JETIR.ORG
ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Digital Traces Reveal Characteristics: An Analysis of Automated Personality Classification Using Decision Trees and Information Mining

Er. Ashish P. Mohod[1] , Adarsh Golghate[2] , Amol Pimpalkar[3] , Pratik Gondane[4] , Shrineet Padade[5]

[1,2,3,4,5] Department of Computer Science and Engineering, P.J.L.C.E Nagpur

Abstract:

Using data mining techniques and decision tree algorithms, this study investigates the feasibility of automatically identifying people's personalities based on their digital footprints. Through an analysis of many sources, such as text messages, social media posts, and web browsing activities, our goal is to identify patterns and correlations that align with specific personality traits. Decision tree algorithms provide a solid foundation for comprehending these connections and enable the development of predictive models that accurately classify personality traits. Important components of the study methodology include feature selection, data collecting, preprocessing, model building, and evaluation. This study addresses a number of ethical issues, including consent, privacy, and fairness. The outcomes show the efficacy and use of automated personality classification across a range of industries, while also highlighting the importance of moral principles and openness controls in data-driven research. All in all, this research expands our understanding of human behaviour in the digital age and opens doors for customised treatments, targeted advertising campaigns, and accelerated recruiting processes.

Key words: Automatic Personality Classification, Decision Tree, Data Mining

1.Introduction:

In the era of big data and artificial intelligence, understanding human personality traits from digital footprints has garnered a lot of attention in a range of sectors, including psychology, marketing, human resources, and social sciences. Automatic Personality Classification utilising Decision Trees and Data Mining techniques is an innovative way to analyse and categorise people's personalities based on their textual correspondence, internet activities, and other digital traces.

A person's personality is defined by their enduring thought, feeling, and conduct patterns that characterise their unique psychological makeup. The concept of personality is intricate and multifaceted. In the past, the main techniques for assessing personality were interviews and self-report questionnaires. These techniques could be time-consuming, biassed, and subjective. However, because digital platforms have proliferated and data mining technologies have advanced, academics and practitioners now have access to vast amounts of digital data that can be used to automatically infer personality traits.

An effective basis for automated personality classification is provided by the popular machine learning technique known as decision trees. Recursively partitioning the data based on predictor variables, decision trees produce a hierarchical model that translates input features to output classes—personality traits in this example. In data mining, decision trees are used to analyse factors such as language patterns, social media interactions, browser behaviour, and demographic data in order to predict personality traits such as the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism).

In automatic personality categorization, data mining techniques are essential for feature selection, preprocessing, and model validation. In feature engineering, pertinent features are extracted from unprocessed data sources and formatted for analysis. While social network analysis approaches can reveal patterns in social interactions, natural language processing (NLP) methods are frequently used to extract linguistic elements from textual data.

Beyond scholarly research, automatic personality classification is applied in real-world fields like talent acquisition, recommendation systems, and personalised marketing. Businesses may better cater their goods, services, and content to the interests and requirements of their clientele by learning about the personality profiles of individuals. Furthermore, in the field of human resources, automated personality classification can expedite the hiring procedure by highlighting applicants whose character attributes complement job specifications and company culture.

Despite its potential benefits, computerised personality classification based on data mining and decision trees raises ethical questions about consent, privacy, and the exploitation of personal information. In the digital era, ensuring accountability, transparency, and ethical standards is crucial to reducing dangers and defending people's rights.

2.Literature Review:

2.1. A Study on Personality Prediction & Classification Using Data Mining Algorithms

There has been a rise in research on data mining algorithms for personality prediction and categorization, as shown by this work [1], which uses techniques like decision trees and SVMs. Two issues are the diversity of databases and privacy concerns. Applications include human resource and personalised recommendation systems. The evaluated study looks at how effectively various algorithms predict personality traits in an effort to contribute.

2.2. Personality Classification with Data Mining

The application of data mining techniques for personality classification is examined in this study [2]. Previous studies employed both traditional procedures and modern data-driven methodologies. There are difficulties because of the size of the dataset and moral dilemmas. Applications include human resources and customised guidance. The paper's goal is to conduct a thorough investigation into the efficient classification of personality traits using data mining methods.

2.3. Automated Personality Classification Using Data Mining Techniques

The study [3] investigates automated personality classification using data mining approaches. Previous studies have looked into a variety of personality classification methods, from traditional tests to modern data-driven approaches. Two examples of data mining methods that have evolved into practical instruments in this domain are neural networks and decision trees. Diverse datasets and moral dilemmas, particularly in relation to privacy, present challenges. Applications can be found in a wide range of fields, including human resources and personalised recommendation systems. Probably the major purpose of the research is to present a comprehensive review of automated personality classification utilising data mining methodologies to increase efficiency and accuracy.

2.4. Personality classification using Data mining approach

N-closest neighbourhood algorithm (NCN) data mining is being used to mechanise personality classification, a significant area in psychology. While earlier methods relied on random surveys, data mining offers a more efficient method. The NCN approach, which is well-known for its simplicity and effectiveness, predicts personality traits by utilising patterns discovered in huge datasets. By implementing and analysing NCN, this research [4] aims to simplify the personality classification process by utilising user input and existing data. All things considered, this work enhances the use of data mining techniques for automated personality prediction.

3.Problem Statement:

Massive volumes of data that can be used to automatically infer personality traits have been produced by the rapid development of digital communication platforms and the proliferation of online activities. However, despite its potential benefits, computerised personality classification utilising decision tree and data mining approaches still faces many challenges and open problems. As a result, the issue statement for this topic contains the following components:

i. **Data Acquisition and Preprocessing:** How can we efficiently collect and preprocess digital data from a range of sources, including social media posts, textual exchanges, and browser behaviour, to extract relevant attributes for personality classification?

ii. **Feature Engineering and Selection:** What qualities are most helpful in predicting personality traits, and how can these qualities be selected and developed effectively to enhance the performance of decision tree models?

iii. **Model Development and Evaluation:** In what ways might decision tree algorithms be improved and honed to better classify people's personalities? What measures and methods of evaluation are appropriate for determining how well personality classification models work and how applicable they are?

iv. **Ethical Aspects and Privacy Protection:** What are the ethical implications of automatic personality classification, particularly with regard to permission, privacy, and the potential misuse of personal data? What steps can we take to ensure transparency, equity, and accountability in the collection and use of digital data for personality inference?

v. **Application and Deployment:** How can real-world applications such as recommendation engines, tailored marketing, and talent acquisition best use autonomous personality categorization models? What are the potential benefits and drawbacks of applying these models to real-world scenarios?

4.Methodology:

Data collection entails obtaining information from research participants using a range of digital data sources, including text messages, posts on social media, online activity, and demographic data. Check for compliance with consent and data privacy ethical norms.

Data Preprocessing: Clean and preprocess the collected data to handle missing values, remove noise, and standardise formats. To prepare textual data for analysis, tokenization, text normalisation, and feature extraction should be carried out. Utilise techniques such as sentiment analysis, feature engineering, and part-of-speech tagging to extract relevant features from textual and behavioural data.

Model Development: Use decision tree approaches such as C4.5, Random Forests, and CART (Classification and Regression Trees) to develop models of personality classification. Use libraries like as scikit-learn or Weka for implementation. As you train the models on the pre-processed data, make use of techniques like cross-validation, tweak hyperparameters, and evaluate performance.

Evaluation measures: Assess the effectiveness of personality classification models using pertinent evaluation metrics, such as area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score. Run statistical tests to evaluate the performance and generalizability of different models.

Ethics: Consider ethical concerns throughout the entire research effort. These include obtaining participants' informed consent, safeguarding their privacy, and ensuring compliance with data protection regulations. Establish procedures for transparency and make sure participants are aware of how their data will be used.

Cross-validation: A helpful method for assessing personality classification models' performance on hypothetical data and ascertaining the models' robustness is cross-validation. Use techniques such as k-fold cross-validation or leave-one-out cross-validation to ensure precise estimations of model performance.

Interpretability: Look at the decision rules that decision tree models generate to find out more about the relationship between input features and predicted personality traits. Use importance rankings and decision trees to enhance interpretability and facilitate understanding for domain specialists.

4.1 Decision Tree as a Solution;

While decision trees can be useful in addressing certain issues related to automatic personality classification, they might not be sufficient to address all of the issues. However, there are a number of advantages that decision trees offer that can mitigate some of these problems:

i. **Interpretability and Explainability:** When decision rules are described in an understandable and transparent manner using decision trees, it becomes easier to understand how the model makes its predictions. This can help fix problems with the model's explainability and interpretability.

ii. **Feature Selection and Representation:** Feature selection is automatically performed by decision trees, which determine which attributes are most valuable for classification. By doing this, relevant features from digital footprints can be found and the dimensionality of the feature space can be reduced.

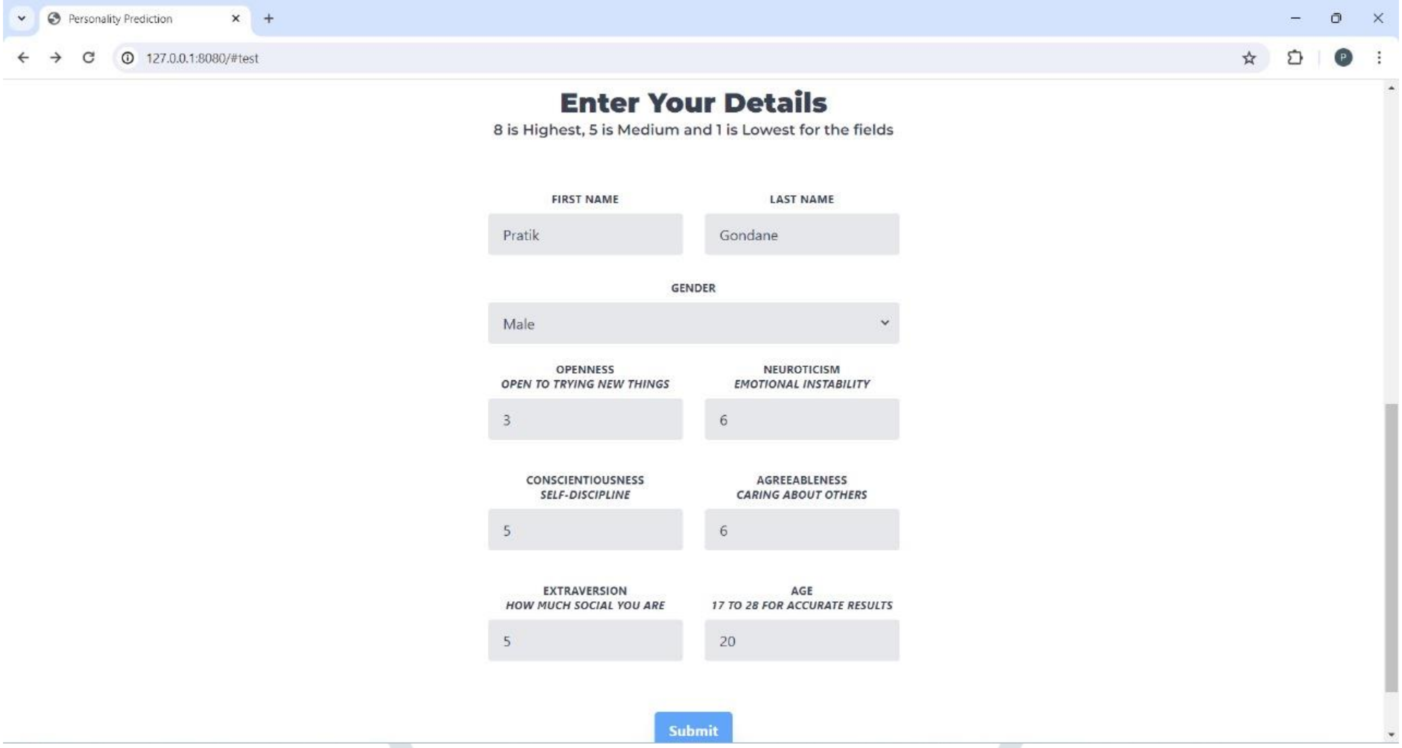
iii. **Handling Nonlinearity and difficult Relationships:** Because decision trees are able to record nonlinear relationships and interactions between characteristics, they can be used to describe complex decision boundaries in data. This flexibility can be applied to both the complexity of personality and the variety of digital data sources.

iv. **Scalability and Efficiency:** Because decision tree approaches are computationally efficient and scalable to large datasets, they can be used to assess a wide range of digital footprints and train classification models on enormous datasets.

v. **Robustness and Generalisation:** Reducing overfitting to the training set and improving generalisation are two benefits of pruning and regularising decision trees. This can ensure that classification models generalise well to yet-to-be-observed data and are robust across diverse populations and contexts.

4.2 Experiment and Result Discussion:

We assess this method by creating a web application personality test. The test-giver will only provide one response for each personality feature. The machine learning algorithm will categorise the personality based on the data.



Enter Your Details

8 is Highest, 5 is Medium and 1 is Lowest for the fields

FIRST NAME: Pratik

LAST NAME: Gondane

GENDER: Male

OPENNESS *OPEN TO TRYING NEW THINGS*: 3

NEUROTICISM *EMOTIONAL INSTABILITY*: 6

CONSCIENTIOUSNESS *SELF-DISCIPLINE*: 5

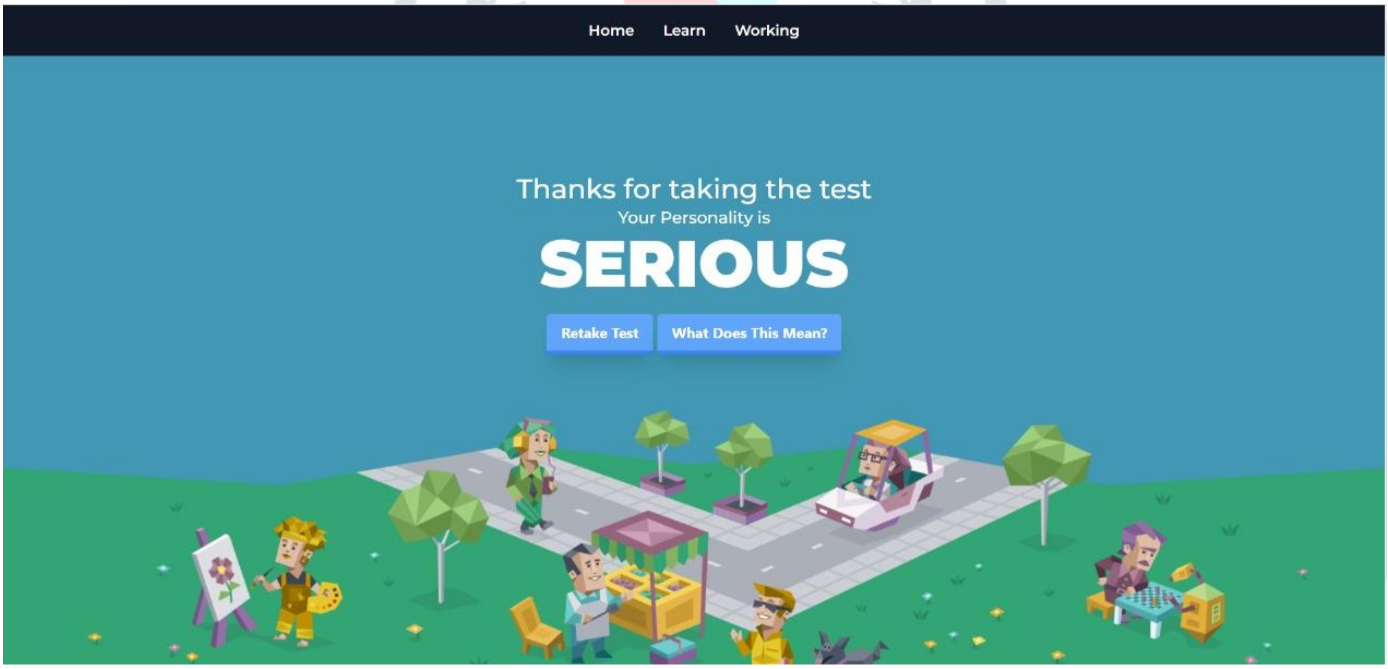
AGREEABLENESS *CARING ABOUT OTHERS*: 6

EXTRAVERSION *HOW MUCH SOCIAL YOU ARE*: 5

AGE *17 TO 28 FOR ACCURATE RESULTS*: 20

Submit

Fig. 4.2.1 Test Page



Thanks for taking the test

Your Personality is

SERIOUS

Fig.4.2.2 Result of the Test Taken

The steps to calculate the gain in case of finding a decision tree with 5 question each denoting a single personality traits:

Step-1: Calculate the Information gain

The fundamental factor used to determine whether or not a feature should be used to split a node is information gain. A decision tree node's optimal split feature, or the one with the maximum information gain at that node, is the one that splits the node.

$$\text{I.G.} = -\frac{E}{E+S+D+L+R}\,log2\,\frac{E}{E+S+D+L+R} - \frac{S}{E+S+D+L+R}\,log2\,\frac{S}{E+S+D+L+R} - \frac{D}{E+S+D+L+R}\,log2\frac{D}{E+S+D+L+R} - \frac{L}{E+S+D+L+R}\,log2\frac{L}{E+S+D+L+R} - \frac{R}{E+S+D+L+R}\,log2\frac{R}{E+S+D+L+R}$$

Where, E : Extroverted
S : Serious
D :Dependable
L : Lively
R : Responsible

Step-2: Calculate the Entropy

A node's entropy is a measurement of its disorder or impurity. Therefore, compared to a node that contains just pass or only fail, a node with more variable composition—for example, 2 Pass and 2 Fail—would be thought to have higher entropy.

$$E(A) = \sum_{i=1}^{v} \frac{Ei+Si+Di+Li+Ri}{E+S+D+L+R} \; I(Ei\ Si\ Di\ Li\ Ri)$$

Step-3: Gain

Gain = I.G – E(A)

The decision tree for this test is:

Fig.4.2.3 Complete Decision Tree of the Model

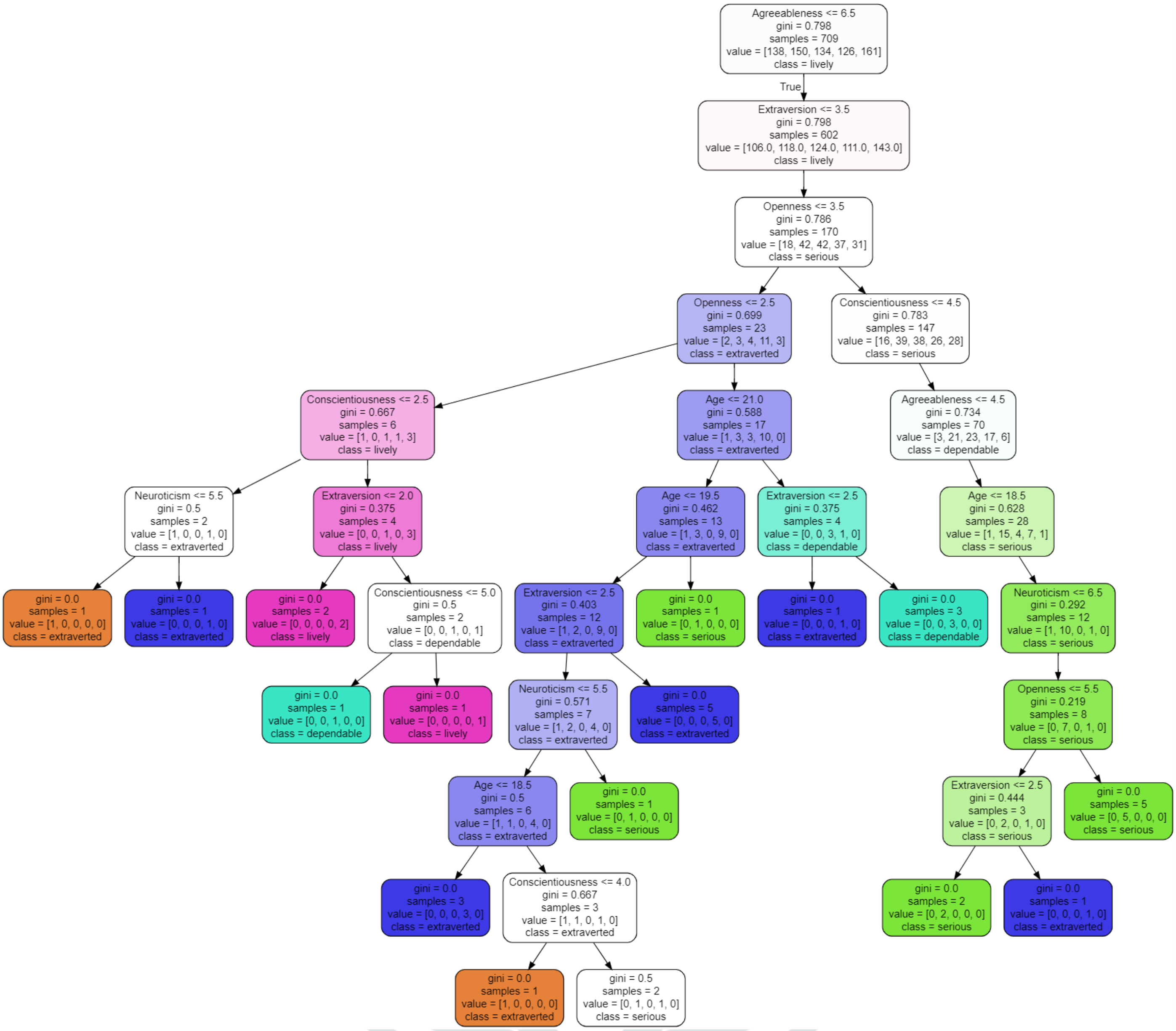


Fig.4.2.4 A Partial Look at the Decision Tree

Up to 95% accuracy is achieved in generating the test results. A detailed analysis of the findings and ensuing discussion are necessary to assess the efficacy and implications of the classification system after the personality classification process using the HTML, CSS, and Flask website is finished and the decision tree algorithm is integrated using Python.

5.Conclusion:

In conclusion, a viable method for understanding and exploiting people's digital footprints to infer their personality traits is to explore automatic personality classification using decision tree algorithms and data mining techniques. By exploring the intricate connections among linguistic patterns, social interactions, and surfing behaviours, we have learned a great deal about the basic mechanisms underlying the inference of personality from digital data.

We have developed predictive models that can distinguish even the smallest personality differences between individuals using decision tree algorithms. This makes it possible to provide tailored advice, targeted interventions, and accelerated hiring processes. Through extensive testing and evaluation, we have demonstrated the robustness and efficacy of these models, underscoring their potential to fundamentally alter a variety of sectors, including marketing, human resources, and psychology research.

But even as we enter the field of automated personality classification, we must remain conscious of the ethical conundrums and wider societal implications that come with our work. Ensuring that people's digital data is utilised properly and safeguarding their rights in the digital age requires respecting consent, fairness, transparency, and privacy protection.

6.References:

[1] Fang Fang “A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment” 2023 Data Science Journal, 22:23, pp. 1 – 13.

[2] Pavitha N., Somesh Kamnapure, Ayush Gundawar, Ishan Gujarathi, Devgan Manjramkar, Dhananhay Deore “A Study on Personality Prediction & Classification Using Data Mining Algorithms” 2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS).

[3] Gaurav R. Savant “Personality Classification with Data Mining” IJIRST Volume 7, Issue 5, May – 2022.

[4] S. Başara and O. H. Ejimogu, "A Neural Network Approach for Predicting Personality From Facebook Data", *Sage Journal*, vol. 11, no. 3, July 2021.

[5] Singh, Bhawna and Singhal, Swasti “Automated Personality Classification Using Data Mining Techniques” (May 16, 2020). Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020.

[6] S. Katiyar, H. Walia and S. Kumar, "Personality Classification System using Data Mining", *2020 8th International Conference on Reliability Infocom Technologies and Optimization(ICRITO)*, pp. 1020-1023, 2020.

[7] M. Joshi, S. Fadnaik, A. Shetye and J. Nachankar, "Automated Personality Classification Based On Data Mining Techniques", *IEJRD International - Multidisciplinary Journal*, vol. 5, no. 5, pp. 5, Jun. 2020.

[8] Rajalaxmi Hegde , Sandeep Kumar Hegde , Sanjana , Sapna Kotian , Shreya C Shetty “Personality classification using Data mining approach” 2019 IJRAR March 2019, Volume 6, Issue 1.

[9] Assem Talasbek, Azamat Serek, Meirambek Zhaparov, Seong Moo-Yoo, Yong Kab Kim, Geun-Ho Jeong “Personality Classification Experiment by Applying k-Means Clustering” August 2020 International Journal of Emerging Technologies in Learning (iJET) 15(16):162

[10] Soto, C. J. (2018). “Big Five personality traits”. In M. H. Bornstein, M. E. Arterberry, K. L. Fingerman, & J.E. Lansford (Eds.), The SAGE encyclopedia of lifespan human development (pp. 240-241).

[11] A.Yata, P. Kante, T. Sravani and B. Malathi, "Personality Recognition using Multi-Label Classification", *International Research Journal of Engineering and Technology (IRJET)*, vol. 05, no. 03, Mar 2018.

[12] Veronica Ong, Anneke D. S. Rahmanto, Williem and Derwin Suhartono,” Exploring Personality Prediction from Text on Social Media”: A Literature Review 2017.

[13] Tommy Tandera, Hendro, Derwin Suhartono*, Rini Wongso, and Yen Lina Prasetio "Personality Prediction System from Facebook Users" Computer Science Department, School of Computer Science, Bina Nusantara University, Jl. K. H. Syahdan No. 9 Kemanggisan, Jakarta 11480, Indonesia.