Preprints.org

Article

Classification of Personality Types Using Predictive Models

Sahil Fareed and Manzoor Hussain *

Article

Classification of Personality Types Using Predictive Models

Sahil Fareed [1] and Manzoor Hussain [2,*]

[1]   University of Sialkot, Sialkot, Pakistan

[2]   Indus University, Karachi, Pakistan

*   Correspondence: manzoor.hussain@indus.edu.pk

Abstract

Personality prediction is a remarkable field with applications in psychology, marketing, and human-computer interaction. This study used supervised machine learning techniques to predict a person's personality type. The goal of the research was to accurately predict personality types using machine learning algorithms. To achieve this, we applied Gradient Boosting, Decision Tree, Random Forest, and Naive Bayes classifiers. Among these machine learning models, the Random Forest algorithm demonstrated the highest accuracy in predicting personality types, achieving an impressive accuracy rate of 87.64%. This study highlights an effective machine learning approach for personality prediction and offers valuable insights for future research in this domain.

Keywords: personality; prediction; machine learning; random forest; RapidMiner

1. Introduction

Personality prediction is a very important concept across various fields of life, especially in areas such as hiring, self-improvement, mental health evaluation, and marketing [1]. Traditionally, personality has been evaluated using interviews, questionnaires, or psychometric tests. While these methods are useful and impactful, they are often slow and subjective [2,3].

In recent years, the introduction of new technologies has significantly improved personality prediction processes. Machine learning has made it possible to predict a person's personality using datasets collected from different social media platforms [28–32]].

This research focuses on personality type prediction using a machine learning approach with a measurable dataset obtained from Kaggle. This method is more efficient and reliable than traditional systems. The Kaggle dataset used in this study contains eight features: Age, Gender, Education, Introversion Score, Sensing, Thinking Score, Judging Score, and Interest [18,19]. These features serve as an input for predicting personality types using machine learning algorithms.

Previous studies have applied machine learning models [25–27] using datasets collected from platforms like Facebook, Twitter, and Instagram. Although those models performed efficiently, the Kaggle dataset used in this study achieved higher accuracy and provides a more comprehensive view of personality prediction.

In this study, we implemented several machine learning algorithms, including Gradient Boosting, Decision Tree, and Random Forest. Each model has its own strengths. The aim of this study is to fill the gap in existing systems [21–24] by evaluating the performance of these machine learning algorithms using metrics such as accuracy, precision, recall, and F1-score [20].

The structure of this research paper is as follows (Figure 1):

1.   Section 1: Introduction, where we present the problem statement and proposed solution.
2.   Section 2: Literature Review, which discusses previous research in this area.

3. Section 3: Methodology, which explains the use of different algorithms to achieve better accuracy.
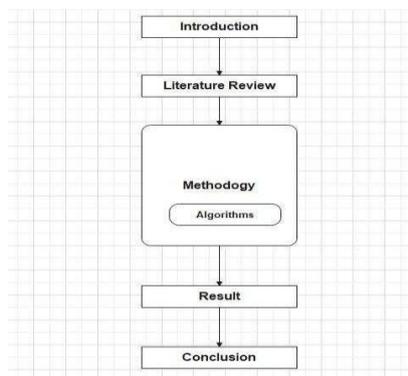
Figure 1. Research Paper Overview.

2. Literature Review

This paper reviews research on using social media data to predict users' personality types. Several studies have investigated the relationship between users' personalities and their interactions on platforms like Facebook and Twitter.

One study focused on Facebook users, examining how the amount of time spent on the platform, types of liked content, and news consumption patterns relate to personality traits. The XGBoost machine learning algorithm achieved the highest personality prediction accuracy of 74.2%. Among all traits, the highest prediction accuracy, 78.6%, was recorded for the Extraversion trait using individual SNA features [1].

Another study explored personality prediction using Twitter data, applying Linear Discriminant Analysis (LDA), Multinomial Naive Bayes, and AdaBoost to classify users based on the Big Five personality model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The models analyzed user tweets. Naive Bayes achieved the highest accuracy of 73.43% for the Openness trait, while LDA and AdaBoost showed the highest recall of 0.72 for the same trait [2].

A third paper evaluated the performance of GPT-3.5 and GPT-4 models (ChatGPT) in classifying personality types using data from PersonalityCafe, a forum focused on personality discussions. GPT-3.5 achieved 73% accuracy, while GPT-4 reached 76%. The study acknowledged a limitation due to the domain-specific nature of the dataset [3].

Another research study combined K-Means Clustering and Gradient Boosting to predict personality traits from general online data [16–19]. The combined model achieved an average accuracy of 86.3%. Techniques such as TF-IDF, lemmatization, and hyperparameter tuning were used to improve performance [4].

Research has also explored personality prediction using smartphone usage data. A study collected data from 624 volunteers over 30 days, analyzing patterns in communication, music listening [13–15], app usage, mobility, and time-of-day activity. Two models Elastic Net and Random Forest were tested, with Random Forest outperforming Elastic Net. The average prediction accuracy was 40%, indicating the potential of mobile data but also its limitations compared to social media platforms [5].

Further studies categorize three generations of personality prediction research using machine learning:
a) Early MLPA studies focused on digital footprints like social media activity.
b) Optimizing predictive validity used large samples to improve model accuracy.
c) Comparison to traditional assessments examined whether machine learning models could outperform self-reports and peer reports. This involved three key stages: data collection, data extraction, and personality prediction [6].

One paper built upon the "Big Five Experiment" by the Online Privacy Foundation, which used basic statistics to identify links between Facebook activity and personality traits. A more recent study improved on this with advanced data mining and machine learning techniques. Using a dataset of 537 Facebook users who completed a 45-question Big Five survey, researchers achieved 75% accuracy in predicting traits among the top 10% most Open individuals and 64% accuracy for the bottom 5% most Extraverted. However, the small sample size limited generalization [1].

Another study collected data from various platforms such as Twitter and Facebook, labeling it using IBM Personality Insights. Multiple machine learning and deep learning algorithms were tested, including kNN, J48, Random Forest, SVM, Naive Bayes, K-Means, Agglomerative Clustering, Decision Trees, and Simple Logistic Regression. These methods showed varying success in personality prediction [2].

SVM were particularly effective for predicting Conscientiousness, achieving 37.1% for Perceptual Curiosity, while Random Forests were used for broader classification tasks [8]. Some studies go beyond digital behavior. One paper focused on physiological responses, specifically eye movements, to predict personality traits. Using the Naive Bayes algorithm, it achieved 85.1% accuracy on a small dataset [10].

Another approach applied NLP techniques to improve personality prediction based on the Myers-Briggs Type Indicator (MBTI). The study used a Kaggle dataset with 8,675 entries containing MBTI types and social media posts to train models [11].

A broader study collected personality data from diverse sources, including government agencies, e-commerce websites, psychometric testing platforms, and marriage websites, integrating various machine learning models for improved prediction. NLP techniques were also applied [12].

3. Proposed Methodology

Personality prediction plays an important role across various domains such as recruitment, education, mental health, and marketing. Traditionally, personality has been assessed through interviews, questionnaires, and CV evaluations. However, advancements in technology have revolutionized this process. Today, personality can be predicted using data collected from social media platforms such as Facebook, Twitter, and Instagram.

Data Description

In this study, we used a publicly available dataset from Kaggle, chosen for its recency and completeness. The dataset contains 8 features: Age, Gender, Education, Introversion Score, Sensing, Thinking Score, Judging Score, and Interest. These features serve as predictors for identifying personality types, with the target variable being the personality type (e.g., ENTP, INTP, ESFP, ENFJ, ISFP, ESTJ, INFP, ISFJ, ESTP, and ENFP).

The dataset was imported into the RapidMiner tool, a widely used machine learning platform that supports data preprocessing, model training, and evaluation. For this study, we primarily used the SVM algorithm, which achieved an accuracy of 85%. However, it is worth noting that the dataset has limitations due to its size, which may not fully reflect the diversity of personality traits in the general population [7].

A related study [10] predicted personality traits using eye movement data collected in a controlled laboratory setting. That dataset consisted of 48 participants (42 females and 8 males). Using RapidMiner, the study achieved F1-scores of 40.3 for Neuroticism, 48.6 for Extraversion, 45.9 for Agreeableness, and 43.1 for other traits. Table 1 shows selected features.

Table 1. prediction of personality.

| Feature	Description
Age	Age of the person
Gender	Gender of the person (e.g., Male, Female)

Education	Indicates whether the person is educated or not
Introversion	Level of activity or engagement on social media
Sensing Score	Measures how practical or detail-oriented a person is
Thinking Score	Reflects the person's logical reasoning ability
Judging Score	Indicates preference for planning and organization
Interest	Person's areas of interest or hobbies

4. Result

We evaluated the performance of four machine learning algorithms: Gradient Boosting, Decision Tree, Random Forest, and Naive Bayes. As described earlier, the dataset was split into 70% for training and 30% for testing. Each algorithm was applied individually. Among them, Random Forest achieved the highest prediction accuracy. This algorithm is particularly effective [33] for handling complex and noisy data. Decision Tree performed the worst, likely due to the dataset's limitations. Table 2 and Figure 2 shows final results

Table 2. Comparison of Accuracy.

Algorithms	Accuracy
Gradient boosting	87.41%
Decision Tree	87.55%
Random Forest	87.64%
Naive Bayes	72.32%
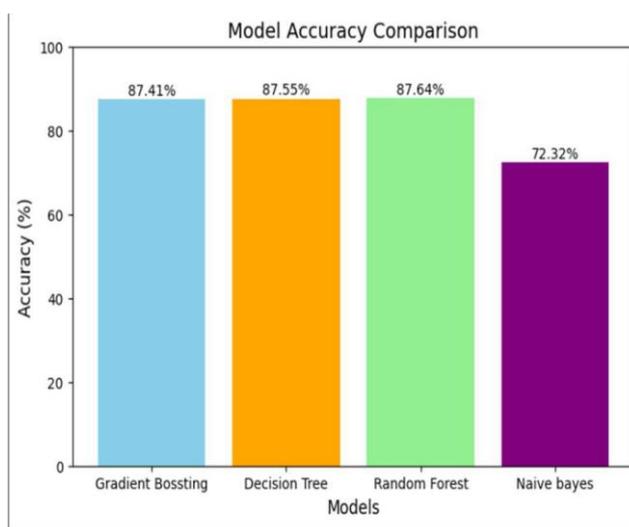
Figure 2. Model Accuracy Comparison.

The benchmark studies predicts the personality types of a person. This all model predicts the personality of a person. This data set has 8 features on the base of this feature the personality predict.

Table 3. Literature Review Accuracy Comparison.

Paper	Accuracy
[1]	74.2-78.6%
[2]	73.43%
[3]	73-76%
[7]	64%
[9]	85%

[12]	85%

5. Conclusions

This research paper explores personality prediction using various machine learning algorithms. The dataset used for this purpose was obtained from Kaggle and contains eight impactful features that contribute significantly to predicting an individual's personality. The paper analyzes which model is most suitable for this task. In the future, results could be improved by adjusting model parameters and focusing on personalized features within the dataset. Applying advanced feature engineering techniques may further enhance prediction accuracy.

References

1. Kale, P. S. R., Pawar, S. R., Behare, T. M., Hingwe, S. H., & Khuje, S. A. (2024). Issue 2. JETIR. Retrieved from http://www.jetir.org

2. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the Facebook social media platform. IEEE Access, 6, 61959–61969. https://doi.org/10.1109/ACCESS.2018.2876502

3. Kosan, M. A., Karacan, H., & Urgen, B. A. (2022). Predicting personality traits with semantic structures and LSTM-based neural networks. Alexandria Engineering Journal, 61(10), 8007–8025. https://doi.org/10.1016/j.aej.2022.01.050

4. Murphy, M. (2024). Artificial intelligence and personality: Large language models' ability to predict personality type. Emerging Media. https://doi.org/10.1177/27523543241257291

5. Mushtaq, Z., Ashraf, S., & Sabahat, N. (2020). Predicting MBTI personality type with K-means clustering and gradient boosting. In 2020 23rd IEEE International Multi-Topic Conference (INMIC). https://doi.org/10.1109/INMIC50486.2020.9318078

6. Stachl, C., et al. (2020). Personality research and assessment in the era of machine learning. European Journal of Personality, 34(5), 613–631. https://doi.org/10.1002/per.2257

7. Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. Personality and Social Psychology Review, 23(2), 190–203. https://doi.org/10.1177/1088868318772990

8. Sai Abhishak, I., & Vashisht, S. (2024). A study was conducted to predict personality traits using machine learning techniques on a dataset obtained from social media. Retrieved from https://ssrn.com/abstract=4833913

9. Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. Frontiers in Human Neuroscience, 12. https://doi.org/10.3389/fnhum.2018.00105

10. Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., & Eetemadi, S. (2020). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In Proceedings of the IEEE International Conference on Data Mining (ICDM), 1184–1189. https://doi.org/10.1109/ICDM50108.2020.00146

11. Berkovsky, S., et al. (2019). Detecting personality traits using eye-tracking data. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI). https://doi.org/10.1145/3290605.3300451

12. Agarwal, D., & Karthikeyan, M. M. (2023). Personality prediction using machine learning. Retrieved from http://www.irjmets.com

13. Chincholkar, A., Bhosale, D., Adsul, S., Bodkhe, A., & Kadam, R. (2023). A comprehensive survey on personality prediction using machine learning techniques. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 12(11). https://doi.org/10.17148/ijarcce.2023.121120

14. Rehman, A. U., et al. (2024). A machine learning--based framework for accurate and early diagnosis of liver diseases: A comprehensive study on feature selection, data imbalance, and algorithmic performance. International Journal of Intelligent Systems, 2024(1). https://doi.org/10.1155/2024/6111312

15. Mir, A., et al. (2024). A novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques. ESC Heart Failure. https://doi.org/10.1002/ehf2.14942

16. Azeem, M., Ullah, A., Ashraf, H., Jhanjhi, N. Z., Humayun, M., Aljahdali, S., & Tabbakh, T. A. (2021). Fog-oriented secure and lightweight data aggregation in iomt. IEEE Access, 9, 111072-111082.

17. Ahmed, Q. W., Garg, S., Rai, A., Ramachandran, M., Jhanjhi, N. Z., Masud, M., & Baz, M. (2022). Ai-based resource allocation techniques in wireless sensor internet of things networks in energy efficiency with data optimization. Electronics, 11(13), 2071.

18. Khan, N. A., Jhanjhi, N. Z., Brohi, S. N., Almazroi, A. A., & Almazroi, A. A. (2022). A secure communication protocol for unmanned aerial vehicles. CMC-Computers Materials & Continua, 70(1), 601-618.

19. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing.

20. Jabeen, T., Jabeen, I., Ashraf, H., Jhanjhi, N. Z., Yassine, A., & Hossain, M. S. (2023). An intelligent healthcare system using IoT in wireless sensor network. Sensors, 23(11), 5055.

21. Shah, I. A., Jhanjhi, N. Z., & Laraib, A. (2023). Cybersecurity and blockchain usage in contemporary business. In Handbook of Research on Cybersecurity Issues and Challenges for Business and FinTech Applications (pp. 49-64). IGI Global.

22. Shahnazari, K., & Ayyoubzadeh, S. M. (2025). Who Are You Behind the Screen? Implicit MBTI and Gender Detection Using Artificial Intelligence. *arXiv preprint arXiv:2503.09853*.

23. Hanif, M., Ashraf, H., Jalil, Z., Jhanjhi, N. Z., Humayun, M., Saeed, S., & Almuhaideb, A. M. (2022). AI-based wormhole attack detection techniques in wireless sensor networks. Electronics, 11(15), 2324.

24. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in industry 4.0. In Cyber Security Applications for Industry 4.0 (pp. 97-109). Chapman and Hall/CRC.

25. Naz, A., Khan, H. U., Bukhari, A., Alshemaimri, B., Daud, A., & Ramzan, M. (2025). Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. *Artificial Intelligence Review*, *58*(8), 239.

26. Stracqualursi, L., & Agati, P. (2025). Predicting MBTI personality of YouTube users. *Scientific Reports*, *15*(1), 7221.

27. Humayun, M., Almufareh, M. F., & Jhanjhi, N. Z. (2022). Autonomous traffic system for emergency vehicles. Electronics, 11(4), 510.

28. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 305-310). IEEE.

29. Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2023). Key applications of state-of-the-art technologies to mitigate and eliminate COVID-19. Authorea Preprints.

30. Kiel, L., Lind, M., Bo, S., Jørgensen, C. R., Bøye, R., Frederiksen, C. K., & Spindler, H. (2025). Associations between pathological personality traits, functional impairment, and personality disorder: Controlling for basic personality traits and identity disturbance. *Personality Disorders: Theory, Research, and Treatment*.

31. Khalil, M. I., Humayun, M., Jhanjhi, N. Z., Talib, M. N., & Tabbakh, T. A. (2021). Multi-class segmentation of organ at risk from abdominal ct images: A deep learning approach. In Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021 (pp. 425-434). Singapore: Springer Nature Singapore.

32. Humayun, M., Jhanjhi, N. Z., Niazi, M., Amsaad, F., & Masood, I. (2022). Securing drug distribution systems from tampering using blockchain. Electronics, 11(8), 1195.

33. Imran, N., Zhang, J., Yang, Z., & Ali, J. (2025). mm-FERP: An effective method for human personality prediction via mm-wave radar using facial sensing. *Information Processing & Management*, *62*(1), 103919.