Indian Journal of Data Mining (IJDM)
ISSN: 2582-9246 (Online), Volume-3 Issue-1, May 2023

Myers-Briggs Personality Prediction

Rohith Muralidharan, Neenu Kuriakose, Sangeetha J

Abstract: ***The Myers-Briggs Type Indicator (MBTI) is one of the most commonly used tool for assessing an individual's personality. This tool allows us to identify the psychological proclivity in the way they take decisions and perceive the world. MBTI has it's applications spread across several fields which include career development and personal growth. This test consists of a set of questions which are specifically designed to evaluate and measure an individual's choices based on four dichotomies - Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Myers-Briggs Personality Prediction project aims to develop and deploy a system using machine learning which is capable of predicting one's MBTI personality type based on their online written interactions such as social media posts, comments, blogs etc. This project has significant implications for various applications, including improving customer experience, optimizing team dynamics, and developing personalized coaching programs. Through this project, we hope to gain a deeper understanding of how language use and personality type are related and to develop a robust tool for personality prediction.***

Keywords: Python, Machine Learning, XGBoost, Supervised learning.

I. INTRODUCTION

A psychological evaluation instrument called the Myers-Briggs Type Indicator (MBTI) aids people in determining their psychological preferences for how they see the outside world and make judgements. The MBTI was created by Katharine Cook Briggs and Isabel Briggs Myers, her daughter, and is based on Carl Jung's idea of psychological types. The MBTI is frequently employed in many different contexts, such as job development, team building, and personal development. An individual's preferences on four dichotomies are measured by a series of questions in the assessment: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P) [12].

The difference between an individual's preference for exterior world that includes other people, events, and objects (Extraversion) or the interior world consisting of thoughts, feelings, and reflections (Introversion) is measured by their extraversion (E) or introversion (I). A person's choice for either definite information gained through the five senses (Sensing) or conceptual knowledge and patterns (Intuition) is measured by the Sensing vs. Intuition scale (S vs. N). A person's preference for choosing options based on logic and unbiased criteria (Thinking) or on personal values as well as emotions (Feeling) is measured by the Thinking (T) vs. Feeling (F) scale. The Judging vs Perceiving scale compares a person's preference for a rigid and planned lifestyle (Judging) to one that is adaptable and unplanned (Perceiving).

Each dichotomy in the MBTI evaluation produces a score between the two opposing poles, and based on the sum of these values, people are assigned to one of 16 personality types. Based on four-letter codes like ESTJ, INFP, or ENTP, the 16 personality types are divided into groups. The validity, dependability, and simplifying of personality in the MBTI evaluation have all drawn criticism. The MBTI has been criticized by some academics for lacking empirical backing and for failing to effectively express the complexity of one's personality through its dichotomies. The MBTI is still a popular evaluation instrument despite its detractors, and its acceptance can be due to its usability and accessibility. The use of the MBTI spans a number of disciplines, including career counselling, organizational development, and personal development. To assist people in understanding their own and other people's communication preferences and styles, it is frequently used in team-building exercises. The application of the MBTI evaluation to machine learning and natural language processing has also been studied by researchers in recent years. Researchers can accurately estimate a person's MBTI personality type by looking at their written communications, such as emails or social media posts.

Using the MBTI with NLP and ML has important ramifications for several applications, such as enhancing team dynamics, creating individualized coaching programs, and bettering customer experience. Organizations may adjust their coaching and communication to better suit the requirements and preferences of each individual by identifying their personality type.

The goal of this Myers-Briggs personality prediction project is to create a machine learning model that can identify a person's MBTI personality type based on their textual communication. We want to construct a powerful tool for personality prediction and to acquire a deeper understanding of the relationship between language use and personality type by studying a huge corpus of social media postings and emails. We hope to advance the field of personality assessment and prediction with this project and show the MBTI assessment's potential in machine learning and natural language processing. The MBTI has been extensively utilized in several areas, including team building, career development, and personal development. It has been used to help people understand their interaction and learning styles, to help people recognize their strengths and shortcomings, and to increase collaboration and teamwork.

Manuscript received on 28 April 2023 | Revised Manuscript received on 08 May 2023 | Manuscript Accepted on 15 May 2023 | Manuscript published on 30 December 2023.

**Correspondence Author(s)*
Rohith Muralidharan*, Department of Computer Science, St. Albert's College (Autonomous), Ernakulam, India. Email: rohith519@gmail.com, ORCID ID: 0009-0007-9967-9293
Neenu Kuriakose, Department of Computer Science, St. Albert's College (Autonomous), Ernakulam, India. Email: neenuannaa@gmail.com, ORCID ID: 0000-0002-9942-2160
Sangeetha J, Department of Computer Science, St. Albert's College (Autonomous), Ernakulam, India. Email: sangiprathap@gmail.com

To administer the MBTI exam, however, one must complete a questionnaire, which can be time-consuming and prone to answer biases. In this era of advanced technology and social media, we can prominently see more and more people expressing their opinions and ideas on social media or blogs rather than in person. People are so comfortable sharing their opinions and expressing themselves online that they hesitate to interact in a similar way when in person. Thus, understanding their personality is a challenge when trying to assess what their thinking style is. This is important in order to understand the personality type of people, this data can further be used for developing marketing strategies by ad agencies, for understanding customer feedbacks, to understand what type of thinking prevails amongst a group of people regarding a particular product or topic. Myers-Briggs Personality Prediction allows us to use machine leaning models to evaluate the type of personality of a person based on the social media post, comment or any type of textual data written by individuals.

The primary objective of this study is to find a machine learning model that provides us with higher accuracy than the existing machine learning models - Logistic regression, Naïve Bayes, and Random forests. We implement XGBoost in order to accomplish the primary objective. We also try to structure the XGBoost Model in such a way that it provides us maximum accuracy possible. The project undergoes series of steps which includes Data collection, Data Pre-processing, Exploratory Data Analysis, Predictive Modeling and Model Evaluation. The Project goes through these stages to effectively give us a best model.

II. RELATED WORKS

This study by Amirhosseini et al. (2020) presented a unique machine learning strategy for automatic meta programming recognition and personality type prediction using the MBTI personality type indicator. The Natural Language Processing Toolkit (NLTK) and XGBoost, an optimised distributed gradient boost library developed in Python for the execution of machine learning algorithms within the Gradient Boosting framework, were used in the development process. The XGBoost model's performance and accuracy were assessed using the same dataset as the most recent and effective current approach. The findings demonstrate that, in comparison to other approaches already in use, the methodology described in this study offers higher accuracy and reliability. Regarding the paper's knowledge addition, the methodology's presentation substantially increased the precision of identifying the four sets of MBTI personalities [1].

Basto et al. (2021) proposed this paper which uses data-centric approach with Natural Language Processing to predict personality types based on MBTI [2]. It also implemented hyperparameter tuning using grid search and gradual feedback. Bhagat et al. (2023) discusses an approach for predicting personality using deep learning on twitter data. It uses deep learning technology to analyse tweets and understand and predict the personality of the twitter user [3][14][15][16][17][18].

Choong EJ et al. (2021) set out to see how well individual features and classifiers worked for the J/P dichotomy in character computation. After comparing five machine learning techniques, the LightGBM model was proposed for the task of predicting J/P dichotomy in MBTI character computation. There is minimal difference in the MBTI Judging-Perceiving prediction ability on social media among character-level TF, TF-IDF and word-level TF, TF-IDF, according to this study. LIWC (Linguistic Inquiry and Word Count) prediction accuracy is frequently off. Character-level features are favored over word-level features since they have lower predictive power and interpretability. Five classifiers were compared to attempt to predict the MBTI Judging-Perceiving type. LightGBM and SVM have identical prediction performance, while being clearly superior to each of the three classifiers. Because SVM failed to attain convergence on just one of the datasets, this study eventually recommended LightGBM for its greater resilience [4].

M. Maulidah et al. (2021) proposes an interesting approach to classify an individuaol's MBTI personality using Long Short-Term Memory (LSTM) algorithm with random oversampling technique. The LSTM was clubbed with RMSprop optimizer to obtain the results [5].

N. Cerkez et al. (2021) proposes a method for classifying personality by an experimental attempt to classify using long short-term memory (LSTM) and convolutional neural network (CNN). The paper discusses the use of classification methods such as ZeroR, Gaussian Process, neighbour classifier and so on [6].

In this study, Ryan G et al. (2023) used the Python computer language to make MBTI personality type predictions based on text. The research's suggested approach made use of Word2Vec embedding, SMOTE, and six machine learning classifiers: Linear Support Vector Classification, Logistic Regression, Random Forest, Stochastic Gradient Descent, CatBoost and Extreme Gradient Boosting. The ability of each classifier to predict an individual's MBTI personality type was developed and evaluated separately. The results showed that the best machine learning model for predicting MBTI type features in this investigation was logistic regression (LR), with an average F1 score of 0.8282. The F1 score increased to 0.8337 as a result of the applied SMOTE technique, while dimension 3 (N/S) had the highest score, 0.8821. The acceptable threshold for the F1 score varies based on the application, however an F1 score that is close to 1 is often regarded as high for data categorization. The fact that this result was superior to the results of the other models considered shows that the recommended method may be used to deepen our understanding of MBTI and may be used in a range of situations where personality classification is required [7].

The Myers-Briggs Type Indicator (MBTI) is used in this study by Sakdipat Ontoum et al. (2022) to predict people's personalities from text using a variety of machine learning approaches, such as "Naive Bayes," "Support Vector Machines," and "Recurrent Neural Networks." Additionally, "CRISP-DM"—an acronym for "Cross-Industry Standard Procedure for Data Mining"—is implemented to implement direct the learning process in this project. Since "CRISP-DM" is an iterative development technique, agile methodology—a quick iterative software development technique—was applied in order to shorten the development cycle. Using information from social networks and a machine learning system, this study succeeded in to predict personality.

Furthermore, the best model for estimating personality based on "MBTI" is the "Recurrent Neutral Networks" machine learning method [8]. Stein et al. (2019) discusses the validity of MBTI theory and also talks about the shortcomings in this theory [9]. Vignesh Ramachandran et al. (2020) provides a comprehensive outlook on how MBTI can be used in medical education. MBTI being a powerful tool has proved to be efficient and promising in training the individuals by understanding and improving their personality trait. It describes how it can be used in non-cognitive aspects of medical training [10]. In a paper proposed by Z. Mushtaqet al. (2020), a way to predict user's personality by analysing social media post is discussed. They used K-Means Clustering and Gradient Boosting to identify the MBTI personality of the users based on their social media activity [11].

III. METHODOLOGY

A. Data Collection

Data collection is the first and foremost stage of any machine learning project. The type of dataset we choose decides how accurate or favorable results we obtain. Grabbing datasets form trusted sources helps us to ensure the integrity, genuine nature and reliability of data. It is important that the dataset is chosen wisely and enough data is present in the dataset for testing and training purposes. The dataset used in this project is the MBTI Dataset which was obtained from Kaggle. It consisted of over 8675 rows and two columns namely - type and posts [13]. The dataset consisted of textual or comments from social media, feedbacks, comments and so on and a predefined personality type assigned to that comment. Data collection is a very important step as the existence of entire project is based on this. The data can be collected by various means such as questionnaire, feedback forms, interviews and so on. Ethical issues are also raised by the collection of data for the Myers-Briggs Personality Prediction study. The idea depends on gathering personal information from people, such as emails or social media posts, which creates privacy issues. In order to resolve these issues, researchers must gain participants' informed consent and make sure that the data gathering procedure complies with all applicable ethical standards and laws. In this particular project we have not mentioned name or any personal details of the writers of those comments or feedback, this helps to maintain and protect the individual's privacy. Several tactics can be used to get around the difficulties in gathering data for the Myers-Briggs Personality Prediction project. Utilizing crowdsourcing platforms is one such tactic that enables academics to swiftly and effectively collect enormous amounts of data. Platforms for crowdsourcing can also offer a variety of participants, ensuring a representative sample. Utilizing current datasets, such as publicly accessible social media posts or email exchanges, is another tactic. These datasets may be utilized to build machine learning models for predicting personality types after being preprocessed to remove any personally identifying information. A crucial component of the Myers-Briggs Personality Prediction project is data collection. The dataset must effectively balance the four MBTI dichotomies and be true to the target community.

B. Pre-processing

The next step after data collection is the preprocessing. It involves cleaning and preparing dataset for the machine learning algorithms to train. This strep ensures data quality and allows us to improve the accuracy of the model and reduce risk of overfitting. The first part of data preprocessing is data cleaning which involves removing unwanted or irrelevant, noisy data that may affect model accuracy. The noise should be removed to ensure that the data is clean. We also need to make sure that there are no duplicate data entries in the dataset. If found out, we need to remove these duplicates as well. In this project, we do have some irrelevant data but there are no duplicates found. The irrelevant data is present in the post's column. To clean this, we need to performs several preprocessing steps such as converting them to lower case, removing repeated words or letters, removing unwanted spaces, removing hyperlinks, removing words that resemble the assigned personality type, removing unwanted full stops and punctuation marks and so on. These help in the model building procedure.

Data normalization, the following stage, entails translating the data into a standardized format. This process is crucial for data comparability and lowers the chance of bias brought on by variations in data formatting. Data normalization in the context of the Myers-Briggs Personality Prediction research could entail converting communication samples to a specific language, format, or length. Following the normalization of the data, feature extraction is carried out to locate pertinent elements that can be employed to gauge an individual's MBTI personality type. This can entail extracting linguistic data from the Myers-Briggs Personality Prediction project, such as phrase length, pronoun usage, or sentiment analysis. Feature engineering, which entails generating new features from the existing data to enhance model performance, is a crucial stage in the preprocessing process. This phase is crucial in situations where the extracted features might not be enough to accurately forecast the target variable. Following the discovery of the features, feature selection entails choosing the features that will be used to train the model. This procedure increases the interpretability of the model and lowers the risk of overfitting. Label Encoder is a technology used in this project that aids with feature selection.

C. Label Encoder

It is a technique which allows us to convert categorical data into numerical data. LabelEncoder assigns a unique label to each category in the categorical variable column, this makes it easy for the machine learning algorithm to process the data. In order to encode categorical data into a format that machine learning models can use, the preprocessing technique LabelEncoder is frequently used alongside with other preprocessing approaches, such as OneHotEncoder.

D. Natural Language Processing

NLP which stands for Natural Language Processing plays an important role in this project. Since we posts are texts from comments o social media posts, these can be efficiently analyzed by NLP. NLP has a potential to deal with textual data and has the capability to interpret human language and associated responses. NLP consists of many techniques including text mining, sentiment analysis, machine translation, and speech recognition.

NLP allows us to interpret the textual data along with its application in Machine learning. It allows us to understand human language and gather insights from it. Some NLP techniques use in this project are:

E. Count Vectorizer

A common method in natural language processing (NLP) for turning a collection of text documents into a token count matrix is called "CountVectorizer." It is a quick and effective way to extract features from text data. Each word in a document is counted for frequency by CountVectorizer, which then visualizes the document as a vector of word counts. This technique can be used to find the most frequently occurring terms in a corpus and to spot linguistic variances between various social groupings. Before using machine learning models on text input, CountVectorizer is frequently used as a pre-processing step. It is a typical NLP feature extraction method that is applied to a variety of tasks, such as sentiment analysis, topic modelling, and text categorization.

F. TF-IDF

Term Frequency-Inverse Document Frequency is a typical text analysis method in natural language processing (NLP). It is a statistical tool which assesses a word's significance to a group or corpus of documents. It operates by counting the frequency of each term (word) appears in the document and multiplying that number by the term's IDF (inverse document frequency) throughout the corpus. The IDF term assists in locating terms that are particular to a given document and are thus more useful for analysis. Applications including text classification, document retrieval, and clustering frequently use TF-IDF. It is an effective method for studying vast amounts of textual data and can produce useful information.

G. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial phase in this project since it enables us to understand the data, spot trends, and choose the best preparation procedures. The EDA process uses a range of methodologies, including as feature engineering, statistical analysis, and data visualization. A useful method for studying the dataset and spotting trends that may not be obvious from summaries of numbers alone is data visualization. For instance, scatter plots can be used to show how two variables relate to one another, while histograms and box plots can show how many variables are distributed. Data visualization can be used to find trends in language use across various personality types, such as variances in vocabulary, sentence length, or grammatical structures.

Additionally, statistical analysis might offer insightful analyses of the dataset. As an illustration, correlation analysis can be used to find variables that are closely associated to the target variable, and hypothesis testing can be used to discover whether language use differs significantly between personality types. The outcomes of these analyses can help when choosing the characteristics to include in the machine learning model. Along with these methods, there are a few factors particular to this project which should be considered during EDA. In order to prevent the algorithm used for machine learning from favoring any one personality type, it is important to balance the data set between the four MBTI dichotomies.

The dataset was thoroughly analyzed to understand the distribution of post as shown in Figure 1 based on each of the 16 personality types which are - ENFJ', 'ENFP', 'ENTJ', 'ENTP', 'ESFJ', 'ESFP', 'ESTJ', 'ESTP', 'INFJ', 'INFP', 'INTJ', 'INTP', 'ISFJ', 'ISFP', 'ISTJ', 'ISTP' which were developed from the initial 4 sets of personality traits- Introversion (I) vs Extroversion (E), Intuition (N) vs Sensing (S), Feeling (F) vs Thinking (T) and Judging (J) vs Perceiving (P). We also plot a swarm plot as shown in Figure 2 to understand the words per comment of each personality trait. We also have a joint plot as shown in Figure 3 to showcase the variance of word counts over words per comment as well as for each personality trait. We also use the dataset to effectively find out the distribution over density of top 50 posts as shown in Figure 4. We also find out the most common words among all the posts and plot a word cloud based on it which is shown in Figure 5 [12]. We then plot a word cloud for most common words for posts of each personality trait that is been demonstrated in Figure 6.

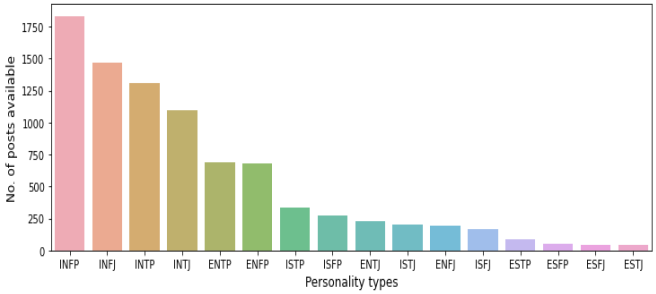

Fig. 1. Distribution of Posts for Each of the Distinct 16 Personalities

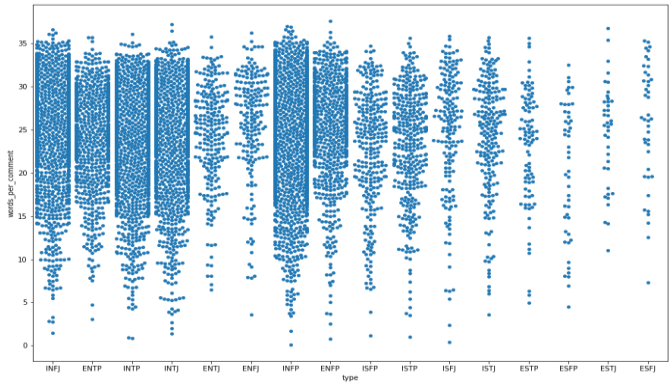

Fig. 2. Swarm Plot of the Words Per Comment of Each Personality Trait

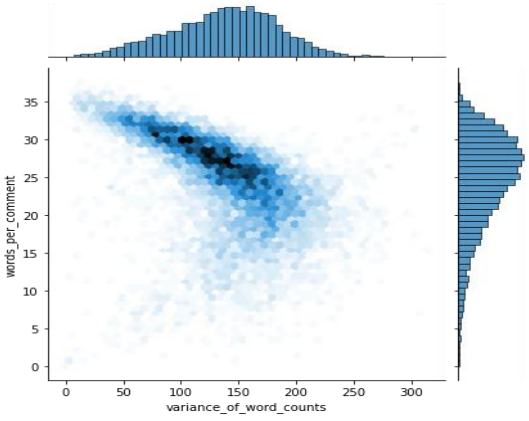

Fig. 3. Joint Plot of the Variance of Word Counts Over Words Per Comment

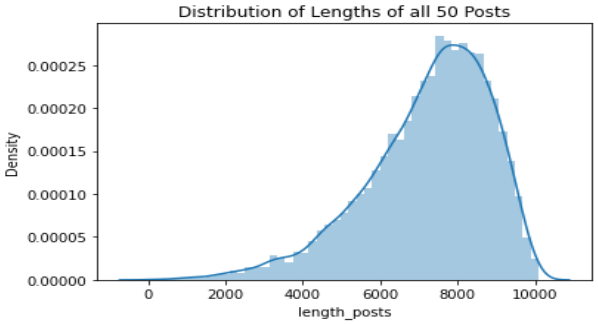

Fig. 4. Distribution of Lengths of Top 50 Posts

Fig. 5. Word Cloud of Most Common Words

Fig. 6. Word Cloud for Most Common Words of Each of the 16 Distinct Personality Types

H. Predictive Modelling

Predictive Modelling is the process of building machine learning models based on various machine learning algorithms. The model is used to predict labels for any data provide in the future. Predictive modelling identifies crucial patterns within the dataset which is used by the algorithm to predict or forecast the results. The accuracy of predictions entirely depends upon the discovery of important and useful patterns. In this project, 4 algorithms were used namely - Logistic Regression, Naive Bayes, Random Forest and XGBoost. Logistic Regression, Naive Bayes, Random Forest were machine learning algorithms used in the existing system, while XGBoost algorithm has been included in this to obtain a better result than previously used algorithms and compare them based on performance.

Any dataset before feeding it to the machine learning model follows a conventional technique called test train split, which is splitting the dataset into portions out of which one portion is used for the purpose of training the machine learning model so that it can be ready for prediction of future instances while the other is used check or test how well the model has been trained by evaluating its performance by providing the data to the model for prediction. The test train split used in this project is 70 -30, which means 70% data was used for training and the remaining 30% data was used for testing.

I. Logistic Regression

A binary variable that is dependent and a number of independent variables are analyzed using the statistical modelling technique known as logistic regression. Since the dependent variable can only have one of the two possible results, such as a yes/no or 0/1 response, it is binary. Logistic regression is frequently used in a number of disciplines, including the social sciences, medical research, and business. The logistic regression model determines the likelihood that the variable that is dependent will have a specific value based on the numerical results of the independent variables. The logistic regression model produces the log-odds, or logit, for the dependent variable.

The interpretability of logistic regression is one of its main benefits. The logistic regression model's independent variable coefficients show how the dependent variable's log-odds vary when the independent variable changes by one unit. As a result, it is feasible to determine which independent factors significantly affect the dependent variable as well as the strength and direction of their influence.

J. Random Forest

It is a well-liked machine learning strategy for problems like classification and regression. It is a particular form of ensemble learning technique that combines different decision trees to provide predictions with greater accuracy. The flexibility of random forests allows them to handle categorical, textual, and numerical data among other forms of data. The random forest method creates several decision trees using various randomly selected subsets of the training data. Each decision tree splits a random subset of features at each node after it has been trained on a subset of the features. By combining all the decision trees' predictions, either through majority voting (for classification) or averaging (for regression), the final prediction is determined.

Random forest is capable of handling high-dimensional data containing a lot of features, which is one of its key advantages. In order to lessen overfitting and increase generalization, the method chooses a random selection of characteristics for each tree. The measure of feature relevance that is provided by this feature selection procedure can be used to pinpoint the features that are most pertinent to the prediction task. As a result of its strong robustness to missing or noisy data, random forest is excellent for real-world applications where ideal data quality may not always be possible. The method may deal with missing values by imputing them based on the existing information.

K. Naïve Bayes

The Bayes theorem, a statistical theory that calculates the likelihood of an event based on knowledge of the circumstances that might be linked with it, is the foundation of the Naive Bayes machine learning technique, which is used for classification tasks. Naive Bayes is frequently used in natural language processing (NLP) and text classification applications such as spam filtering, sentiment analysis, and language recognition.

The assumption of feature independence is what the "Naive" in Naive Bayes refers to. According to this presumption, the presence or absence of one feature does not depend on the presence or absence of any other characteristic, which indicates that the algorithm views each feature as having a separate effect on the result.

In order to determine the class of a new instance, the method first determines the probability of every attribute given a specific class. The conditional probability of every feature given a class is calculated, and the conditional probability of that class given the characteristics is then obtained by multiplying the conditional probabilities of each feature by the class. The instance is then given the category with the highest probability. Both categorical and continuous data can be handled by naive bayes, but the data must first be preprocessed and formatted appropriately. While the technique assumes a normal distribution and computes the mean and standard deviation of each feature for continuous data, it uses a frequency table to determine probabilities for categorical data.

L. XGBoost

Extreme Gradient Boosting, sometimes known as XGBoost, is a potent open-source machine learning package that has become very well-known recently. It is intended to increase the efficiency of gradient boosting methods by making the models faster and more accurate. Many machine learning tasks, including regression, classification, and ranking issues, make extensive use of XGBoost. The fact that XGBoost can work with big datasets and high-dimensional feature spaces is one of its key features. Through the removal of pointless branches and the use of parallel processing strategies, the approach is intended to maximize the decision tree ensemble. Due to its ability to manage huge and complicated datasets, XGBoost is a great option for many practical machine learning applications.

In order to increase the accuracy of the XGBoost, Hyperparameter tuning is used. Hyperparameter tuning is a critical step in machine learning, where the optimal hyperparameters of the model need to be found to achieve the best performance. XGBoost provides several hyperparameters that can be tuned to improve the model's performance. Grid search and random search are commonly used hyperparameter tuning techniques for XGBoost. Grid search exhaustively searches over a pre-defined range of hyperparameters, while random search selects hyperparameters at random from a distribution. Both techniques can be time-consuming and computationally expensive, but the benefits of hyperparameter tuning can result in significant improvements in model performance. XGBoost is a powerful machine learning algorithm that offers several benefits, including its ability to handle large datasets, regularization techniques, feature importance metrics, and early stopping criteria. Its popularity has made it a popular choice for machine learning tasks that require high accuracy and speed.

M. Model Evaluation

An important element of any type of machine learning project is model evaluation. A personality prediction model's performance on the test data set determines how accurate it is. The performance model can be assessed using a variety of evaluation measures. The most popular evaluation metrics are F1 score, Accuracy, Precision, and Recall. Evaluation of the personality prediction model's robustness and generalizability is also crucial. The term "robustness" describes a model's capacity to function well on data sets with various features. The capacity of the model to perform effectively on sets of data that it has not been trained on is referred to as generalization ability. In this project the comparison of the accuracy metrics for the entire model as well as for predicting the 4 sets of personality traits - Introversion (I) vs Extroversion (E), Intuition (N) vs Sensing (S), Feeling (F) vs Thinking (T) and Judging (J) vs Perceiving (P). The accuracy for each model has been evaluated separately so that comparing the accuracy measures is easy.

N. Confusion Matrix

An effective method for evaluating how well a model developed using machine learning is performing is a confusion matrix. The number of true positives, false positives, true negatives, and false negatives for each table class is shown in a classification issue. True positives are instances that were accurately classified as positive, while false positives are instances that were incorrectly classified as positive. Situations that should not have been classified as negative but were yet regarded as true negatives are known as false negatives. The confusion matrix can be used to calculate a range of assessment measures, such as accuracy, precision, recall, and F1 score, which provide insight into how well the model works. Data scientists can assess the confusion matrix to determine the benefits and drawbacks of the model and make modifications to improve the model's accuracy and dependability. The confusion matrix is an essential tool for determining if classification models are effective and meet the necessary requirements.

IV. RESULT

1. Accuracy Comparison for Each Model

Model Name	Accuracy (%)
Logistic Regression	58.23
Random Forest	34.49
Naïve Bayes	22.83
XGBoost	57.56
XGBoost after Hyperparameter tuning	64.80

Table 1. Accuracy Comparison of Each Machine Learning Algorithm

The Table 1 shows an exclusive comparison of the algorithms used in this project. Initially Logistic Regression, Random Forest and Naïve Bayes were used which gave an accuracy of 58.23%, 34.49% and 22.83% respectively. The proposed system makes use of XGBoost algorithm, which initially gave an accuracy of 57.56%. The model was improved by using Hyperparameter tuning after which the accuracy increased to 64.80%.

2. Personality Wise Accuracy for Logistic Regression

Personality Set	Accuracy (%)
I / E	77.22
N / S	86.25
F / T	73.30
J / P	64.35

Table 2. Personality Wise Accuracy for Logistic Regression

The Table 2. Shows the accuracy of Logistic regression over the 4 dichotomies used in this project, which is basically the 4 sets of personality traits. The accuracy percentage of each dichotomy is specified in the table.

3. Personality Wise Accuracy for Random Forest

Table 3. Personality Wise Accuracy for Random Forest

Personality Set	Accuracy (%)
I / E	77.30
N / S	86.21
F / T	69.27
J / P	63.00

The Table 3. Shows the accuracy of Random Forest over the 4 dichotomies used in this project, which is basically the 4 sets of personality traits. The accuracy percentage of each dichotomy is specified in the table.

4. Personality Wise Accuracy for Naïve Bayes

Table 4. Personality Wise Accuracy for Naïve Bayes

Personality Set	Accuracy (%)
I / E	77.49
N / S	86.21
F / T	68.04
J / P	62.16

The Table 4. Shows the accuracy of Naïve Bayes over the 4 dichotomies used in this project, which is basically the 4 sets of personality traits. The accuracy percentage of each dichotomy is specified in the table.

5. Personality Wise Accuracy for XGBoost

Table 5. Personality Wise Accuracy for XGBoost

Personality Set	Accuracy (%)
I / E	75.53
N / S	85.94
F / T	67.15
J / P	62.24

The Table 5. Shows the accuracy of XGBoost over the 4 dichotomies used in this project, which is basically the 4 sets of personality traits. The accuracy percentage of each dichotomy is specified in the table

6. Personality Wise Accuracy for XGBoost after Hyper-Parameter Tuning

Table 6. Personality Wise Accuracy for XGBoost after Hyperparameter tuning

Personality Set	Accuracy (%)
I / E	77.22
N / S	86.29
F / T	70.92
J / P	64.81

The Table 6. Shows the accuracy of XGBoost after Hyperparameter tuning over the 4 dichotomies used in this project, which is basically the 4 sets of personality traits. The accuracy percentage of each dichotomy is specified in the table. The short forms in the given tables are expanded as:

- Introversion / Extroversion – I/E
- Intuition / Sensing – N/S
- Feeling / Thinking – F/T
- Judging / Perceiving – J/P

7. Confusion Matrix

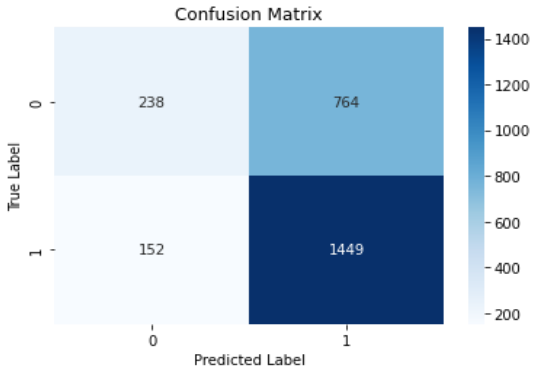

Fig 7. Confusion Matrix of XGBoost after Hyperparameter tuning.

The above Figure 7. Shows the confusion matrix which uses the most accurate model, i.e., XGBoost after Hyperparameter tuning. Confusion matrix was constructed using the testing set which consisted of 30% of the total data set. The top left number 238 represents the True Positive cases, top right number 764 represents False positive cases, the bottom left 152 represents False negative cases and the bottom right 1449 represents True Negative cases.

V. FUTURE ENHANCEMENT

Future improvements to the Myers-Briggs Personality Prediction project have the potential to increase both its precision and usefulness. Some of these improvements that might be used in the future include:

- Multimodal Input: At the moment, the project's analysis of textual communication samples is what it concentrates on in order to identify a person's MBTI personality type. The accuracy of the model can be increased by including additional modalities, such as audio, video, or pictures.
- Fine-grained Personality Prediction: The current project forecasts the four main categories—or dichotomies—of the MBTI personality types. Future improvements might instead concentrate on predicting the more nuanced characteristics of personality, like the facets inside each dichotomy.
- Longitudinal Analysis: Currently, the study only examines communication samples from just one point in time for longitudinal analysis. However, examining language usage variations through time can offer insightful information about how personalities grow and develop.
- Cross-Cultural Analysis: In order to comprehend how language usage and personality types differ between cultures, the study could be expanded to incorporate a cross-cultural analysis.
- Privacy Issues: The current effort depends on gathering personal information from people, which raises moral dilemmas. Future improvements might concentrate on creating methods that protect data privacy while still producing accurate predictions, like federated learning or homomorphic encryption.

VI. CONCLUSION

At the end of the project, it is concluded that the Myers-Briggs Personality Prediction project has a very significant capability to be used in various sectors including customer experience improvement, optimize team dynamics, personalized coaching programs and so on. The primary objective of the object to find a model using XGBoost that could provide us with a better and accurate results was accomplished. The accuracy of the algorithms used in the existing system was surpassed with help of a model created using XGBoost along with hyperparameter tuning. The aim of the project is to extract the personality type of an individual based on his/her text-based comments. blocks or feedbacks. This was successfully accomplished with improved performance and accuracy using the XGBoost method. Data collection, preprocessing, feature extraction, model training, and evaluation are only a few of the project's difficulties. To make sure that the project is technically feasible and to generate reliable predictions, these issues need to be carefully considered. Despite these difficulties, the project has several standout advantages. It makes use of the well-known MBTI assessment technique to offer insightful data about a person's psychological preferences and language use. Additionally, it uses scalable, highly accurate machine learning methods like XGBoost that produce helpful feature importance rankings. To increase its accuracy and scalability going ahead, the Myers-Briggs Personality Prediction project can profit from ongoing study and development. To ensure the project's societal acceptability and utility, ethical issues related data protection and the proper use of personality tests must also be considered. With important ramifications for numerous industries, the Myers-Briggs Personality Prediction project is an innovative application of machine learning and personality psychology. By working on this project, we can better understand the connections between the use of language and personality type and create a powerful tool for personality prediction.

DECLARATION STATEMENT

Funding	No, I did not receive.
Conflicts of Interest	The article bears No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	The dataset is freely avalaible in kaggle. The link for the same is - https://www.kaggle.com/datasets/datasnaek/mbti-type
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. Amirhosseini MH, Kazemian H. Machine Learning Approach to Personality Type Prediction Based on the Myers–Briggs Type Indicator®. Multimodal Technologies and Interaction. 2020; 4(1):9. https://doi.org/10.3390/mti4010009
2. Basto, Carlos. 'Extending the Abstraction of Personality Types Based on MBTI with Machine Learning and Natural Language Processing'. ArXiv [Cs.CL], 2021. arXiv. http://arxiv.org/abs/2105.11798.
3. Bhagat, Ayushi and Aylani, Amit and Shahasane, Aditi and Shinde, Shraddha and Limaye, Heramba, Personality Prediction By Using Deep Learning On Twitter (April 9, 2023). Available at SSRN: https://ssrn.com/abstract=4413612 or http://dx.doi.org/10.2139/ssrn.4413612
4. Choong EJ, Varathan KD. 2021. Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in online social forum. PeerJ 9:e11382 https://doi.org/ 10.7717/peerj.11382 https://doi.org/10.7717/peerj.11382
5. M. Maulidah, and H. F. Pardede, "Prediction Of Myers-Briggs Type Indicator Personality Using Long Short-Term Memory," Jurnal Elektronika dan Telekomunikasi, vol. 21, no. 2, pp. 104-111, Dec. 2021. doi: 10.14203/jet.v21.104-111 https://doi.org/10.14203/jet.v21.104-111
6. N. Cerkez, B. Vrdoljak and S. Skansi, "A Method for MBTI Classification Based on Impact of Class Components," in IEEE Access, vol. 9, pp. 146550-146567, 2021, doi: 10.1109/ACCESS.2021.3121137. https://doi.org/10.1109/ACCESS.2021.3121137
7. Ryan G, Katarina P, Suhartono D. MBTI Personality Prediction Using Machine Learning and SMOTE for Balancing Data Based on Statement Sentences. Information. 2023; 14(4):217. https://doi.org/10.3390/info14040217 https://doi.org/10.3390/info14040217
8. Sakdipat Ontoum, Jonathan H. Chan. Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning. 2022; https://doi.org/10.48550/arXiv.2201.08717
9. Stein, R, Swan, AB. Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. Soc Personal Psychol Compass. 2019; 13:e12434. https://doi.org/10.1111/spc3.12434
10. Vignesh Ramachandran, Asad Loya, Kevin P. Shah, Shreya Goyal, Esha A. Hansoti, Andrew C. Caruso, Myers-Briggs Type Indicator in Medical Education: A Narrative Review and Analysis, Health Professions Education, Volume 6, Issue 1, 2020, Pages 31-46, ISSN 2452-3011, https://doi.org/10.1016/j.hpe.2019.03.002.(https://www.sciencedirect.com/science/article/pii/S245230111830124X)
11. Z. Mushtaq, S. Ashraf and N. Sabahat, "Predicting MBTI Personality type with K-means Clustering and Gradient Boosting," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-5, doi: 10.1109/INMIC50486.2020.9318078. https://doi.org/10.1109/INMIC50486.2020.9318078
12. https://medium.com/@dolly19304/personality-prediction-using-myers-briggs-type-indicator-56888416e87c
13. https://www.kaggle.com/datasets/datasnaek/mbti-type
14. Behera, D. K., Das, M., & Swetanisha, S. (2019). A Research on Collaborative Filtering Based Movie Recommendations: From Neighborhood to Deep Learning Based System. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 4, pp. 10809–10814). https://doi.org/10.35940/ijrte.d4362.118419
15. Nikam, S. S., & Dalvi, Prof. R. (2020). Fake News Detection on SocialMedia using Machine Learning Techniques. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 7, pp. 940–943). https://doi.org/10.35940/ijitee.g5428.059720
16. Wanjau, S. K., Wambugu, G. M., & Oirere, A. M. (2022). Network Intrusion Detection Systems: A Systematic Literature Review o f Hybrid Deep Learning Approaches. In International Journal of Emerging Science and Engineering (Vol. 10, Issue 7, pp. 1–16). https://doi.org/10.35940/ijese.f2530.0610722
17. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319
18. Kanani, P., & Padole, Dr. M. (2019). Deep Learning to Detect Skin Cancer using Google Colab. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 2176–2183). https://doi.org/10.35940/ijeat.f8587.088619

AUTHORS PROFILE

Rohith Muralidharan, is currently pursuing MSc Computer Science, he completed his B.Sc. Computer Science from KMM College of Arts and Science, Thrikkakkara, Ernakulam. His area of interests includes Python, Data Science, Machine Learning.

Ms. Neenu Kuriakose, joined the Department of Computer Science, St. Albert's College (Autonomous) in July, 2021 with four years of experience in Software Industry and teaching undergraduate courses in Computer Science.

She did her Bachelor's Degree in Computer Application from Rajagiri College(MG University) in 2014, Masters from UC College Aluva in 2017 and M.Phil from Amrita University in 2021.She is currently pursuing PhD. Her areas of interest include Internet of Things (IoT), Machine Learning, Computer Security and Block-chain Technology. She has published 10+ papers in professional journals. She holds two patents, entitled as "Block-chain enabled intelligent IoT architecture with AI "and "Intelligent IoT based smart irrigation system using cloud computing". She earned an Indian book of record in 2nd August 2021. She is an active national mentor of Yuva IncubatED team. Ms.Neenu received Best Research scholar award from novel research academy(2021) and Indo-Asia Best Researcher award in Computer Security era(2021).She currently conducts the Department Certificate Program in Internet of Things (IoT) . She also actively participates in continued learning through conferences and professional research.

Ms. Sangeetha J, working as an Assistant Professor and Assistant Controller of Examination, Department of Computer Science in St. Albert's College (Autonomous) having over a decade of experience in academics and industry. She joined the Albertian family in 2011. She has done her M.Phil in Computer Science and pursuing her Ph.D. She has specialized in Data Mining and has extensive knowledge in the full life-cycle of software development process including requirement gathering, design, coding, testing and debugging. She has designed various software applications for various agencies including financial enterprises and educational consultancies. She has published papers in many national and international publications. She presented papers in various national and international seminars/conferences. Her area of research is data mining.