A Comprehensive Examination of Machine Learning Models in Predicting 16 Personality Traits

Aroma Khan   
*KIET Group of Institutions Delhi-NCR   
Ghaziabad, India*aroma.2024cse1163@kiet.edu

Preeti Garg\*  
*KIET Group of Institutions Delhi-NCR   
Ghaziabad, India*  
preeti.itgarg@gmail.com Ashish Kumar  
*KIET Group of Institutions Delhi-NCR   
Ghaziabad, India*  
ashish.2024it1118@kiet.edu

Rohit Vashisht

KIET Group of Institutions Delhi-NCR   
Ghaziabad, India  
rohit.vashisht@kiet.edu

Harshit Menaria  
*KIET Group of Institutions Delhi-NCR   
Ghaziabad, India*  
harshit.2024csit1169@kiet.edu

*Abstract*— This work implemented various machine learning frameworks to predict the personality traits using the 16 traits included in the MBTI (Myers-Briggs Type Indicator). The data used in the research includes demographic information and self-identification. The data set is preprocessed to convert categorical factors, such as gender, into numerical values. Multinomial logistic regression using variables such as age, gender, and recall scores based on 16 MBTI personality traits was used to solve the multinomial classification problem. The data is used to train the model, and the Newton-CG solver is used to optimize the model parameters to improve accuracy and convergence. The effectiveness of the model was evaluated by validating it on demographic data providers' testing data and MBTI 16 personality traits. The purpose of this study is to estimate the personal characteristics of each individual and explain the results easily for understanding. The results of this study demonstrate the suitability of logistic regression and MBTI 16 personality types for accurate personality prediction. The results demonstrate the potential of machine learning to provide insight into human behavior and preferences across a variety of domains, including emotions, human resources, and personalized recommendations.

Keywords—Machine Learning, Personality Traits, LightGBM, Accuracy, Recall, Random Forest

# Introduction (*Heading 1*)

It's common practice to determine someone's personality based on their nature. These were formerly completed manually, which took a long time to determine the person's personality. Businesses that prioritize their customers, including those in retail, finance, communications, and marketing, are the ones who employ data mining the most frequently these days. The data was analyzed using a variety of methods, including surveys, questionnaires, interviews, classroom exercises, data from social media networks and retail websites on user experiences and challenges. On the other hand, traditional methods are scale- and time-limited. The user's personality will be made clear by our suggested technology. Determining an individual's nature to determine their personality is not a novel concept. The way a person interacts with the outside world and their environment can also be influenced by their personality. Among other things, personality may be utilized as a differentiator in career counseling, health counseling, and hiring processes. It's common practice to analyze a person's behavior to infer their personality. It required a lot of time and work to forecast personalities manually [1].

No personality type is "better" or "best" than any other. Some of the personality types are explained here.

## Introversion (I) and Extraversion (E)- The extraversion-introversion dichotomy explains human behavior and interactions with the outside world. In contrast to introverts, who value thoughtful conversations, feel refreshed after spending time alone, and enjoy in-depth, meaningful dialogues, Extroverts enjoy frequent social interactions, are goal-oriented, and experience a boost in energy after interacting with others. All of us display some extraversion or introversion, although most individuals lean more toward one or the other overall.

## Sensing (S) – Intuition (N)- Individuals spend time feeling and intuiting depending on the environment, much like extraverts and introverts do. While individuals who value intuition pay attention to perceptions and patterns, those who value sensing concentrate on reality and get knowledge from their senses. Those who favor sensing concentrate on facts and practical experiences, whereas they like to imagine possibilities and abstract ideas.

## Thinking(T) – Intuition(N) - Reasoning people are more likely to regard facts and unbiased information. When making a choice, they are usually dispassionate, rational, and consistent. When making decisions, those who value feeling are more inclined to take other people and their feelings into account.

## Judging (J) – Perceiving (P)- Judging-inclined people like structure and decisive outcomes. Individuals with a tendency for perception are more receptive, adaptive, and flexible. The other scales are impacted by these two inclinations.

With a focus on employing a LightGBM since it performs better than the other three models—Random forest, Multinomial Naive Bayes, and Logistic Regression—this study examines the possible uses of machine learning to personality prediction models. The model uses a collection of retrieved relevant parameters to predict personality traits. The approach consists of four steps: data preparation, feature selection, model training, and assessment. To translate the categorical personality labels into a machine learning algorithm-understandable format, we employ techniques such as label encoding. In addition, we separated the dataset into training and testing sets to guarantee the model's performance for generalization. 16 Personalities based on these factors (I E S N T F J P)[11] are shown in Table 1.

1. 16 Personality Traits

|  |  |  |  |
| --- | --- | --- | --- |
| **S.N.** | **TYPE** | **ABBR.** | **DESCRIPTION** |
| 1. | ISTJ | The Inspector | tactful and sensible, devoted, tidy, and conventional. |
| 2. | ISTP | The Crafter | Extremely autonomous and eager to learn firsthand from different experiences. |
| 3. | ISFJ | The Protector | kind and committed, ever prepared to defend those they love. |
| 4. | ISFP | The Artist | calm and adaptable, taciturn and creative. |
| 5. | INFJ | The Advocate | Analytical and creative, one of the rarest kinds. |
| 6. | INFP | The Mediator | High-minded and idealistic, they work to improve the world. |
| 7. | INTJ | The Architect | extremely rational, imaginative, and analytical. |
| 8. | INTP | The Thinker | renowned for having a vibrant inner life; quiet and introverted. |
| 9. | ESTP | The Persuader | extroverted, theatrical, and appreciative of the moment. |
| 10. | ESTJ | The Director | tenacious, rule-abiding, idealistic, and prone to taking the lead. |
| 11. | ESFP | The performer | gregarious, impulsive, and loves the limelight. |
| 12. | ESFJ | The Caregiver | extroverted and kind, with a propensity to see the best in people. |
| 13. | ENFP | The Champion | vivacious and captivating, flourishes in imaginative environments. |
| 14. | ENFJ | The Giver | Sensitive and devoted, well-known for compassion and giving. |
| 15. | ENTP | The Debater | Extremely creative, enjoys being surrounded by ideas, yet sometimes find it difficult to complete tasks. |
| 16. | ENTJ | The Commander | Outgoing, self-assured, and exceptionally skilled at planning and project management. |

# LITERATURE REVIEW

A number of researchers have worked in the field of personality prediction using various Machine Learning techniques. A comparative analysis of some of the works is described in Table 2.

1. Comparative Anaysis of state of art techniques

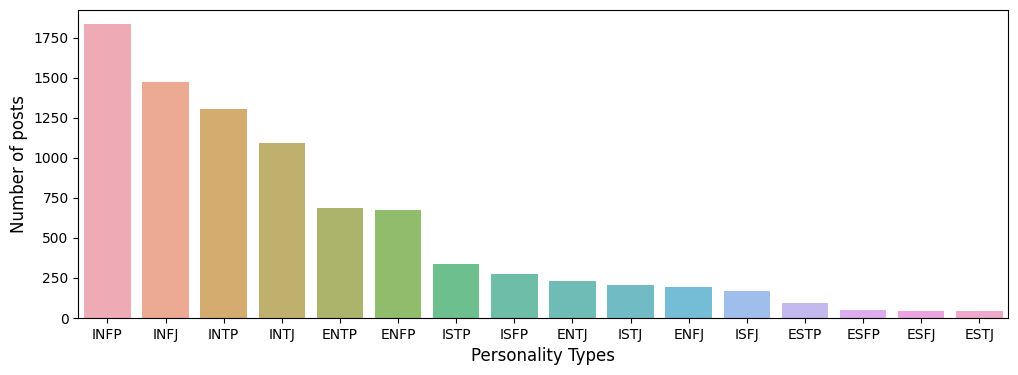
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ref** | **Goal** | **Technique/**  **Algorithm** | **Description** | **Results** |
| [1] | Personality prediction using social media applications. | Pearson Correlation | A novel Binary-Partitioning Transformer (BPT) equipped with Term Frequency & Inverse Gravity Moment (TF-IGM) is employed in this work to forecast social media applications' effects on human behavior and personality. | F1- score: 0.762  Accuracy: 78.43% |
| [2] | Personality traits for detection improvement | * Decision Tree * Naïve Bayes * Random Forest * SVM * Neural Networks | In order to enhance driver assistance systems, the study use machine learning to evaluate physiological data from healthy volunteers in hazardous urban environments. | 1.Agreeableness:  Accuracy:92.3%  F1-score: 91.8%  2.Conscientiousness:  Accuracy:90%  F1-score:89.2%  3.Extraversion:  Accuracy:90%  F1-score:89.8%  4.Neuroticism:  Accuracy:89.5%  F1-score:89.5%  5.Openness:  Accuracy:89.8%  F-score: 89.8% |
| [3] | Personality Prediction through Handwriting with Big Five model architecture. | * SVM * KNN * Decision Tree | The work predict the personality of the person using their handwriting and various Machine learning models are used here. | Accuracy:  1.SVM- above 90%  2.KNN- 88.6%  3.Decision Tree- above 99% |
| [4] | Text based Personality Prediction. | Multi model deep learning architecture and multiple pre-trained language models. | With 3.8 billion active social media users, personal recommendation systems, online marketing, and recruitment can all benefit from this system. | Mean Accuracy: 77.34%  Mean F1-score: 0.749 |
| [5] | Personality Prediction | XG-Boost | This study examines the prediction of personality traits among Facebook users by utilizing social network architecture and Big 5 model elements. | Accuracy: 74.2% |
| [6] | Product recommendation using personality prediction. | * Interest Mining * Item mapping * Discover Meta Paths * Recommend | Using interest mining and meta-path discovery, Meta-Interest is a personality-aware product recommendation system that enhances accuracy and recall in cold-start scenarios. | 1.Meta Interest:  Precision- 0.854  Recall- 0.868  2.DGRec:  Precision- 0.845  Recall- 0.855 |
| [7] | Personality Prediction for electronic marketplace. | Generalized Linear model | In order to forecast user personality and determine product preferences based on personality, this article suggests a methodology for analyzing social media data. | Mean Accuracy: 66.6%  Mean Precision: 0.670 |
| [8] | Personality assessment in psychology. | KGrAt-Net architecture. | This work investigates the use of KGrAt-Net, to the analysis of text-based Automatic Personality Prediction (APP) performance across domains, with an emphasis on knowledge graph representation. | 1.Knowledge Graph Attention Network:  Precision- 64.82%  Recall-81.44%  Accuracy-70.26%  2.Knowledge Graph Attention Network with EmbeddingsPrecision- 69.27%  Recall-80.58%  Accuracy-72.41% |

# DATA SET

In the proposed work a dataset of 8000 rows is used which represents an individual. The individuals are categorized into one of 16 distinct personality types. Every row has a number of important characteristics, such as:

* Personality Type: It provides the particular personality type of an individual.
* Description of Social Media postings: On the basis of the post uploaded on any social networking site such as facebook, twitter, blog post, a summary is generate to identify the personality type of a person.
* Timestamp: The time and date the social media post was created is shown by the timestamp.
* Engagement metrics: Metrics that show how popular and influential a piece of content is. They include things like likes, shares, comments, and responses for each article.
* Text Sentiment: Sentiment analysis scores that represent the negative, positive or neutral nature of the material.

The plotted Graph of data set is shown in Fig1.



1. Graph of Dataset

# implementation

A number of machine learning models as explained below are implemented to predict the personality of the individuals.

## Multinomial Naïve Bayes: Naive Bayes classifier is based on Bayes theorem [9]. Here A number of personality types have low precision, recall, and F1-scores; the accuracy on the training dataset is also rather low. This implies that the model might not be operating optimally with the available data. The classification results are shown in Figure 2. The X-Asis shows the 16 persoinality traits. The Figure shows that some classes had 0 accuracy, recall, or F1-score, suggesting that the model failed to classify examples of those classes properly. The model appears to have trouble making generalizations about certain personality types.

1. Classification using Multinomial Naïve Bayes

## Random Forest Classifier: Random forest is a computationally efficient method of ensemble learning, categorized under homogeneous base learner category. Its structure offers two advantages: computationally, it can handle regression and classification issues at high speed, making it a fast classic classifier. Statistically, it has features like feature prioritization, weight attribution, and unsupervised learning ability, making it suitable for high-dimensional issues [10]. The classification results are shown in Figure 3. Compared to the Multinomial Naive Bayes model, the accuracy on the training dataset is greater; nonetheless, performance problems persist, particularly with respect to specific personality types. For several classes, the recall, F1-score, and accuracy are still quite poor.

1. Classification using random forest

## LightGBM Classifier: LightGBM is a gradient-boosting framework that uses decision trees to lower memory utilization and improve model efficiency. These tactics constrain the primary histogram-based technique. The two GOSS and EFB approaches that are covered in the next section comprise the features of the LightGBM Algorithm. When combined, they give the model an advantage over rival GBDT frameworks and allow it to operate well.[7]. The classification results are shown in Figure 4. From Figure it can be observed that The LightGBM model demonstrates very high accuracy on the training dataset (approximately 98.8%). Each personality type's accuracy, recall, and F1-score are included in the classification reports along with macro and weighted averages. In terms of accuracy, the Random Forest and Multinomial Naive Bayes models are surpassed by the LightGBM model. The results on the test dataset demonstrate a great capacity to generalize to new data.

1. Classification using LightGBM

## Logistic Regression: For classification problems, supervised machine learning algorithms like logistic regression are used to estimate the likelihood that an instance will belong to a certain class. It forecasts the result of a categorical dependent variable by estimating the probability for the specified class using a sigmoid function. A discrete or category value with probability values between 0 and 1 must be the result. Similar to linear regression, logistic regression predicts two maximum values, 0 and 1, using a logistic function with a "S" shape. This curve shows the probability of obesity or malignant cells in a mouse. Because it can provide probabilities and categorize newly discovered data using both continuous and discrete datasets, logistic regression is important [5]. The classification results are shown in Figure 5.

1. Classification using Logistic Regression

The precision score value for all 16 traits is shown in Table 3 and is plotted in Fig 6 using all the classifiers.

1. Precision score comaprison table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Naive Bayes** | **Random Forest** | **Logistic Regression** | **LightGBM** |
| **ENFJ** | 0.29 | 0.39 | 0.36 | 0.97 |
| **ENFP** | 0.63 | 0.58 | 0.77 | 0.99 |
| **ENTJ** | 0.23 | 0.25 | 0.53 | 0.99 |
| **ENTP** | 0.56 | 0.61 | 0.77 | 0.99 |
| **ESFJ** | 0.09 | 0 | 0.26 | 0.9 |
| **ESFP** | 0.09 | 0 | 0.16 | 0.94 |
| **ESTJ** | 0 | 0 | 0.25 | 0.96 |
| **ESTP** | 0.16 | 0.8 | 0.36 | 0.97 |
| **INFJ** | 0.75 | 0.65 | 0.86 | 1 |
| **INFP** | 0.71 | 0.63 | 0.89 | 0.98 |
| **INTJ** | 0.49 | 0.59 | 0.81 | 0.99 |
| **INTP** | 0.62 | 0.57 | 0.86 | 0.99 |
| **ISFJ** | 0.3 | 0.46 | 0.54 | 1 |
| **ISFP** | 0.21 | 0.47 | 0.65 | 0.99 |
| **ISTJ** | 0.31 | 0.55 | 0.58 | 1 |
| **ISTP** | 0.44 | 0.63 | 0.74 | 0.98 |

1. Comparative analysis of Precision Score

# result and discussion

The proposed work is assessed using a number of performance metrics, including Accuracy, Sensitivity, Specificity, Precision, Negative Predicted value, and F1 Score.

## Accuracy

This refers to the extent to which the system model is precise. This speaks to the degree of accuracy of the system model. It is, in general, the ratio of expected observations to total observations. The percentage of accurate forecasts to all predictions is known as accuracy.  
The accuracy is calculated using Eq1. The accuracy calculated are shown in Table 4.

(1)

1. Accuracy values

|  |  |  |
| --- | --- | --- |
| **Models Implemented** | **Training-Dataset Accuracy** | **Testing-Dataset Accuracy** |
| LightGBM | 99% | 98.88% |
| Logistic Regression | 75.77% | 57.88% |
| Random Forest | 30.15 % | 28.19% |
| Multinomial Naive Bayes | 25.04% | 22.97% |

From the table it can be concluded that The LightGBM and Logistic Regression models exhibit comparable accuracy levels on the training dataset. On the test dataset, the LightGBM and Logistic Regression models perform similarly, with same accuracy. In comparison to the other two models, the Multinomial Naive Bayes model performs less accurately on both the training and test datasets.

## Precision

How accurately a machine learning model predicts the future, is one indicator of the algorithm's efficacy. Precision is calculated by finding the values of True Positive and False Positive. The precision is calculated using Eq2. The accuracy calculated are shown in Table 5.

(2)

1. Precision values

|  |  |  |
| --- | --- | --- |
| **Models Implemented** | **Training-Dataset Precision** | **Testing-Dataset Precision** |
| LightGBM | 97.4% | 52% |
| Logistic Regression | 63% | 42% |
| Random Forest | 14 % | 15% |
| Multinomial Naive Bayes | 31% | 31% |

## F1-Score

An assessment statistic for machine learning called the F1 score quantifies the accuracy of a model. It integrates a model's accuracy and recall ratings. The F1-score is calculated using Eq3. The F1-score calculated are shown in Table 6.

(3)

1. F1-Score values

|  |  |  |
| --- | --- | --- |
| **Models Implemented** | **Training Dataset F1-score** | **Testing Dataset F1-score** |
| LightGBM | 97.4% | 63% |
| Logistic Regression | 61% | 42% |
| Random Forest | 12.5 % | 11.3% |
| Multinomial Naive Bayes | 34.5% | 27.5% |

## Recall

Recall is dependent on True positive value and False Negative values. Recall is used to identify how good the model is to find the positive results. The recall value is higher if the model predicts the high positive samples. The Recall is calculated using Eq4. The Recall calculated are shown in Table 7.

(4)

1. Recall VAlues

|  |  |  |
| --- | --- | --- |
| **Models Implemented** | **Training-Dataset Recall** | **Testing-Dataset Recall** |
| LightGBM | 96% | 74% |
| Logistic Regression | 56% | 51% |
| Random Forest | 20 % | 14% |
| Multinomial Naive Bayes | 16% | 14% |

The comparative analysis of all the techniques is shown in Table 8. The Table shows that the highest accuracy value (98.88%) achieved is using the LightGBM model.

1. comparative analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **LightGBM** | **Logistic Regression** | **Random Forest** | **Multinomial Naive Bayes** |
| Accuracy | 98.88% | 57.88% | 28.19% | 22.97% |
| Precision | 52% | 42% | 15% | 31% |
| F1-Score | 63% | 42% | 11.30% | 27.50% |
| Recall | 74% | 51% | 14% | 14% |

# Conclusion

In conclusion, this research delved into the promising realm of predicting personality through the application of diverse machine learning models. Through meticulous analysis and experimentation, we explored the efficacy of various algorithms in capturing and interpreting intricate facets of human personality. Our findings underscore the potential of machine learning in unraveling the complexities of individual traits, offering valuable insights into the realms of psychology and behavioral science. The results indicate that LightGBM machine learning models exhibit commendable accuracy in predicting personality traits, thereby showcasing their utility as predictive tools. However, the work can be extended on large dataset also. In future work can be done by designing hybrid machine learning models to predict the personality of individuals.

##### References

1. Murari Devakannan Kamalesh, Bharathi B, Personality prediction model for social media using machine learning Technique, Computers and Electrical Engineering, Volume 100,2022,107852,ISSN 0045-7906.
2. Morgane Evin, Antonio Hidalgo-Munoz, Adolphe James Béquet, Fabien Moreau, Helène Tattegrain, Catherine Berthelon, Alexandra Fort, Christophe Jallais, Personality trait prediction by machine learning using physiological data and driving behavior, Machine Learning with Applications, Volume 9, 2022, 100353, ISSN 2666-8270,
3. Samsuryadi, Rudi Kurniawan, Julian Supardi, Sukemi, Fatma Susilawati Mohamad, "A Framework for Determining the Big Five Personality Traits Using Machine Learning Classification through Graphology", Journal of Electrical and Computer Engineering, vol. 2023, Article ID 1249004, 15 pages, 2023. https://doi.org/10.1155/2023/1249004
4. Christian, H., Suhartono, D., Chowanda, A. et al. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. J Big Data 8, 68 (2021).
5. M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behaviour on the Facebook Social Media Platform," in IEEE Access, vol. 6, pp. 61959-61969, 2018, doi: 10.1109/ACCESS.2018.2876502.
6. S. Dhelim, H. Ning, N. Aung, R. Huang and J. Ma, "Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery," in IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 86-98, Feb. 2021, doi: 10.1109/TCSS.2020.3037040.
7. Buettner, R. Predicting user behavior in electronic markets based on personality-mining in large online social networks. Electron Markets 27, 247–265 (2017).
8. Ramezani, M., Feizi-Derakhshi, MR. & Balafar, MA. Text-based automatic personality prediction using KGrAt-Net: a knowledge graph attention network classifier. Sci Rep 12, 21453 (2022).
9. Z. Tan, Y. Zhang, C. Zhang, R. Huang, P. Lei and X. Duan, "Research on The Text Emotion of Multinomial Naïve Bayes Integration Algorithm," 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 2019, pp. 107-111, doi: 10.1109/IMCEC46724.2019.8984049.
10. Mohammad Savargiv, Behrooz Masoumi, Mohammad Reza Keyvanpour, "A New Random Forest Algorithm Based on Learning Automata", Computational Intelligence and Neuroscience, vol. 2021, Article ID 5572781, 19 pages, 2021.
11. Woods RA, Hill PB. Myers Brigg. [Updated 2022 Sep 18]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available