given text data, we will also spend a lot of time on feature extraction. Then the model is trained and got an accuracy around 90% for each columns. For future usage, the model is saved in the drive or to the local machine.

**1. Introduction**

* 1. **MBTI:**

The characteristic set of actions, ideas, and emotional patterns that result from a person's biology and environment are referred to as their personality. Although there isn't a single definition of personality that is universally accepted, the majority of theories centre on motivation and psychological interactions with one's surroundings. Personality is described as "The features that predict a person's conduct" in trait-based personality theories like those put forth by Raymond Cattell.On the other hand, more behavioral approaches use habits and learning to define personality. However, most theories consider personality to be somewhat stable.

The MBTI model is another name for the Myers-Briggs Type Indicator Personality model [13]. An introspective self-report questionnaire called the Myers-Briggs Type Indicator (MBTI) identifies various psychological tendencies in how people view the world and make decisions. The MBTI's fundamental premise is that each of us has unique preferences for how we interpret our experiences, and that these preferences are what drive our interests, needs, values, and motivation.



Machine learning is a type of artificial intelligence that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning [algorithms](https://www.techtarget.com/whatis/definition/algorithm) use historical data as input to predict new output values.

Machine learning is made up of three elements:

* The computational set of rules in the middle of making determinations.
* Variables and features that make up the decision.
* Base knowledge for which the answer is understood that permits (trains)

There are mainly four technologies that machine learning used to work:

### 1. Supervised Learning

### 2. Unsupervised Learning

### 3. Semi-supervised learning

### 4. Reinforcement learning

**1.2 Text Classification:**

[Text classification](https://monkeylearn.com/text-classification/) also known as text tagging or text categorization is the process of categorizing text into organized groups. By using [Natural Language Processing](https://monkeylearn.com/natural-language-processing/) (NLP), text classifiers can automatically analyse text and then assign a set of pre-defined tags or categories based on its content.

Text Classification Example:

Text classification is becoming an increasingly important part of businesses as it allows to easily get insights from data and automate business processes. The example for text classification is as shown in the figure 3. Some of the most [common examples](https://monkeylearn.com/text-classification-examples) and use cases for automatic text classification include the following:

* [Sentiment Analysis:](https://monkeylearn.com/sentiment-analysis) the process of understanding if a given text is talking positively or negatively about a given subject (e.g. for brand monitoring purposes).
* [Topic Detection:](https://monkeylearn.com/topic-analysis/) the task of identifying the theme or topic of a piece of text (e.g. know if a product review is about Ease of Use, Customer Support, or Pricing when analyzing customer feedback).
* Language Detection: the procedure of detecting the language of a given text (e.g. know if an incoming support ticket is written in English or Spanish for automatically routing tickets to the appropriate team).

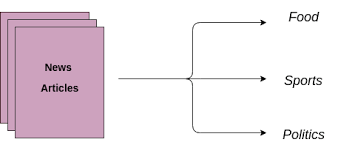


Fig. 1.2: Text Classification Example

**Supervised Learning**

In supervised learning, data is already tagged with correct results- this allows for machines to learn from well-labeled data. Whenever new information enters into the system, an algorithm can analyze this sample data and predict correct outputs using previously labeled examples.

It is classified into two different categories of algorithms. These are as follows:

* **Classification:** **The input data are classified by classification models. Techniques for classification foresee separate reactions. For instance, the tumour may be malignant or benign, while the email may be spam. Common uses include credit scoring, speech recognition, and medical imaging.**

Common algorithms for performing classification include support vector machines (SVMs), boosted and bagged decision trees, *k*-nearest neighbors, Naive Bayes, discriminant analysis, logistic regression, and neural networks.

* **Regression:** **Techniques for regression forecast continuous responses, such as variations in temperature or the demand for power. Power load forecasting and algorithmic trading are examples of common applications.**

Common regression algorithms include linear, nonlinear models, regularization, stepwise regression, boosted and bagged decision trees, neural networks, and adaptive neuro-fuzzy learning

**1.3 Problem Statement:**

The basic goal of the research is to compare Machine Learning algorithms utilising the MBTI dataset at a specific type of categorical data. For convenience of use, the personalities are divided into 4 separate columns, and the results are then compared in terms of accuracy.

**2. Related work**

In recent years, research on personality prediction has exploded. The multiple advantages of determining a user's personality would be a significant factor in this increase. Before I actually started working on creating our own system for personality analysis, I consulted a number of international studies and articles.

In paper [1], they presented a technique to predict personality types using 1.2 million tweets that had MBTI annotations. Binary word n-gram is used for feature selection. Although there were no changes in the Sensing-Intuition and Performance Reduced for Perceiver-Judgers dimensions of the Logistic Regression Model, linguistic features produced better results. In paper [10], an experiment was conducted using data from 150 users and my personality dataset, which was manually collected. To predict personality, the experiment built multilayer perceptron, GRU, and long short-term memory, combined CNN 1D and LSTM algorithms, and used linguistic features such as SPLICE (structured Programming for Linguistic Cue Extraction) and LIWC (Linguistic Inquiry and Word Count). The model performance showed the accuracy of 79.49%.

In this paper [2], the authors employed PCA (principal component analysis), correlation, and grey prediction models, as well as multiple regression and the multitasking model, to extract the data from social media and personality questionnaires for the dataset. This demonstrated that the grey prediction of MAE values had good performance and that the prediction accuracy ranged from 0.8 to 0.9.

The big five model was used by Tadesse et al. [3]to categorise a person's personality. Their research demonstrates that extraverted users write shorter, simpler sentences and use more positive and emotive terms than neurotic users. On the other side, neurotic users employ phrases that convey negative emotion and compose longer entries and sentences. On the other hand, users with high conscientiousness choose discussing upbeat topics. People who are highly open utilise function words more frequently than content terms.

Users who are in agreement frequently use more question marks and interrogative language.

**3.Proposed Methodology**

**3.1 Dataset:**

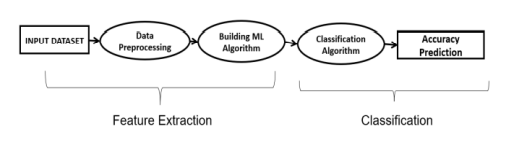
We used the MBTI dataset [11], which was gathered from the my Personality forum, to categorise participants according to their MBTI. Each row of the 8,600 rows of unique data specifies the type and posts, where type is the four-coded MBTI type, followed by the first 50 posts made by the user of the indicated type. There are three pipe characters ('| | |') separating each post. A large imbalance between introverts and extroverts was found when the number of users for each characteristic was analysed (6678 (I) VS 1999 (E)). A similar situation was found when the number of users for the sensing and intuition qualities was analysed (7478 (N) VS 1197 (S)). Judging/perceiving (5241 (J) VS 3434 (P) and feeling/thinking (4694 (F) VS 3981 (T)) pairs, in contrast, were balanced.



**Fig.1.1 Dataset**

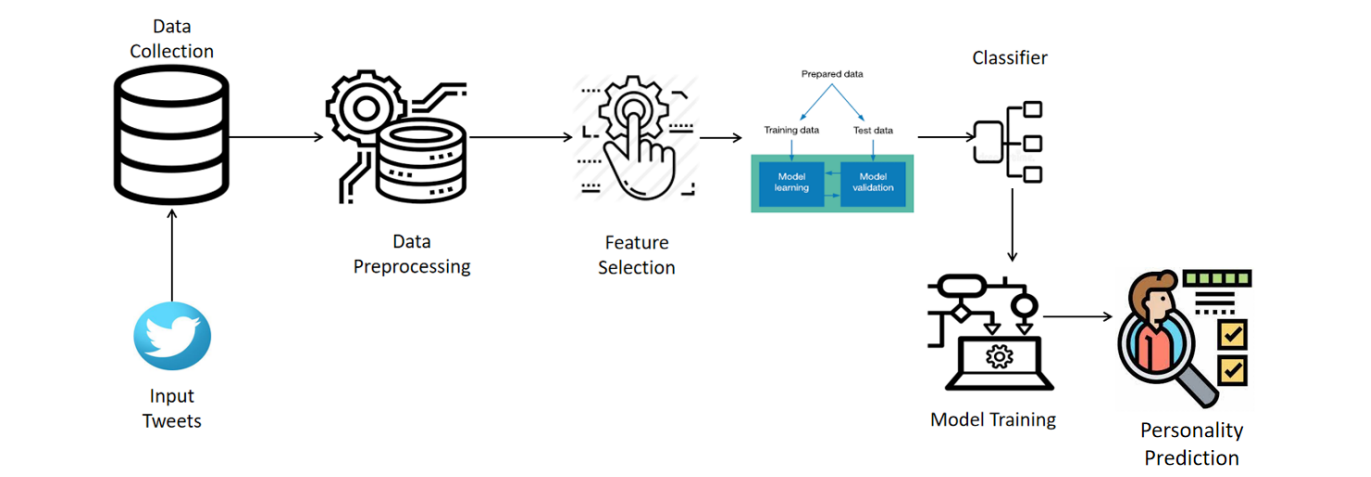
**3.2 Dataflow:**

The datasets that are divided into categories based on how respondents responded to the four "MBTI" dimensions now include four additional columns: "Intuition (N) - Sensation (S)", "Feeling (F) - Thinking (T)", and "Perception (P) - Judgment (J)" are all illustrations of introversion. The purpose of the process is to improve the accuracy of "KNN" and "Random Forest." The four columns on the four "Recurrent Nepal Networks" dimensions were constructed using "PD. obtain dummies ()," and those variables were used as one-hot encoding. values. deletion of particular words and characters Since the information in the text comes from "personalitycafe.com," a chat/forum where all communication is done through writing. We removed some data pieces with links to websites because our model is universal.



**Fig 3.2 DataFlow Diagram**

**3.3 Model Creation:**



**Fig 3.2 System architecture**

**3.3.1 Training Testing:**

The dataset was split into training and test sections testing to assess the MBTI personality model's correctness. We used the "traintest split" method from the "Scikit-learn" package to divide the data into 25% for testing and 75% for training. The testing dataset is a set of unknown data used solely to judge how well a certain classifier performs.

**3.3.2 Classification Process:**

A person's personality may fall into more than one personality class. The aim was to partition the 16 classes into four binary classes, and a model was developed for each type of indication. Random Forest and K-Nearest Neighbor were employed as the classifiers.

**Random Forest Prediction:**

Improvement in prediction with the predictive accuracy of the dataset and prediction with the algorithm classification in the random forest model. The number of subsets in the dataset that each have unique properties and can use the specified model's average. The supervised learning method in the provided model for the random forest in the machine learning algorithm for the data that can be easily expressed and implemented is the most widely used method.

The idea of likeness facilitates the process in a specific approach to learning techniques and can be reflected in the way through which the multiple categorization has the ability to execute and enhance a given machine learning model.

**K-Nearest Neighbor Prediction Model:**

The K nearest neighbour approach can be used to store data in a dataset that is already accessible. It classifies new data points according to the criteria needed for prediction, can be applied to content, and displays data in a way that is simple to categorise. Data is presented, followed by classification using the KNN algorithm with the goal of predicting accuracy and the most appropriate category.

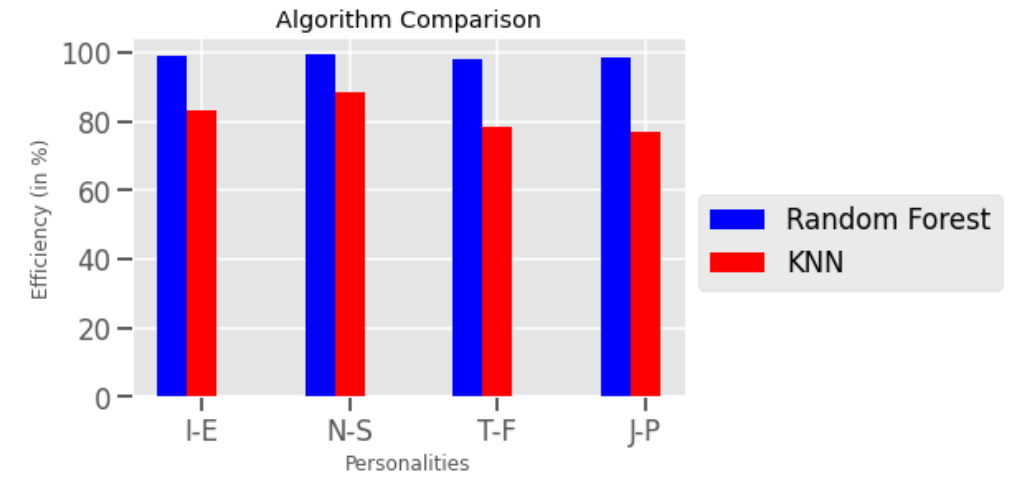
The KNN algorithm may predict personality using training data that has the ability to represent new ways of data prediction that can be stored at new data points, and it can then be categorised using categories with similar values. Removing the unwanted URLs is necessary because they are merely hyperlinks and don't hold any data.

**4.Result**

Specifically, Introversion (I), Extroversion (E), Intuition (N), Sensing (S), Thinking (T), Feeling (F), Judging (J), and Perceiving (P) are the 16 different personality traits used to predict personality. In this experiment, the machine learning method for categorization has been the main focus. Random Forest, a machine learning technique, exhibits higher accuracy practically across the board than K Nearest Neighbors Classifier.

|  |  |  |
| --- | --- | --- |
| Algorithm | Random Forest | K Nearest Neighbor |
| I/E | 99.19% | 83.26% |
| N/S | 99.38% | 88.46% |
| T/F | 98.19% | 78.22% |
| J/P | 98.3% | 76.98% |

**Table 1:Comparative Analysis of result**



**Fig 4:Graphical comparison of results**

**Project Link:** <https://colab.research.google.com/drive/1xSFx1onMy885wh_IqCxhURmsfspUQq90?usp=sharing>

**5.Conclusion**

The system was able to gather the user's tweets after receiving his Twitter handle. The used algorithms were effective in processing the tweets and generating the necessary results. The data is processed using the NLTK software, which provides an integrated list of stop words in English. The http links, symbols, and numbers are removed using regular expressions. The Random Forest Classifier was used in the suggested way to predict personality, and the outcomes were the best.

**BIBILOGRAPHY**

[1] S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, “A comparative Study of Different Classifiers for Myers - Brigg Personality Prediction Model, ” IRJET, vol.05, pp.1410 - 1413, 2018.

[2] Chaowei Li, Jiale Wan, Bo Wang, “Personality Prediction of Social Network Users”, IEEE, 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, pages 84 - 87.

[3] Michael Mesfin Tadesse,Hongfei Lin, and Bo Xu. Personality Predictions Based on User Behavior on the Facebook Social Media Platform. IEEE, 2018

[4] <https://www.javatpoint.com/machine-learning-techniques>

[5] <https://towardsdatascience.com/>