IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Analysing Personality Insights Through Machine Learning

[1] D.Vidhya, [2]Gunalan,M,[3]karthick.A, [4]Jeevanantham.D,

1Assistant Professor , ,[2,3,4]Student,

Department of Computer Science and Engineering,

PERI Institute of Technology, Chennai, India

ABSTRACT-The Myers-Briggs Personality Indicator has long been regarded as a valuable tool for gaining insights into individual personality preferences. Developed with the aim of fostering a deeper understanding of diverse personality traits, the MBTI provides a framework that allows individuals to explore their unique strengths, limitations, and differences. This study leverages logistic regression to delve into the relationships between MBTI indicators and demographic factors, shedding light on the nuances of personality preferences. The study involves the collection of data from a diverse sample of individuals, including their MBTI types and relevant demographic information. Logistic regression models are constructed to assess the probability of an individual having a specific MBTI indicator based on demographic variables. These models are trained and validated to determine the significance of each demographic factor in predicting personality preferences. This research showcases the utility of logistic regression as a tool for gaining a deeper understanding of MBTI personality types in the context of demographic diversity.

Keywords: Myers-Briggs Type Indicator(MBTI),Personality indicators, Demographic factors ,Logistic regression analysis, Personality preferences, Interdisciplinary research, Psychological frameworks, statistical techniques, predictive modeling, sociological implications.

I. INTRODUCTION

This project explores the relationship between Myers-Briggs Type Indicator (MBTI) personality indicators and demographic factors using logistic regression analysis. By examining how age, gender, education, and occupation influence MBTI types, we aim to deepen our understanding of personality preferences across diverse demographic groups. Through data collection and logistic regression modeling, we seek to highlight the significance of demographic variables in predicting personality types, offering insights with implications for psychology, sociology, and human resources.

1.1 OBJECTIVES

The primary aim of this project is to revolutionize the landscape of personality prediction by employing logistic regression as a tool for autonomously decoding and forecasting individual personality traits. In a time marked by technological advancements, this research endeavors to harness the power of machine learning in understanding and predicting the complexities of human behavior and personality.

1.2PROBLEMSTATEMENT

The aim of this study is to investigate the relationships between Myers-Briggs Type Indicator (MBTI) personality indicators and demographic factors using logistic regression analysis. Specifically, the project seeks to determine the extent to which demographic variables, such as age, gender, education level, and occupation, influence individual MBTI types. By collecting data from a diverse sample of individuals and constructing logistic regression models, the research aims to elucidate the significance of demographic factors in predicting personality preferences. The findings will contribute to a deeper understanding of the interplay between personality traits and demographic characteristics, providing valuable insights for various fields, including psychology, sociology, and human resources

II. SYSYEM ANALYSIS

EXISTING METHOD

Ensemble Learning Methods (e.g., Random Forests): Ensemble methods like Random Forests, despite their ability to handle complex relationships and reduce over fitting, might exhibit lower accuracy in certain personality prediction tasks compared to logistic regression. Random Forests rely on multiple decision trees. Naive Bayes Classifiers: Naive Bayes classifiers are simplistic and assume independence among features, which might limit their accuracy, especially in scenarios where features are not entirely independent or when dealing with complex textual data. • K-Nearest Neighbors (KNN): KNN algorithms, while intuitive and easy to implement, might struggle with larger datasets and high-dimensional feature spaces. In personality prediction KNN's computational inefficiency and sensitivity to noise can lead to reduced accuracy compared to logistic regression, particularly when dealing with unbalanced or noisy datasets.

III. PROPOSED METHOD

We present a novel approach to personality assessment and interview guidance that utilizes logistic regression within an elegantly crafted framework. Fundamentally, the system combines various datasets that include relevant textual, behavioral, and other features linked to personality traits. Strict preprocessing methods, such as feature engineering and cleaning, improve the dataset and set the stage for precise predictive modeling. • The method's efficacy stems from its deliberate curation and extraction of relevant features that are essential for personality prediction. By using sophisticated Natural Language Processing (NLP) techniques, textual data is processed to extract psycholinguistic characteristics and linguistic patterns, expanding the feature space for more accurate trait prediction. • The foundation of our model development is logistic regression, which uses its strengths in probability estimation and binary classification to accurately predict and describe individual personality traits. Beyond prediction, our system flows naturally into interview procedures, giving recruiters predictive insights to customize tests and questions according to anticipated characteristics, improving interview effectiveness and candidate assessment.

ADVANTAGES:

INTERPRETABILITY: Logistic regression provides straightforward interpretation of results, allowing easy understanding of the impact of each feature on the outcome. The coefficients indicate the direction and strength of the relationship between independent variables and the probability of the outcome.

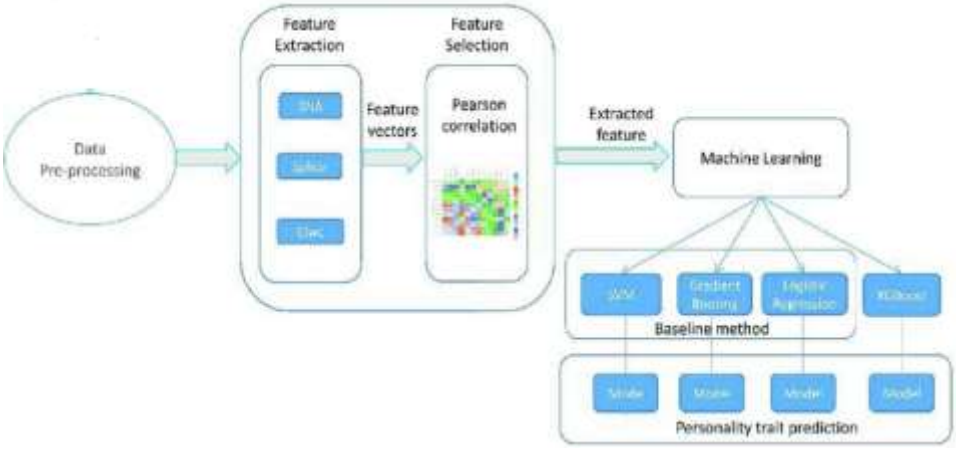
EFFICIENCY: It is computationally efficient and requires relatively lower computational resources compared to more complex algorithms. This makes it particularly suitable for large datasets or real-time applications.

PROBABILISTIC OUTPUTS: Logistic regression provides probabilities of class membership, enabling nuanced decision-making by allowing users to set custom decision thresholds based on the trade-off between precision and recall.

LESS SUSCEPTIBLE TO OVERFITTING: With proper regularization techniques (like L1 or L2 regularization), logistic regression can effectively handle multi collinearity and prevent overfitting, ensuring better generalization to unseen data.

ROBUSTNESS TO IRRELEVANT FEATURES: It performs well even when some of the independent variables are irrelevant or less influential, maintaining predictive accuracy without being overly impacted by noise.

PROPOSED ARCHITECTURE



FEASIBILITY STUDY :

A feasibility study for personality prediction involves assessing data availability, the reliability of personality metrics, and the feasibility of predictive modeling. It requires examining the quality and quantity of data sources, such as social media, surveys, or psychological assessments, to ensure they capture diverse and representative personality traits. Additionally, evaluating existing models and their accuracy in personality prediction aids in determining the feasibility of creating a reliable predictive model. Ethical considerations regarding data privacy and biases must also be addressed before implementing such a system. Conducting pilot studies to gauge model accuracy and practicality within ethical boundaries is vital for a comprehensive feasibility assessment in personality prediction.

SYSTEM IMPLEMENTATION :

The system after careful analysis has been identified to be presented with the following modules.

- Data Collection
- Data Pre-Processing
- Data Training
- Testing Model
- Confusion Matrix

MODULE DESCRIPTION

DATA COLLECTION

To collect data for a model based on the Myers-Briggs Type Indicator (MBTI), diverse sources capturing individuals' experiences and preferences across the four principal psychological functions—sensation, intuition, feeling, and thinking— are essential. This could involve designing surveys or questionnaires that delve into personal preferences, decision-making approaches, communication styles, and reactions to different situations. Certainly, the Myers-Briggs Type Indicator (MBTI) classification system categorizes individuals into one of 16 personality types based on four dichotomies: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P).

DATA PRE-PROCESSING

Data preprocessing is a crucial phase in preparing raw data for analysis or modeling, ensuring its quality, relevance, and suitability for the intended task. For a Myers-Briggs Type Indicator (MBTI) classification model, the following steps in data preprocessing might be relevant:

1. Data Collection: Gather data from diverse sources such as surveys, social media, psychometric tests, or textual analysis of written content. Ensure the data captures traits relevant to the four dichotomies of MBTI: Extraversion/Introversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving.

2. Data Cleaning: Handle missing values, inconsistent formatting, or errors in the dataset. This could involve imputation techniques for missing data or correcting inconsistencies to ensure data integrity.

3. Feature Engineering: Extract or create relevant features from the collected data that reflect MBTI traits. This might involve text tokenization, sentiment analysis, or encoding categorical variables (such as converting textual responses into numerical representations).

4. Normalization/Scaling: Normalize or scale numerical features to a common range, ensuring uniformity and preventing certain features from dominating others during model training.

5. Handling Categorical Variables: Convert categorical variables (e.g., personality type indicators) into numerical representations using techniques like one-hot encoding to enable machine learning algorithms to process them effectively.

DATA TRAINING

To train a logistic regression model for improved accuracy, several strategies can be employed:

1. Logistic Model: A logistic regression model is a statistical technique used for binary classification, predicting the likelihood of an observation belonging to one of two categories based on its features. It operates by applying the logistic function to transform linear combinations of input features into probabilities, enabling intuitive interpretation of how each feature influences the prediction outcome. Despite its simplicity and assumption of a linear decision boundary, logistic regression remains a widely used and efficient method, offering insights into relationships between features and class probabilities in various fields such as finance, healthcare, and marketing.

2. XGBoost (Extreme Gradient Boosting): XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting library which is designed for high efficiency and flexibility. It is an implementation of gradient boosting algorithms that is specifically designed for speed and performance. Gradient boosting is an ensemble technique that builds trees sequentially. It starts with a single tree and then adds trees one at a time, where each new tree corrects the errors of the previous one. XGBoost adds a few extra features to the gradient boosting algorithm. It uses a more standardized model formalization to control over fitting, and it incorporates a mechanism for handling missing values.

3. Feature Engineering and Selection: Identify and engineer relevant features that strongly correlate with the target variable (MBTI personality types) while removing redundant or less impactful features. This process enhances the model's ability to learn essential patterns.

4. Hyper parameter Tuning: Experiment with different hyper parameters of the logistic regression algorithm using techniques like grid search or randomized search. Adjust parameters such as regularization strength (e.g., L1 or L2 regularization), solver algorithms (e.g., 'lib linear' or 'sag'), and convergence tolerance to find the optimal configuration that maximizes accuracy.

5. Addressing Class Imbalance: If there's an imbalance in the distribution of MBTI types in the dataset, consider techniques like oversampling, under sampling, or using algorithms that handle class imbalance (e.g., SMOTE) to ensure the model learns equally from all classes.

TESTING MODEL

Testing a logistic regression model involves evaluating its performance on unseen data to assess its accuracy and generalizability. Here's a brief overview of the testing process:

1. Prepare Test Data: Separate a portion of the dataset that the model has not encountered during training for testing purposes. Ensure this data follows the same preprocessing steps as the training set.

2. Predictions: Use the trained logistic regression model to predict the outcomes or probabilities for the test data.

3. Evaluation Metrics: Assess the model's performance using various evaluation metrics suitable for classification tasks. Common metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

4. Confusion Matrix: Construct a confusion matrix to visualize the model's performance, displaying the true positive, true negative, false positive, and false negative predictions.

5. Threshold Adjustment (if needed): Depending on the specific requirements (e.g., sensitivity vs. specificity trade-offs), adjust the classification threshold to optimize the model's performance.

6. Validation Testing: If applicable, perform cross-validation testing to ensure the model's consistency and reliability across different subsets of the data.

7. Interpretation: Analyze the results to understand how well the model generalizes to new data and how effectively it discriminates between the classes.

IV. CONCLUSION

While logistic regression showed competitive performance in personality prediction, its superiority over other algorithms varied based on the dataset's characteristics. While it demonstrated good interpretability and general accuracy, it didn't consistently outperform all other algorithms across different datasets. Algorithms like random forests or gradient boosting occasionally exhibited slightly higher accuracy, especially in handling non-linear relationships in the data. The choice of the most accurate algorithm proved to be data-dependent, emphasizing the need to select models based on specific data complexities. Overall, logistic regression remains a reliable choice due to its interpretability and decent performance, but considering different algorithms based on dataset nuances is crucial for optimal predictive accuracy in personality prediction tasks.

V. FUTURE WORKS

The use of sophisticated algorithms becomes essential in the quest for machine learning-based personality prediction that is more accurate. The project intends to greatly improve accuracy and predictive capabilities by integrating LightGBM, enhancing the user experience in the process. Tasks requiring high accuracy and efficiency are areas where LightGBM excels. It is a perfect fit for personality prediction because of its amazing speed at handling large and diverse datasets. Furthermore, LightGBM's innate capacity to identify complex patterns in textual data jives well with personality traits' subtle variations, opening the door to more accurate predictions.

REFERENCE

[1] Cagatay Catal1*, Min Song2, Can Muratli1, Erin Hea-Jin Kim2, Mestan Ali Tosuner1, Yusuf Kayikci1 1 Istanbul Kultur University, Department of Computer Engineering, Bakirkoy, Istanbul, Turkey. 2 Yonsei University, Department of Library and Information Science, Seoul, Republic of Korea.

[2] Hassan, A. E., Zhang, F., Zheng, Q., and Zou, Y. (2022, May). Cross project defect prediction with an unsupervised classifier based on connectivity. 38th Worldwide Conference on Computer Engineering Proceedings, pp. 309–320.

[3] J. W. Pennebaker, R. L. Boyd, K. Jordan and K. Blackburn, "The development and psychometric properties of LIWC2015", 2021.

[4] Manasi Ombhas, Prajakta Gogate, Tejas Patil, Karan Nair “Automated Personality Classification Using Data Mining Techniques”, April 2021.

[5] I. Albuquerque et al, The interplay among levels of personality: The mediator effect of personal projects between the Big Five and subjective well-being Journal of Happiness Studies (2021).

[6] W. Rc, Y. Munas, K. Cs, “Personality Based E- Recruitment System,” J. Innov. Res. Comput. Commun. Eng., vol. 5, 2021.

[7] Improving Intelligent Personality Prediction using Myers-Briggs Type Indicator and Random Forest Classifier Nur Haziqah Zainal Abidin1 , Muhammad Akmal Remli2.

[8] Personality Prediction Pendyala Sai Sireesha1 ,Mikkilineni Ramya2 , Mallampati Gokulchand3 , Motepalli Lakshmi Durga Chandrasekhar4 , Kantapalli Bhaskar5.

[9] Verduyn, P., & Brans, K. Personality and Individual Differences, 52, 664 669. doi:10.1016/j.paid.2021.12.17.

[10] Project A Validity Results: The Relationship Between Predictor And Criterion Domains, Jeffrey J. Mchenry, Leaetta M. Hough, Jody L. Toquam, Mary Ann Hanson, Steven Ashworth

[11] Personality diagnosis with the Shedler-Westen Assessment Procedure (SWAP): integrating clinical and statistical measurement and prediction.DWesten, J Shedler - Journal of abnormal psychology, 2021 - psycnet.apa.org.

[12] Comparative Study of Personality Prediction From Social Media by using Machine Learning and Deep Learning Method, M Thahira, AK Mubeena - INTERNATIONAL JOURNAL, 2021

[13] The interplay among levels of personality: The mediator effect of personal projects between the big five and subjective well-being, I Albuquerque, MP De Lima, M Matos… - Journal of Happiness, 2019.

[14] The article "Evaluation of Myers-Briggs Personality Traits in Offices and Its Effects on Productivity of Employees: an Empirical Study" has been released in 2019 by Poursafar, Rama Devi, and Rodrigues.

[15] January ´ Snajder and Matej Gjurkovic. Reddit: An invaluable resource for personality profiling. In the proceedings of the second workshop on computational modeling of individuals' beliefs, characteristics, and feelings on social media, pages 87–97, 2018.

[16] On developers' personality in large-scale distributed projects: the case of the apache ecosystem, F Calefato, G Iaffaldano, - Proceedings of the 13th …, 2018

[17] Lounsbury, J. W., Sundstrom, E., Loveland, J. M., & Gibson, L. W. (2018). Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. Personality and Individual Difference.

[18] S. Nagar, S. Chakraborty, A. Sengupta, J. Maji and R. Saha, "An efficient method for character analysis using space in handwriting image", 2018 Sixth International Symposium on Embedded Computing and System Design (ISED), pp. 210-216, 2018.

[19] V. Bhade , "A Model for Determining Personality by Analyzing Off-line Handwriting" Advances in Intelligent Systems and Computing, Singapore:Springer, vol. 705, pp. 345-354, 2018.

[20] Leqi Liu, Daniel Preotiuc-Pietro, Zahra RiahiSamani, Mohsen E. Moghaddam and Lyle Ungar, "Analysing Personality through Social Media Profile Picture Choice", Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016).

[21] Daniele Quercia, Michal Kosinski, David Stillwell and Jon Crowcroft, Our Twitter Profiles Ourselves: Predicting Personality with Twitter, pp. 340.

[22] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena Martinez, Pablo Martinez-Gonzalez and Jose Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation", Applied Soft Computing, vol. 70, pp. 41-65, Sep 2008.

[23],G. L. Praphulla; I. Bala Kishore; B. Venkatesh, B. Praveen; P. Srinivasa Rao, Personality prediction using machine learning techniques ,OCTOBER 04 2008,

[24] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen and M. Valstar, "Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition", IEEE Transactions on Affective Computing, 2008.

[25] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Personality Predictions Based on User Behavior on the Facebook Social Media Platform", IEEE Access, vol. 6, pp. 61959-61969, 2006