Myers-Briggs Type Indicator

Apeksha Shah

Department of Computer Science

California State University Fullerton

Fullerton, USA

apeksha@csu.fullerton.edu

Malay Koladiya

Department of Computer Science

California State University Fullerton

Fullerton, USA

malay.koladiya@csu.fullerton.edu

***Abstract*— The paper focuses on analyzing different machine learning algorithms for predicting the personality type of an individual based on the Myers-Briggs Type Indicator (MBTI) from the data collected through social media. MBTI is the most famous and widely used method for predicting personality. In this study, we evaluate the performance and compare the accuracies of different machine learning techniques Such as Logistic Regression, Support Vector Machine, Random Forest, and K-Nearest Neighbor. Our biggest motivation to perform personality prediction is to help the people understand their type and strength or weaknesses which can help them to understand themselves better.**

Keywords—Myers-Briggs Type Indicator, Word clouds, Lemmatization, Logistic Regression, KNN, Random Forest, Support Vector Machine

I. INTRODUCTION

The Myers-Briggs Type Indicator (MBTI) is a method for assessing and identifying an individual's personality or character. The personality of an individual is defined as a collection of traits and characteristics that indicate their unique thoughts, feelings, and behavior [1]. A person's traits change and develop over time and circumstances. In other words, personality is a cluster of features that combine to form a unique personality of an individual.

Woodsmall created meta-programs for use in therapy and business, combining the MBTI with them [2]. Metaprograms are essentially a collection of scales of various types that enable dynamically recognize preferences to react in a certain situation. The Language and Behavior (LAB) profile, created by Roger Bailey in the 1980s, is a study that evokes 14 major meta-programs, separated into two categories: Motivation Traits and Working Traits[2]. In 1988, the findings were presented in the book "Timeline Therapy and the Basis of Personality." Patterns were re-created, and a reduced collection of meta programs was created, consisting of only four basic important meta-programs. The MBTI consists of four basic meta programs. MBTI is a four-dimensional model widely used to check the type of personality of the individuals. MBTI divides everyone into 16 distinct personalities based on four dimensions. The first dimension is defined as Introvert, denoted as I, and Extrovert denoted as E. The second dimension is defined with Intuitive, defined as N, and Sensing, denoted as S. The third dimension is defined with Thinking, denoted as T, and Feeling, denoted as F. The fourth dimension is defined with Judging, denoted as J, and Perceiving denoted as P.

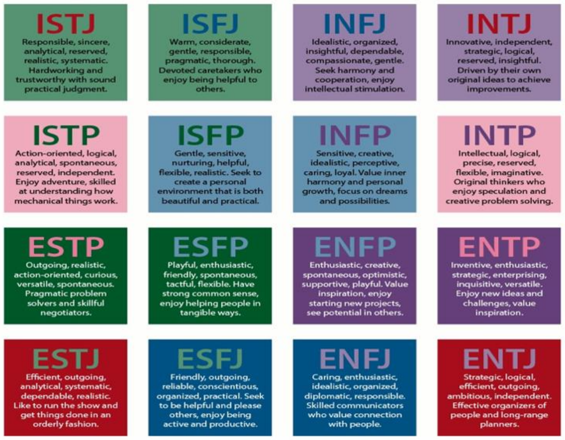
II. BACKGROUND

A. *Machine Learning*

Machine learning is the study of algorithms using data to make the computer or machine able to make predictions and make decisions the way humans learn. Sometimes humans cannot fetch the information or pattern from the huge or vast data for that machine learning to take place. Three types of machine learning systems are Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Supervised and unsupervised learning are popular and widely used algorithms [3]. The statistical dataset is used as the input in any ML algorithm, which gives the output in terms of classified or predicted in the different classes. As the name suggests, the supervised algorithm works under supervision. Labeled data set is used to train the supervised model and tests on the testing dataset, which predicts the correct label for the data. Unlike supervised learning, the unsupervised algorithm detects or recognizes the pattern in a given dataset and tests that model on the test dataset. Logistic Regression and K-Nearest Neighbor (KNN) unsupervised learning algorithms, on the other hand, merely fetch a few features or a pattern from the data. When a new instance is found, it is more chance to put new data into the predefined pattern. Examples include “K-Means,” “Mean Shift,” and “K models” [4]. “Reinforcement learning” comes into the picture when the result of any task should be a sequence of opinions. During the training phase, the artificial robot is trained on the matrix of rewards or fines for the acts it does. This learning process is designed to maximize total reward. Reinforcement learning algorithms include “Q-Learning” and the “Markov Decision Process.”

B. *Types of MBTI Personalities*

The system is divided into four personality dimensions, and there are 16 distinct personality types made from the combination of four core personalities. The four dimensions are following: “Introversion (I) – Extraversion (E),” “Intuition (N) – Sensation (S),” “Feeling (F) – Thinking (T),” and “Perception (P) – Judgment (J).” The 16 personality types are then categorized based on the MBTI, combining the four core personality dimensions. The 16 distinct personalities are shown in figure 1 [5].



ISTJ
Responsible, sincere, analytical, reserved, realistic, systematic. Hardworking and trustworthy with sound practical judgment.

ISFJ
Warm, considerate, gentle, responsible, pragmatic, thorough. Devoted caretakers who enjoy being helpful to others.

INFJ
Idealistic, organized, insightful, dependable, compassionate, gentle. Seek harmony and cooperation, enjoy intellectual stimulation.

INTJ
Innovative, independent, strategic, logical, reserved, insightful. Driven by their own original ideas to achieve improvements.

ISTP
Action-oriented, logical, analytical, spontaneous, reserved, independent. Enjoy adventure, skilled at understanding how mechanical things work.

ISFP
Gentle, sensitive, nurturing, helpful, flexible, realistic. Seek to create a personal environment that is both beautiful and practical.

INFP
Sensitive, creative, idealistic, perceptive, caring, loyal. Value inner harmony and personal growth, focus on dreams and possibilities.

INTP
Intellectual, logical, precise, reserved, flexible, imaginative. Original thinkers who enjoy speculation and creative problem solving.

ESTP
Outgoing, realistic, action-oriented, curious, versatile, spontaneous. Pragmatic problem solvers and skillful negotiators.

ESFP
Playful, enthusiastic, friendly, spontaneous, tactful, flexible. Have strong common sense, enjoy helping people in tangible ways.

ENFP
Enthusiastic, creative, spontaneous, optimistic, supportive, playful. Value inspiration, enjoy starting new projects, see potential in others.

ENTP
Inventive, enthusiastic, strategic, enterprising, inquisitive, versatile. Enjoy new ideas and challenges, value inspiration.

ESTJ
Efficient, outgoing, analytical, systematic, dependable, realistic. Like to run the show and get things done in an orderly fashion.

ESFJ
Friendly, outgoing, reliable, conscientious, organized, practical. Seek to be helpful and please others, enjoy being active and productive.

ENFJ
Caring, enthusiastic, idealistic, organized, diplomatic, responsible. Skilled communicators who value connection with people.

ENTJ
Strategic, logical, efficient, outgoing, ambitious, independent. Effective organizers of people and long-range planners.

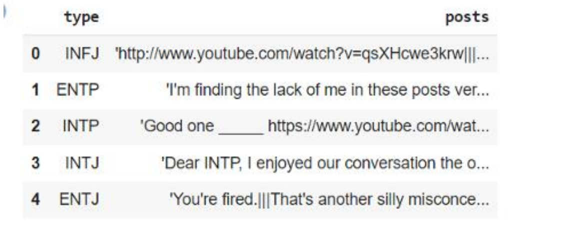
Figure 1: 16 distinct personalities

Each of the 16 keywords characterizes an individual’s personality. The machine learning model learns from the given data (text input dataset) and determines a user’s MBTI type. For instance, people classified as “ESFJs” prefer “Extroversion (E),” “Sensing (S),” “Feeling (F),” and “Judgment (J)” personality traits. We can learn the patterns in the text data of an individual using machine learning algorithms, and according to the labels, we can classify the behavior or needs of any individual whose data is available.

III. DATASET AND DATA PREPROCESSING

A. *Dataset*

For the purpose of this research, we used the dataset of the Myers-Briggs personality type that was publicly available. We obtained the dataset from Kaggle. The dataset consists of 2 columns and 8765 rows. The data available in the first column describes the type of MTBI personality of an individual. The data in the second column of the dataset contains the last 50 social media posts and comments of that individual. Each post or comment entry is separated by three pipe structures (|||), and this data was collected through the PersonalityCafe forum [6]. The dataset is shown in figure 2.



	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...

Figure 2: Kaggle dataset

B. *Data Analysis*

In the data analysis part, graphs were generated for the visual representation of the data. We plot the chart to see
the proportionality of all types of MBTI categories. The graph is shown in figure 3.

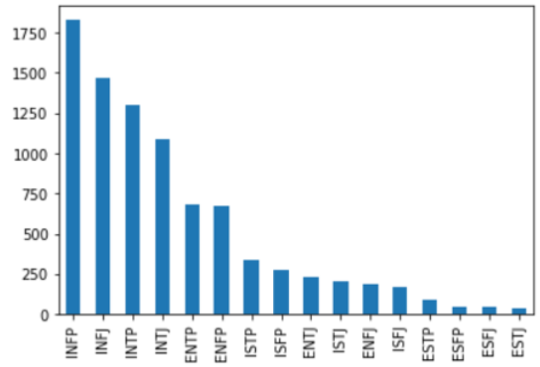


Figure 3: Proportionality of all distinct personalities in the dataset

This graph shows that the classes in the dataset are heavily imbalanced. We observe that the INFP MBTI type is the most common type of personality found in the dataset. Because the dataset contains the comments and posts from social media, we can say that most social media users are perceptive, introverted, emotional, and intuitive.

Then we looked at the most common type of words people use in the post. The common type of words used were ‘do,’ ‘don’t,’ ‘can,’ ‘I,’ ‘me,’ ‘are,’ ‘have,’ ‘be,’ ‘as,’ ‘this,’ etc. Most of these words are useless words or ‘stopwords.’ These words do not contribute to predicting the personality of any individual. Moreover, these words will act as noise in the data for the machine learning algorithms and may decrease the performance of the machine learning models. The common words need to be removed from the post to improve the performance.

To further analyze the data, we formed word clouds to determine the frequency of the words used by people. This helped us to look at the type of words that are frequently used by the people and illustrate how different MBTI personality types of people use language in unique ways. The word clouds are shown in figure 4.

The bigger the size of the word seen in each MBTI type in figure 4, the more frequently that word is used by that MBTI type person. For example, INTP type person uses words like ‘think’ and ‘INTP’ frequently, whereas a person with type ISFJ uses words like ‘LOL,’ ‘yes,’ and ‘sure’ more regularly. From the figure, we can see that people use the MBTI personality type of themselves in the post.

As a result, we must remove all the stopwords, MBTI type, and irrelevant or common words to clean the dataset to improve the proportionality of each type of MBTI personality. In addition to the words, we will also be dropping link/hyperlinks because it does not contribute to an individual’s personality.

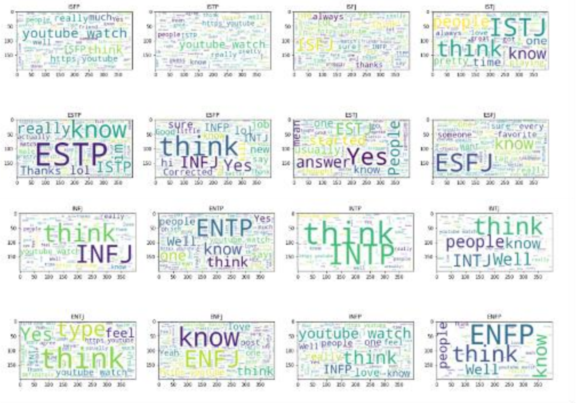


Figure 4: Word Clouds

C. *Data Preprocessing*

Selective Removal of Character and Words

The data contains the links and websites that create the model’s complexity to learning patterns. So, we eliminated several data points that had links to websites. To remove the ‘Stop words,’ we used the Natural Language Tool available in python. ‘Stop Word’ are the commonly used words that can be programmed to be ignored when search query results are retrieved or when indexing entries for searching. These commonly used words are ‘the,’ ‘an,’ ‘a', etc. In addition, we removed punctuations, multiple full stops, any non-words, short or long words, and MBTI type. Then we converted the ‘posts’ data into lower case and removed words with multiple letters repeating.

Lemmatization

We used the Lemmatization technique to group the different modulated words with the same meaning and transform them into their dictionary form (e.g., finally, final, finalized will be replaced with Final). We have used the “nltk.stem.wordNetLemmatizer” in our code to implement Lemmatization on a different post. The stemming technique is similar to the Lemmatization. However, Stemming replaces the words with the most common alphabet in different words, while Lemmatization replaces the words with their dictionary form, which must be some meaningful word.

Converting MBTI type into Four Dimensions

As we see in figure 3, the proportionality of all the MBTI types is imbalanced. We converted the 16 MBTI classes into four dimensions of balanced classes. The first dimension is IE: Extrovert (E) vs. Introvert (I), where E is 1, and I is 0. The second dimension is NS: Intuition (N) vs. Sensing (S), where N is 0 and S is 1. The third dimension is FT: Feeling (F) vs. Thinking (T), where F is 0 and T is 1. The fourth dimension is PJ: Perceiving (P) vs. Judging (J), where P is 0 and J is 1. After converting the 16 personalities into four dimensions, we plot the distribution of each type in the graph to see the proportionality. The graph is shown in figure 5.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc1b6ec1f10>
```

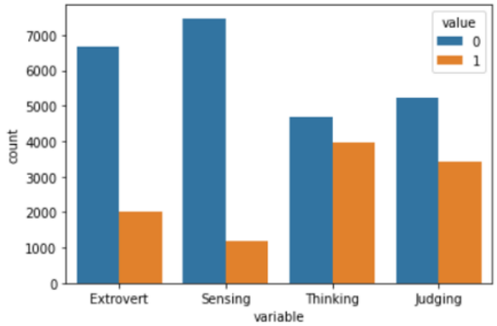


Figure 5: proportionality of each class after binarizing target features

Then using the Pearson correlation coefficient, we generated a heat map to measure the relationship between each class. The heat map is shown in figure

6. Based on the heat map, there were no correlations found.

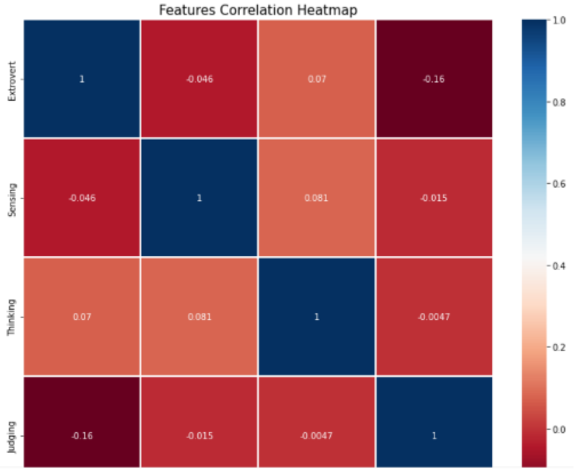


Figure 6: Features Correlation Heatmap

Vectorization

We have used the count and TF-IDF vectorizer to vectorize the dataset. The count vectorizer makes the count matrix by counting the frequency of words between 10% to 70% of posts, and TF-IDF measures how relevant or essential a word is to a collection of posts. After performing the vectorization, the dataset was updated with new columns. Now, our dataset has 8675 rows and 595 columns.

Data Splitting

To train our models on the data, we have split the dataset into two parts: train and test. We used the "train-test splitting" technique offered by Scikit-Learn. We secured 67 % of our data for training and 33% data for testing. Our model was trained on random labeled data and testing on the unknown unlabeled data.

IV. RESULTS AND MODEL EVALUATION

We have implemented four different Machine learning classification algorithms. K-Nearest Neighbor, Logistic Regression, Support Vector Machine, and Random Forest Classifier. For each algorithm, we gave 67% of the data as training data and 33% data as testing data. For K-fold Cross-Validation, the number of splitting iterations was set to 5, test size was set to 0.3, train size was set to 0.7, and random state was set to 27. For K-Nearest Neighbor, the tuning parameter was set to 2. To handle the imbalance in the data, we used the Random-Over-Sampling technique imported from the 'imblearn' python package. Using random oversampling the sampling strategy was 'minority' to target minority classes for resampling.

This section summarizes the outcomes of the study. we have evaluated the prediction performance of four different classifications for each dimension (IE -Introversion / Extraversion, NS - Intuition /Sensation, FT - Feeling– Thinking, and PJ - Perception – Judgment.). Several evaluation methods were used to compare the results such as Precision, Recall, F1, and K-fold cross-validation and confusion matrix.

Table 1 shows the accuracies of four models in all four personality types. For IE, NS, and JP dimensions, Random Forest is the most efficient algorithm with **92.81%**, **98.24%**, **71.96%,** and **72.80%** accuracies, respectively. However, for the FT dimension, SVM is the most efficient algorithm, with **74.43%** accuracy. The reason behind the good performance of Random Forest is, that the algorithm performs efficiently well with large datasets. The Random Forest algorithm combines the output of different subset decision trees which reduces the variance and prevents overfitting that improves the performance. On the other hand, KNN does not perform well on FT and JP.

Models	IE	NS	FT	JP
KNN	91.40%	94.45%	52.50%	50.18%
Logistic Regression	66.69%	69.10%	73.18%	61.45%
SVC	85.25%	94.65%	**74.43%**	70.84%
Random Forest	**92.81%**	**98.24%**	71.96%	**72.80%**

Table 1: Accuracies for all the 4 personality type with all classification algorithms

Models	IE	NS	FT	JP
KNN	67.69%	86.17%	53.08%	56.78%
Logistic Regression	76.96%	86.42%	**72.69%**	63.50%
SVC	**77.04%**	**86.42%**	72.66%	**64.17%**
Random Forest	76.83%	86.42%	68.21%	62.02%

Table 2: Accuracies for each dimension with all classification algorithms using k-fold cross-validation where k = 5

Table 2 represents the accuracy of the same models trained using k-fold cross-validation where the value of k is 5. When using cross-validation for training and testing, the Support Vector Machine performed well with **77.04%** accuracy for the IE dimension, **86.42%** accuracy for the NS dimension, **72.66%** for the FT dimension, and **64.17%** accuracy for the JP dimension.

Tables 3,4,5, and 6 shows the performance metrics for Random Forest

Classes	Precision	Recall	F1	Support
Introvert	0.91	0.95	0.93	2225
Extrovert	0.94	0.91	0.93	2182
Average	0.93	0.93	0.93	4407

Table 3: Random Forest classification report for IE dimension

Classes	Precision	Recall	F1	Support
Intuition	0.97	1.00	0.98	2429
Sensing	1.00	0.97	0.98	2507
Average	0.98	0.98	0.98	4936

Table 4: Random Forest classification report for NS dimension

Classes	Precision	Recall	F1	Support
Feeling	0.72	0.72	0.72	1563
Thinking	0.72	0.72	0.72	1536
Average	0.72	0.72	0.72	3099

Table 5: Random Forest classification report for FT dimension

Classes	Precision	Recall	F1	Support
Perceiving	0.70	0.80	0.75	1748
Judging	0.76	0.65	0.70	1712
Average	0.73	0.73	0.73	3460

Table 6: Random Forest classification report for PJ dimension

Figure 7,8,9, and 10 shows the confusion matrix for each dimension of Random Forest

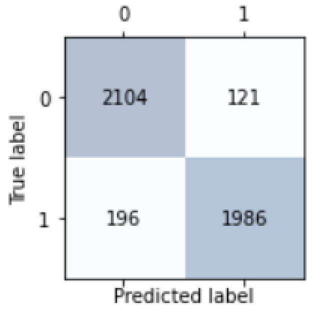


Figure 8: Confusion matrix for Introvert-Extrovert dimension where 0 represent Introvert and 1 represents Extrovert

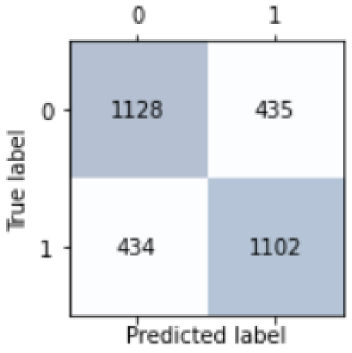


Figure 9: Confusion matrix for Filling-Thinking dimension where 0 represents Filling and 1 represents Thinking

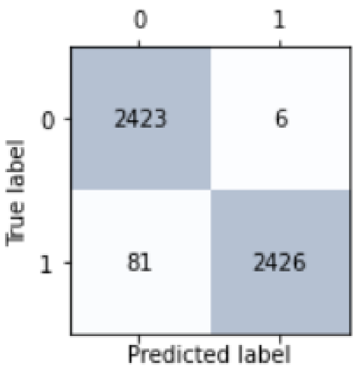


Figure 10. Confusion matrix for Intuition-Sensing dimension where 0 represents Intuition and 1 represents Sensing

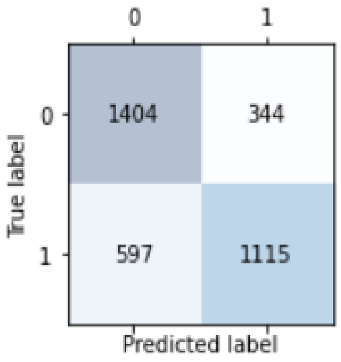


Figure 11: Confusion matrix for Judging-Perceiving dimension where 0 represents Perceiving and 1 represents Judging

Figure 11,12,13, and 14 shows the visual representation of the performance of all classifiers across each

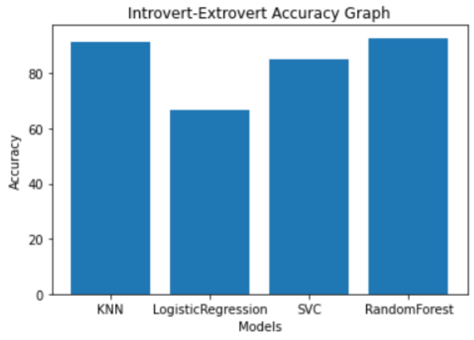


Figure 11: Accuracies of different classifiers for Introvert-Extrovert

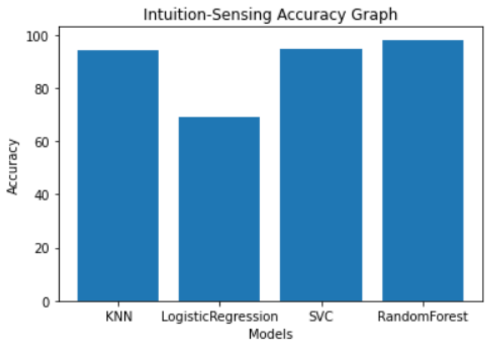


Figure 12: Accuracies of different classifiers for Intuition-Sensing

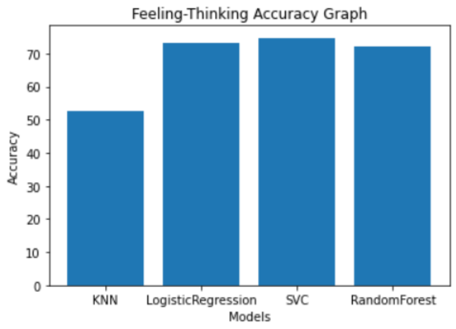


Figure 13: Accuracies of different classifiers for Feeling-Thinking

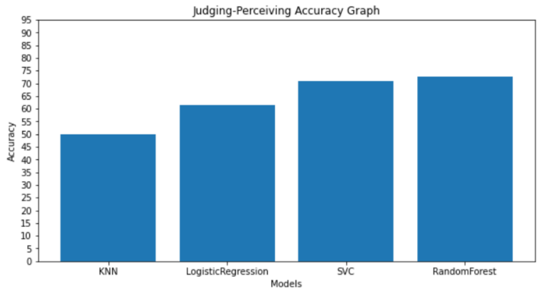


Figure 14: Accuracies of different classifiers for Judging-Perceiving

Overall, we see that the accuracies for IE and NS dimensions are generally higher in all classification algorithms compared to JP and FT dimensions, even when using K-fold cross-validation. The reason for the low accuracies of JP and FT is that there are fewer data points available for those dimensions.

V. CONCLUSION

This study provides the work for automating the personality detection based on the Myers-Briggs Type Indicator program. Several python libraries were used to develop the MBTI personality indicator, such as seaborn, re, sklearn, pandas, matlplotlib, numpy, and natural language processing toolkit. Performance of different machine learning techniques such as Random Forest, Support Vector Machine, KNN, and Logistic Regression was evaluated and compared. On evaluation, Random Forest and Support Vector Machine performed well. The accuracies of Extrovert-Introvert (IE) and Sensing-Intuitive (NS) classes were high in all performed ML algorithms.

The further evaluation concluded that all the machine learning techniques implemented in this work performed well for Intuition-sensing (NS) class. To deal with the imbalanced dataset, we have implemented Random Over Sampling. After implementing these techniques, we evaluated and compared the performance. In the future, we intend to use the “Recurrent Neural Networks” technique to predict the MBTI personality type and compare the results with the existing work [7].

REFERENCES

[1] 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.

[2] Hall, L.M.; Bodenhomer, B.G. Figuring Out People, Design Engineering with Meta-Programs; Crown House Publishing Ltd.: Carmarthe, UK.

[3] Misra, S., Li, H., & He, J. (2019). Machine Learning for Subsurface Characterization (1st ed.). Gulf Professional Publishing.

[4] Sah, S. (2020). Machine Learning: A Review of Learning Types. Machine Learning: A Review of Learning Types. Published. https://doi.org/10.20944/preprints202007.0230.v1.

[5] Beech, J. (2014, January 28). A chart with descriptions of each Myers-Briggs personality type and the four dichotomies central to the theory [Chart].

https://commons.wikimedia.org/w/index.php?title=File:MyersBriggsTypes.png&oldid=538599539.

[6] Mitchelle, J.; Myers-Briggs Personality Type Dataset. Includes a Large Number of People’s MBTI Type and Content Written by Them. Available online:

https://www.kaggle.com/datasnaek/mbti-type

[7] S. Ontoum and J. H. Chan, “Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning,” *arXiv.org*, 21-Jan-2022. [Online]. Available: https://arxiv.org/abs/2201.08717. [Accessed: 15-Mar-2022].