UNIVERSITI MALAYA

WQD7006
Machine Learning for Data Science

Group Project

Prediction of Personality Types from Online Text using Machine Learning

Group Members:

Name	**Matric Number**
Jie Hongsheng	22064728
Low Boon Kiat	17138399
Mark Nicholas Akyapogu	22102968
Yuejing Huang	S2158553
Zhou Yao	S2177633

1.0 Introduction

Personality is the unique amalgamation of behaviours, feelings, and thoughts that defines an individual. Understanding personality offers profound insights into how people think, behave, and interact with the world around them (Ahmed et al., 2010). One popular framework used to categorize and understand personalities is the Myers-Briggs Type Indicator (MBTI), a widely recognized assessment tool that identifies personality types based on four dimensions: extraversion (E) vs. introversion (I), sensing (S) vs. intuition (I), thinking (T) vs. feeling (F), and judging (J) vs. perceiving (P) (Tieger, 2014).

With the abundance of data available on social media and other digital platforms, there exists an immense opportunity to leverage this textual data to discern and predict an individual's personality traits. Such predictions hold substantial promise, particularly in the realm of recruitment and talent acquisition. Predicting personality from online text can provide hiring managers with a deeper understanding of candidates beyond what traditional resumes or interviews offer. This insight enables better matching of candidates to roles, fostering improved job satisfaction and employee performance.

Machine learning algorithms have emerged as powerful tools in predicting personality from textual data. Leveraging natural language processing (NLP) techniques, these algorithms can analyse vast amounts of text to identify patterns associated with specific personality traits. By training models on labelled data sets derived from text samples linked to known personality types, machine learning algorithms can learn to predict personalities from new, unseen text data.

In this project, various machine learning algorithms were employed to develop a classifier that can predict MBTI personality type of an individual by utilizing online text as input. Section 2 reviews literature works on relevant studies; Section 3 presents the methodology; Section 4 evaluates project findings whereas Section 5 summarises the work with a conclusion.

2.0 Background / Literature Review

Understanding personality traits through online text has garnered substantial interest, leading to numerous studies employing machine learning techniques for classifying MBTI personality types. This section reviews significant works that explore the classification of MBTI personalities using textual data from various online sources.

Amirhosseini & Kazemian (2020) proposed a novel methodology for predicting MBTI personality types from textual data obtained from online forum. The proposed methodology started with data pre-processing through stop words removal and lemmatization, followed by vectorization by assigning weights to each piece of text using Term Frequency-Inverse Document Frequency (TF-IDF) and using XGBoost for classification. The model delivered satisfactory performance with an accuracy of 70-80% for each dimension (I/E, N/S, F/T, J/P).

Mushtaq et al. (2020) made improvements to the previous methodology. The authors employed K-Means Clustering to form the clusters followed by using XGBoost for classification. The results showed that the proposed modifications improved the classification accuracy to 85-90% for each dimension. However, there were room for improvement through hyperparameter tuning such as adjusting tree depth and number of iterations.

Ontoum and Chan (2022) explored multiple traditional and deep learning approaches namely Support Vector Machine (SVM), Naïve Bayes (NB), and Recurrent Neural Networks (RNN) for classification of MBTI personality types. The results showed that RNN added with Bi-directional Long Short-Term Memory (BI-LSTM) performed better than SVM and NB with an accuracy of 80-90% for each dimension. However, the overall accuracy of this model was only 49.75% indicating that it was unable to classify all four dimensions accurately.

Ryan et al. (2023) proposed using Word2Vec method to replace the commonly used TF-IDF for vectorization of textual data. The authors implemented several machine learning algorithms for the classification of MBTI personality types including Logistic Regression, Linear Support Vector, Stochastic Gradient Descent, Random Forest, XGBoost and CatBoost. The combination of Word2Vec embedding and Logistic Regression was selected as the best classification model with an average F1-score of 0.83 for each dimension.

All the preceding works employed a divide-and conquer approach for MBTI personality prediction, which was to classify within four dimensions. Sánchez-Fernández et al. (2023) proposed an alternative methodology that focused on predicting single MBTI label, i.e. the target variable consists of 16 classes. The models including NB, Logistic Regression, and three Artificial Neural Networks managed to achieve good accuracy of 90%.

3.0 Methodology

Business Understanding

The MBTI (Myers-Briggs Type Indicator) has a wide range of applications in business for team building, leadership development, communication optimisation and talent management. By measuring the psychological differences between individuals, the MBTI provides a simple and intuitive framework for understanding work styles and communication between employees to improve team effectiveness, optimise leadership development and enhance talent matching.

Data Understanding

Dataset

The dataset is publicly available on "Kaggle" and was sourced from the Personality Cafe forum. It represents an extensive collection of individuals and their corresponding MBTI personality types, encompassing test results from 8,675 respondents with diverse backgrounds. The survey explores the four dimensions of MBTI: Extraversion and Introversion, Sensing and Intuition, Thinking and Feeling, as well as Judging and Perceiving. As show in Fig 1. Each entry in the dataset comprises two columns, with one indicating the MBTI type of each respondent. The second column represents the last 50 messages posted by each respondent on the forum, with messages separated by "|||" (three space characters). In total, the dataset comprises approximately 430,000 posts, providing a rich and varied source for analyzing and understanding the relationships between personality types and online communication patterns.

	type	posts
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...
1	ENTP	'I'm finding the lack of me in these posts ver...
2	INTP	'Good one _____ https://www.youtube.com/wat...
3	INTJ	'Dear INTP, I enjoyed our conversation the o...
4	ENTJ	'You're fired.|||That's another silly misconce...
...	...	...
8670	ISFP	'https://www.youtube.com/watch?v=t8edHB_h908||...
8671	ENFP	'So...if this thread already exists someplace ...
8672	INTP	'So many questions when i do these things. I ...
8673	INFP	'I am very conflicted right now when it comes ...
8674	INFP	'It has been too long since I have been on per...

8675 rows × 2 columns

Fig. 1. "MBTI" Personality Type Dataset

Exploratory Data Analysis

Through exploratory data analysis of the MBTI (Myers-Briggs Type Indicator) dataset, we found some differences in the distribution of individual personality types among the surveyed population。

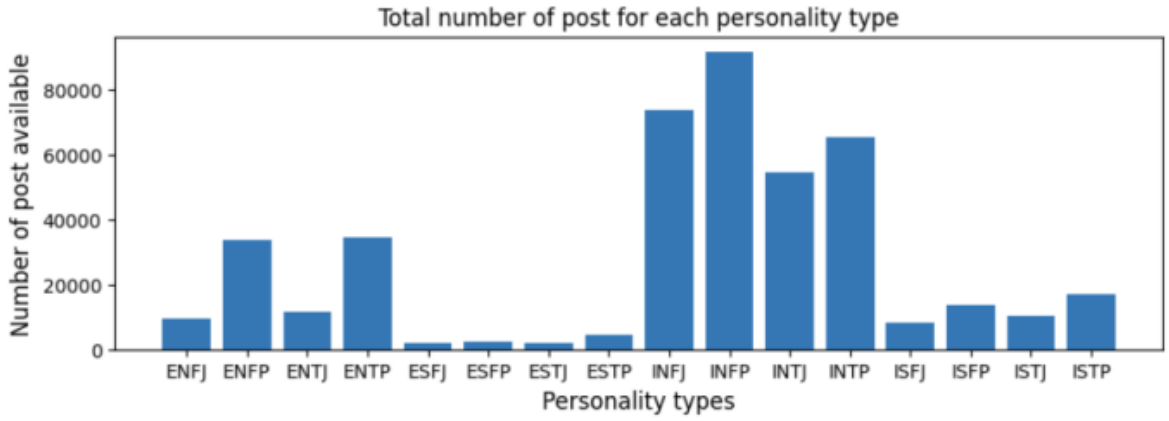


Fig 2. Proportionality diagram of "MBTI" types

The histogram of the number of posts versus each personality type (Figure 1) shows that the number of posts is concentrated in the four types INFJ, INFP, INTJ, INTP. The number of posts in INFP is the highest with more than 90,000, while the number of posts in ESTJ is the lowest. This may indicate that people with the dimensions " Introversion " and " Intuition " are active in social networks.

Fig 3. Unique word cloud for overall “MBTI” types

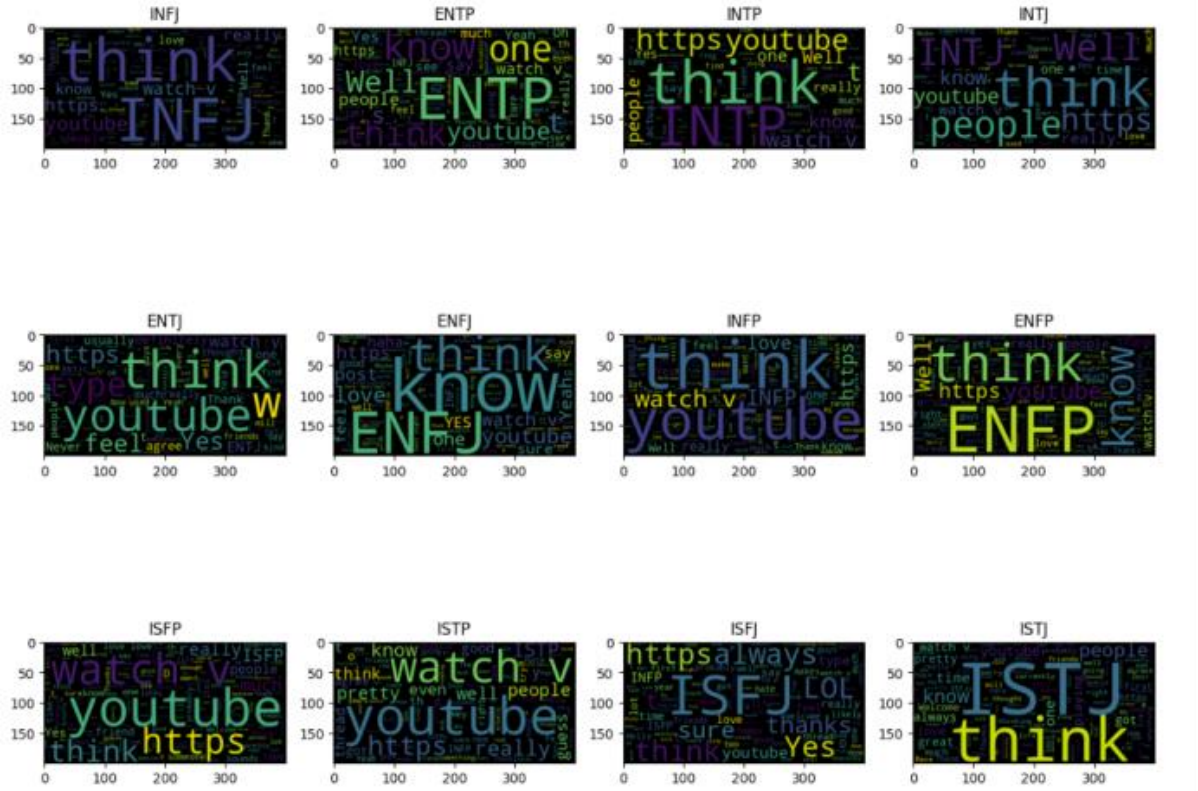


Fig 4. Unique word cloud for individual “MBTI” types

As shown in Figs. 3 and 4. The overall word cloud suggests that words such as "people", "think", and "know" are prevalent in all discussions. The typology word cloud reveals that certain words such as "youtube" and "https" may be associated with more frequent references to video content by users with specific personality types. Commonly used positive emotion words and online phrases indicate the positive and humorous nature of communication between users, while the frequency of verbs such as "think", "believe", "feel", etc. emphasises the importance of personal expressions. The frequency of verbs emphasises the importance of personal expression and introspection. These findings suggest that specific personality types may favour unique discussion topics and communication preferences.

Data Preparation

In the data processing stage, feature construction was first used. In this project, four columns of new features were constructed based on the type feature. The figure below shows the data set after construction. IE means that if the first letter of type is I, it will be set to 0, and if it is E, it will be set to 1. The following columns of data are done in the same way, and are set to 0 or 1 depending on whether they are a specific letter.

	type	posts	IE	NS	FT	JP
0	INFJ	'http://www.youtube.com/watch?v=qsXHcwe3krw|||...	0	0	0	0
1	ENTP	'I'm finding the lack of me in these posts ver...	1	0	1	1
2	INTP	'Good one _____ https://www.youtube.com/wat...	0	0	1	1
3	INTJ	'Dear INTP, I enjoyed our conversation the o...	0	0	1	0
4	ENTJ	'You're fired.|||That's another silly misconce...	1	0	1	0

Fig 5. Dataset after feature construction

For text data, preprocessing is needed to remove special characters, punctuation marks, numbers, etc. Also, text segmentation is required to split sentences into sequences of words or phrases. The techniques used in preprocessing are shown in table.

Table 1. Techniques used in data preprocessing.

No	Technique Name	Description
1	Convert MBTI personality type to binary form	Convert MBTI (Myers-Briggs Type Indicator) personality type into binary form
2	Lower text	The primary purpose of lowering the text to all lowercase is so that words such as "Hello" and "hello" would not be treated as a different word since they are the same word. It helps in reducing the number of words that the dictionary needs to hold at a time (Batista & Alexandre, 2008).
3	Removal of whitespace	Whitespace does not provide any meaning to the text, so it is removed for computational purposes.
4	Stopwords removal	Eliminate common stopwords such as "and", "the", "is" etc., as they generally do not carry significant sentiment-related meaning. After that the remaining text is used for sentimental purposes. It builds the exactness of the text.
5	Removal of punctuation	Removing the punctuation or any non-alphanumeric words from the original text.
6	Remove URL link	Remove all URL links from the text. In natural language processing tasks, URL links often do not contain useful semantic information, and if they are retained, they may interfere with text analysis.
7	Lemmatizing	Lemmatization is a text processing technique in natural language processing that aims to convert words into their base forms (called roots or stems), thereby reducing word variations and making them easier to compare and analyze.
8	Feature Engineering - TF-IDF	Use TF-IDF feature engineering to determine the relevance and importance of words to documents in a document collection

Modelling

After preprocessing and feature extraction, the dataset is split into training and testing data with a ratio of 80:20. In data understanding section, it is discovered that the data for IE and NS dimensions are unbalanced. This

imbalance could affect the performance of machine learning models trained on the data. Hence, we apply stratified K-fold cross-validation (K=5) with GridSearchCV to address the data imbalance. Stratified K-fold ensures that each fold retains the same class distribution as the original dataset. This is crucial when dealing with imbalanced classes, as it prevents certain folds from having disproportionately fewer instances of a particular class. GridSearch CV is a method for hyperparameter tuning, where it defines a grid of hyperparameters for a machine learning model and search for the best combination of hyperparameters. It performs an exhaustive search over the hyperparameter grid while ensuring that the evaluation is robust and not biased due to class imbalance. After identifying the best-tuned values for each hyperparameter, we fit the models to the training data and make predictions on the testing data. The models used for classification of MBTI personality types in this project are Naïve Bayes, Logistic Regression, Random Forest, K-Nearest Neighbours (KNN), Stochastic Gradient Descent (SGD), and Support Vector Classifier (SVC). Table below summarizes the list of hyperparameters tuned for each model.

Table 2. List of hyperparameters tuned for each model.

No.	Model	Hyperparameter
1	Naïve Bayes	The model is simple and doesn't have adjustable parameters like other models do.
2	Logistic Regression	• 'C' (regularization parameter) • 'solver' (different optimization algorithms)
3	Random Forest	• max_depth (maximum depth of tree) • n_estimators (number of trees in the forest) • min_samples_split (minimum number of samples required to split a node)
4	K-Nearest Neighbours (KNN)	• K (number of nearest neighbours)
5	Stochastic Gradient Descent (SGD)	• 'alpha' (regularization parameter) • 'loss' (different log functions)
6	Support Vector Classifier (SVC)	• -

4.0 Results and Discussion

(i) Naïve Bayes

The results from Naïve Bayes model revealed varying accuracies across personality dimensions. While F/T achieved the highest accuracy at 75%, the model struggled most in predicting J/P with an accuracy of 65%. Additionally, I/E and N/S had lower precision for their minority classes, indicating a higher tendency for misclassification in identifying Extroversion and Sensing traits. In imbalanced datasets i.e. I/E and N/S, accurately predicting the minority class is often more challenging, and a lower precision score for such classes implies a higher tendency for misclassification. Overall, Naive Bayes displayed moderate performance metrics (accuracy, precision, recall, and F1-score) across personality dimensions, suggesting that it may not be ideally suited for capturing complex relationships within the dataset. Its assumptions of feature independence might be limiting the model's capacity to fully comprehend and model intricate relationships among variables.

(ii) Logistic Regression

The Logistic Regression model demonstrated consistently high accuracies across personality dimensions. It achieved high accuracy at 80% for I/E although the recall for Extroversion was notably lower at 23%, indicating challenges in accurately identifying Extroverted traits. In the case of N/S, it attained 86% accuracy but showed a very low recall at 3% for Sensing traits, suggesting difficulty in identifying Sensing characteristics. For F/T, the model achieved 80% accuracy with balanced precision and recall for both traits. In J/P, the accuracy reached 73% with reasonably balanced precision and recall for both traits. Overall, the Logistic Regression model demonstrated high accuracy and balanced F1-scores across personality dimensions, indicating its proficiency in modelling linear relationships within the data. The analysis result is supported by Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A curve closer to the upper left corner signifies better model performance, indicating higher True Positive Rate (TPR) and lower False Positive Rate (FPR) across different threshold.

(iii) Random Forest

The results from the Random Forest Classifier demonstrate varying performance across different personality dimensions. For the I/E dimension, the model achieved an accuracy of 77%, yet it struggled to identify Extroversion with a precision and recall of 0%. Similar issues were observed in the N/S dimension, where the model showed high accuracy of 86% but failed to predict Sensing traits with precision and recall of 0.00. In contrast, for the F/T dimension, the model displayed reasonable accuracy of 76% with balanced precision and recall scores for both classes. The J/P dimension demonstrated an accuracy of 65%, presenting a high precision of 86% for Judging but a low recall of 15%, suggesting an inclination towards false negatives in identifying Judging traits. Overall, the model showed considerable shortcomings in identifying certain traits like Extroversion, Sensing, and Perceiving, indicating potential overfitting or sensitivity to specific features within the dataset.

(iv) K-Nearest Neighbours (KNN)

K-Nearest Neighbours, utilized in this study for personality prediction, demonstrated competitive performance across different dimensions. With an overall accuracy of 78%, the model exhibited notable proficiency in capturing personality traits. Notably, it excelled in predicting N/S with an accuracy of 86%, indicating a strong ability to discern between Intuition and Sensing traits. However, challenges were observed in distinguishing between F/T and J/P traits, where the accuracy rates were 68% and 67%, respectively. The model's balanced performance, as reflected in precision and recall scores, suggests its effectiveness in handling diverse personality dimensions. The higher recall for I/E and J/P compared to precision suggests a tendency for fewer false negatives but a slightly increased likelihood of false positives, emphasizing the need for fine-tuning in specific trait identification.

(v) Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent, applied to predict personality dimensions, highlighted robust performance with an overall accuracy of 80%. The model excelled in discerning N/S traits with an accuracy of 86%, although it struggled in accurately predicting Extroversion (I/E) with a precision of 0.98 and a recall of 0.19. The F/T dimension achieved balanced accuracy at 78%, indicating a well-rounded performance in capturing the Thinking and Feeling traits. In J/P, the model displayed a balanced accuracy of 72%, with precision and recall scores suggesting an even identification of Judging and Perceiving traits. These findings underscore SGD's proficiency across diverse personality dimensions, but attention to precision in specific traits, such as Extroversion and Sensing, could enhance its overall performance.

(vi) Support Vector Classifier (SVC)

The Support Vector Classifier, employed for personality prediction, demonstrated commendable accuracy at 80%. Notably, it excelled in capturing N/S traits, achieving an accuracy of 86%, highlighting its ability to distinguish between Intuition and Sensing. However, challenges were evident in predicting J/P traits, where the accuracy dropped to 70%. In terms of precision and recall, I/E dimension exhibited a higher recall for Introversion (98%) but a lower recall for Extroversion (15%). The same applies to N/S dimension, with very low precision and recall for Sensing trait. The F/T and J/P dimensions achieved balanced precision and recall scores, indicating a well-rounded performance. The precision and recall metrics highlight SVC's ability to accurately identify specific traits, but attention is warranted for achieving a more balanced identification, particularly in traits like Extroversion and Sensing.

Table 3. Accuracy of each model across four personality dimensions.

Model	I/E	N/S	F/T	J/P
Naïve Bayes	0.69	0.70	0.75	0.65
Logistic Regression	0.80	0.86	0.80	0.73
Random Forest	0.77	0.86	0.76	0.65
KNN	0.78	0.86	0.68	0.67
SGD	0.80	0.86	0.78	0.72
SVC	0.79	0.86	0.79	0.70

Table 4. Precision of each model across four personality dimensions.

Model	I/E (0,1)	N/S (0,1)	F/T (0,1)	J/P (0,1)
Naïve Bayes	0.75 (0.86, 0.39)	0.82 (0.91, 0.24)	0.75 (0.77, 0.72)	0.65 (0.55, 0.72)
Logistic Regression	0.78 (0.81, 0.68)	0.84 (0.86, 0.67)	0.80 (0.80, 0.79)	0.72 (0.71, 0.73)
Random Forest	0.59 (0.77, 0.00)	0.74 (0.86, 0.00)	0.78 (0.73, 0.83)	0.73 (0.86, 0.64)
KNN	0.75 (0.78, 0.65)	0.88 (0.86, 1.00)	0.71 (0.63, 0.80)	0.66 (0.65, 0.68)
SGD	0.78 (0.80, 0.72)	0.82 (0.86, 0.54)	0.79 (0.86, 0.72)	0.72 (0.69, 0.74)
SVC	0.78 (0.79, 0.73)	0.74 (0.86, 0.00)	0.79 (0.80, 0.77)	0.70 (0.70, 0.70)

Table 5. Recall of each model across four personality dimensions.

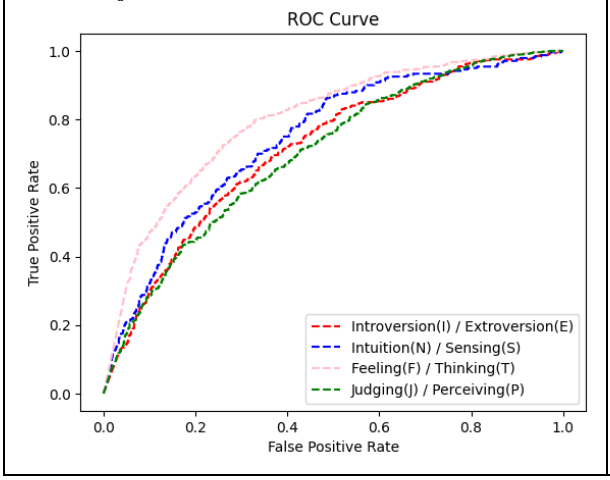
Model	I/E (0,1)	N/S (0,1)	F/T (0,1)	J/P (0,1)
Naïve Bayes	0.69 (0.71, 0.60)	0.70 (0.73, 0.55)	0.75 (0.76, 0.73)	0.65 (0.60, 0.68)
Logistic Regression	0.80 (0.97, 0.23)	0.86 (1.00, 0.03)	0.80 (0.83, 0.75)	0.73 (0.53, 0.86)
Random Forest	0.77 (1.00, 0.00)	0.86 (1.00, 0.00)	0.76 (0.89, 0.61)	0.65 (0.15, 0.98)
KNN	0.78 (0.99, 0.08)	0.86 (1.00, 0.00)	0.67 (0.92, 0.37)	0.67 (0.37, 0.87)
SGD	0.80 (0.98, 0.19)	0.86 (1.00, 0.03)	0.78 (0.71, 0.86)	0.72 (0.55, 0.84)
SVC	0.79 (0.98, 0.15)	0.86 (1.00, 0.00)	0.79 (0.81, 0.76)	0.70 (0.42, 0.88)

Table 6. F1-score of each model across four personality dimensions.

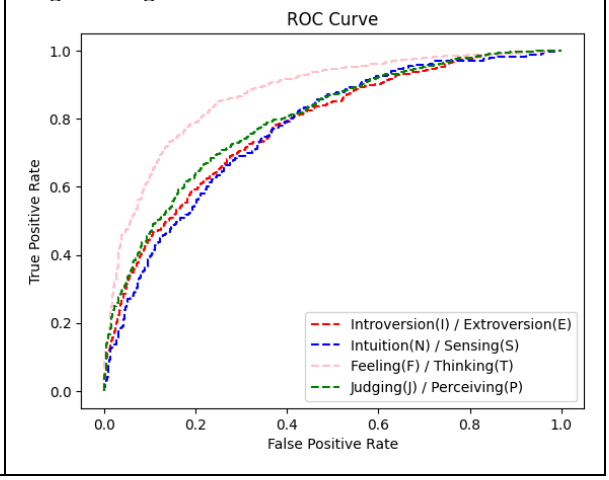
Model	I/E (0,1)	N/S (0,1)	F/T (0,1)	J/P (0,1)
Naïve Bayes	0.71 (0.78, 0.47)	0.74 (0.81, 0.34)	0.75 (0.77, 0.73)	0.65 (0.57, 0.70)
Logistic Regression	0.76 (0.88, 0.35)	0.80 (0.93, 0.05)	0.80 (0.82, 0.77)	0.72 (0.60, 0.79)
Random Forest	0.67 (0.87, 0.00)	0.80 (0.93, 0.00)	0.76 (0.80, 0.70)	0.57 (0.25, 0.77)
KNN	0.71 (0.87, 0.15)	0.80 (0.93, 0.01)	0.64 (0.75, 0.50)	0.65 (0.47, 0.76)
SGD	0.75 (0.88, 0.31)	0.81 (0.93, 0.06)	0.78 (0.78, 0.78)	0.72 (0.61, 0.79)
SVC	0.73 (0.88, 0.24)	0.80 (0.93, 0.00)	0.79 (0.80, 0.77)	0.68 (0.52, 0.78)

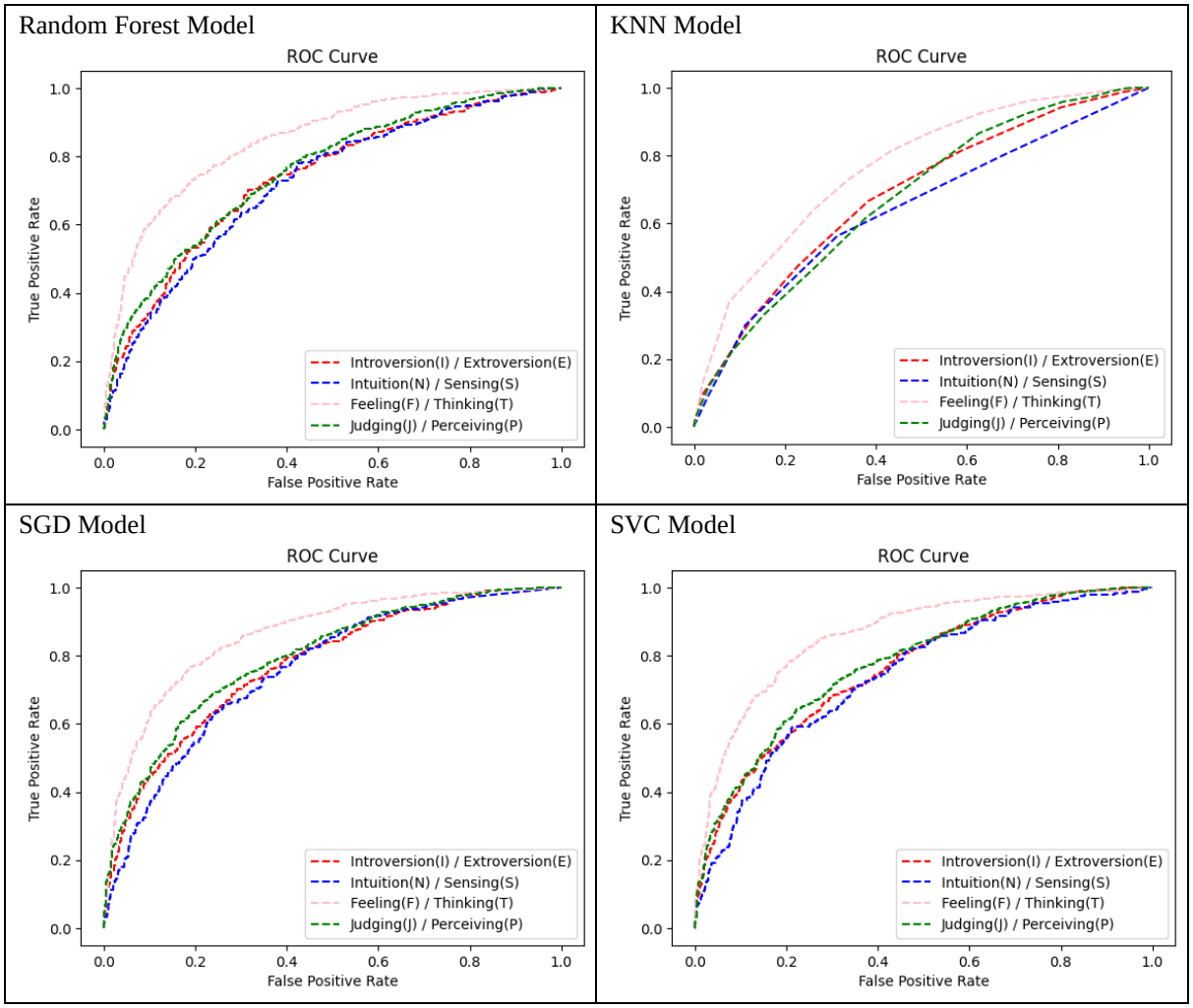
Table 7. AUC-ROC curves of each model across four personality dimensions.

Naïve Bayes Model



Logistic Regression Model





5.0 Conclusion

Across the models assessed for personality prediction, Logistic Regression demonstrated robust overall performance. The model achieved the highest accuracy for I/E (80%) and N/S (86%), reflecting its proficiency in distinguishing these dimensions. Additionally, Logistic Regression demonstrated notable precision and recall scores for most categories, underlining its balanced identification of personality traits. Naïve Bayes, while offering competitive accuracy in some dimensions, struggled with recall for specific classes (e.g., Extroversion in I/E) and exhibited a higher tendency for misclassification as evident from lower precision scores. Random Forest and KNN models showcased varied performances across dimensions, showing strengths in certain aspects but limitations in precision and recall for specific classes. SVC provided stable and moderate performance across the dimensions, emphasizing balanced precision and recall. In summary, Logistic Regression appears to be the best-performing model among the evaluated ones due to its consistently high accuracy, precision, recall, and F1-scores across different personality dimensions (I/E, N/S, F/T, J/P). It consistently demonstrated balanced performance in identifying various personality traits, making it a reliable choice for personality prediction.

To further enhance the accuracy and robustness of personality prediction models, several avenues for future research can be explored. Addressing the challenge of class imbalance, especially in predicting minority classes within personality dimensions, is a crucial area for improvement. Future efforts may involve employing sophisticated techniques like oversampling or undersampling to balance class distributions. This can help prevent biases toward majority classes and improve the models' capacity to predict rarer personality traits accurately. Additionally, further refinement of hyperparameters for the selected models could lead to optimized performance. Conducting exhaustive searches over hyperparameter spaces or employing more sophisticated optimization algorithms can help in achieving superior model performance across various personality dimensions.

6.0 Reference

Ahmed, F., Campbell, P., Jaffar, A., Alkobaisi, S., & Campbell, J. (2010). Learning & personality types: A case study of a software design course. *Journal of Information Technology Education: Innovations in Practice*, *9*, 237–252. https://doi.org/10.28945/1329

Amirhosseini, M. H., & Kazemian, H. (2020). Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, *4*(1), 9. https://doi.org/10.3390/mti4010009

Batista, L. V., & Alexandre, L. A. (2008). Text pre-processing for lossless compression. *Data Compression Conference*. https://doi.org/10.1109/dcc.2008.78

Mushtaq, Z., Ashraf, S., & Sabahat, N. (2020). Predicting MBTI personality type with k-means clustering and gradient boosting. *2020 IEEE 23rd International Multitopic Conference (INMIC)*. https://doi.org/10.1109/inmic50486.2020.9318078

Ontoum, Sakdipat, & Chan, Jonathan. (2022). Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning. https://doi.org/10.48550/arXiv.2201.08717

Ryan, G., Katarina, P., & Suhartono, D. (2023). MBTI personality prediction using machine learning and smote for balancing data based on statement sentences. Information, 14(4), 217. https://doi.org/10.3390/info14040217

Sánchez-Fernández, P., Baca Ruiz, L. G., & Pegalajar Jiménez, M. del. (2023). Application of classical and advanced machine learning models to predict personality on social media. Expert Systems with Applications, 216, 119498. https://doi.org/10.1016/j.eswa.2022.119498

Tieger, P. (2014). Do what you are: Discover the perfect career for you through the secrets of personality type. Little, Brown and Company.