Personality Prediction Using Machine Learning

Dr.Jim Mathew Philip[1], Yamunaa.R[2], Siddharth Ragavendra.S[3], Rithikesh. P[4]
[1]AssociateProfessor, Dept of CSE
[2, 3, 4]Dept of CSE
[1, 2, 3, 4] SriRamakrishna InstituteofTechnology

Abstract- *Personality is useful for recognizing how people lead, influence, communicate, collaborate, negotiate business and manage stress. Personality is one of the important main features that determine how people interact with the outside world. This project is helpful when we have data related to personal behaviour. This personal behaviour data can be useful for identifying a person based on his/her personality traits. The personality characteristics will be already stored in a database. Later when a user enters his personality characteristics his personality is examined in a database and the system will detect the personality of the user, it is based on Big Five Personality Traits. Personality is one feature that determines how people interact with the outside world. This data can be helpful to classify persons using Automated personality classification (APC).*

This learning can now be used to classify/predict user personality based on past classifications. This system is useful to social networks as well as various ad-selling online networks to classify user personality and sell more relevant ads. This system will be helpful for organizations as well as other agencies who would be recruiting applicants based on their personality rather than their technical knowledge. In this project, we propose a system that analyses the personality of an applicant.

Keywords- Personal Behaviour, Big 5 Personality Traits, Automated Personality Classification,

I. INTRODUCTION

Human personality prediction has always been a challenging assignment for everyone, and throughout the recruiting process it is challenging for interviewers to ascertain the genuine personality of the interviewee. Now that it involves the scenario of online interviews, it becomes much more challenging. In the past, this was done manually by taking a lot of time to guess the person's personality. Companies with a strong consumer emphasis, such as those in retail, finance, communication, and marketing, are the main users of data mining today. Surveys, interviews, questionnaires, classroom activities, data from shopping websites, and information from social networks about user experiences and issues are some of the methods used to analyze the data. Our suggested system will reveal details about the user's personality. The system will compare the personality traits with the data in the database based on the user-provided personality features.

The system will classify the user's personality automatically and compare the pattern to the data that has been stored. The system will look over the data in the database and compare the user's personality attributes with the information there. The system will then determine the user's personality.

The system will offer additional features appropriate to the user's personality based on their personality traits. Additionally, personality can be used as a factor in the hiring process, career counselling, health counselling, etc.

Analyzing a person's behaviour to predict their personality is an old trick. It took a lot of time and effort to predict personalities manually. It was a laborious task that would take a lot of human work to do personality analysis based on one's nature. Additionally, this manual analysis produced inaccurate results when attempting to determine a user's personality based on their character and behaviour.

The accuracy of the results is impacted. After all, the analysis was done manually because people are biased by nature and tend to see things in a certain way. A human's personality can be better defined by two categories: verbal and nonverbal. These categories include several key aspects, such as verbal communication abilities, the usage of certain words or facial expressions, and so on. Nonverbal communication includes an individual's posture, speaking tone, and so on. Some additional essential variables for determining a person's personality include an individual's handwriting, social media activity such as updating posts, profile image, reaction to others' postings, and resume analysis by an Interviewer or HR. There are several personality systems for determining an individual's personality, including the MBTI (Myers Briggs), Enneagram, Big 5 personality characteristics model (OCEAN model), and others. The MBTI divides personality into 16 types, the Enneagram into 9 types, and the Big 5 into 5 kinds. The work is focused on the OCEAN model (See Figure: 1) which are as follows:

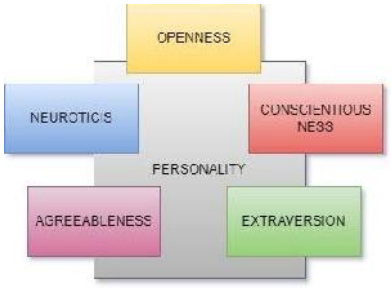

Fig1 Image of Ocean Model

A) Openness

Openness to experience is a general appreciation for workmanship, feeling, experience, unusual thinking, creative mind, curiosity, and a wide range of engagement. People that are open to new experiences are cognitively curious, sensitive to beauty, and willing to try new things. Furthermore, those with high transparency are believed to pursue self-completion deliberately by looking for something significant. Those with poor openness, on the other hand, seek fulfilment via effort and are depicted as rational and information driven occasionally even appearing to be one-sided and closed-minded. Considerable disagreement persists over how to understand and examine the openness element.

B) Conscientiousness

Conscientiousness is a proclivity for self-control and obedience. It has to do with how people command and arrange themselves. High conscientiousness is frequently perceived as demanding and focused. Low conscientiousness is associated with flexibility and immediacy, but it can also manifest as messiness and a lack of loyalty. High conscientiousness scores indicate a preference for structured over uncontrolled nature. The average degree of conscientiousness rises among young individuals and then declines among older ones.

C) Extraversion

Volatility, friendliness, forcefulness, gossip, and huge excitement are all characteristics of extraversion (or extroversion). People with a high level of extraversion are pleasant and gain energy in friendly situations. Being in the company of others makes them feel energetic and stimulated.

Those with low extraversion (or thoughtfulness) tend to be more reserved and have less energy to expend in sociable situations. Public gatherings might seem exhausting and thoughtful individuals require some amount of time of loneliness often and stillness to "re- energize".

D) Agreeableness

A person's living relationship with others, characterized by the degree of compassion and cooperation. Trust, graciousness, selflessness, feeling, and alternative prosocial activities are all characteristics of this quality. Those who place a greater value on this quality are more likely to be pleasant. Those with low levels of this trait are likely to be extremely serious and even calculating.

E) Neuroticism

A person with an undeniable degree of pleasantness on a personality test is often kind, cautious, and warm. They have a positive outlook on human instinct and get along well with others. A person with a low pleasantness rating may priorities their personal benefits over those of others. They will be intimidating, uncooperative, and difficult to reach.

II. EXISTING SYSTEM

A groundbreaking effort on personality extraction from text. They looked at words in a variety of contexts, including diaries, college writing assignments, and social psychology publications, to explore personality traits using linguistic clues. Their findings demonstrate that pleasant people use more articles, whereas introverts and those with low conscientiousness use more words to communicate differences. Neurotics employ more phrases that express unpleasant emotions. Using language elements such as function words, judgmental and assessment phrases, and modal verbs, this strategy categorized neuroticism and extraversion. Their findings demonstrated that neuroticism is associated to the usage of functional lexical aspects, such as assessment lexical taxonomy, although extraversion results were less obvious.

III. PROPOSED SYSTEM

The suggested approach predicts personality using OCEAN and machine learning techniques. Logistic Regression, Decision Tree, and Bagging Classifier are the machine learning methods employed. The Big 5 hypothesis is the most commonly acknowledged model in psychology for describing the underlying structure of human personality. The five factor model (or the Big 5 model) is the most commonly recognized model of personality based on these components. It provides a terminology and a conceptual framework for many of the research findings in individual differences and personality psychology. It condenses a vast number of personal characteristics into five key personality qualities that are represented by the acronym OCEAN. In terms of the OCEAN, the Big 5 model differs in terms of openness to

experience, conscientiousness, extraversion, agreeableness, and neuroticism. It is a comprehensive set of characteristics that may capture personality differences.

V. LITERATURE SURVEY

1] Personality Prediction Through CV Analysis using Machine Learning Algorithms for Automated E-Recruitment Process [2022]

The development of an organization and a person's personal growth is both greatly influenced by their personality. Examining the candidate's curriculum vitae or doing a standard review are two common techniques to determine a person's personality. The traditional approach of hiring people is manually shortlisting resumes of job seekers to the needs of the business. In this piece, a method that automates the process of separating candidates into groups based on eligibility requirements and personality assessments during the hiring process is suggested. To multicriteria, an online application was created that allows candidates to register their personal information and have their personalities assessed via an MCQ test. The algorithm then compares the uploaded Curriculum Vitae trained datasets to determine a user's professional eligibility. This system uses the "Logistic Regression" machine learning method, which aids in making fair selections when selecting candidates to hire. Thus, both candidates and the admin will receive the final results of the personality tests. [1]

2] Detection of Personality Using Machine Learning

Handwriting may be a reliable indicator of a character's identity thanks to data that includes the dynamically collected route, stroke, distance, length, strain, and shape of an individual's signature. The virtual signature in the biometric modality, which employs the anatomical and behavioural characteristics that a person exhibits when signing her/his name, is a new frontier for forensic handwriting examination. Dynamic information, such as velocity and stress, is fundamental and can be expected qualitatively when handwriting examiners regularly have to determine whether a signature is real or simulated. Graphology is the study of the physical attributes and writing styles that may be used to identify an author, denote their state of mind at the time they wrote, or assess their personality. It is frequently seen as a pseudoscience. [2]

VI. ALGORITHM

1. LOGISTIC REGRESSION

Logistic regression is a commonly used algorithm in machine learning for classification tasks, where the goal is to predict the class label of a given data point based on its features. It is a type of generalized linear model that models the probability of the outcome variable (dependent variable) given the values of the input features (independent variables).

The use of logistic regression in prediction involves the following steps:

Data Preparation: The first step in using logistic regression for prediction is to prepare the data by selecting the relevant features and preprocessing the data to remove any missing values, outliers, or other anomalies.

Model Training: The logistic regression model is trained on a labeled dataset, where the outcome variable is binary (0 or 1). The model learns to map the input features to the probability of the positive outcome (1).

Model Evaluation: The performance of the logistic regression model is evaluated using a validation dataset or cross-validation technique. The evaluation metrics used to assess the model's performance include accuracy, precision, recall, F1-score, ROC curve, and AUC.

Prediction: Once the logistic regression model is trained and evaluated, it can be used to predict the class label of new data points. The model computes the probability of the positive outcome (1) based on the input features and uses a threshold value to classify the data point into one of the two classes.

2. Decision Tree

Decision tree is a widely used algorithm in machine learning for classification and regression tasks. It is a non-parametric and supervised learning algorithm that uses a tree-like structure to model the relationship between the input features and the output variable.

3. BAGGING CLASSIFIER

Bagging (Bootstrap Aggregating) is a popular ensemble learning technique in machine learning that involves training multiple models on different subsets of the training data and combining their predictions to improve the performance and reduce the variance of the model. Bagging can be used with any base classifier, such as decision tree, logistic regression, or support vector machine, and is particularly effective when the base classifier is unstable or prone to overfitting.

VII. METHODOLOGY

Machine learning is a subset of artificial intelligence that involves training computers to learn and make predictions or decisions based on data without being explicitly programmed. It involves creating mathematical models that can learn from data and make predictions or decisions based on that data. The models are trained on large datasets and can be used to recognize patterns and make predictions on new, unseen data.

There are three main types of machine learning:

- supervised learning
- unsupervised learning
- reinforcement learning.

Supervised learning is a type of machine learning algorithm that involves training a model to learn from labeled data. Unsupervised learning is a type of machine learning algorithm that involves training a model to find patterns or relationships in unlabeled data. Reinforcement learning is a type of machine learning algorithm that involves training a model to make decisions based on feedback from the environment.

Machine learning algorithms can be used in a wide range of applications, including image recognition, speech recognition, natural language processing, recommendation systems, fraud detection, and predictive analytics. It is a rapidly evolving field with new techniques and applications emerging all the time.

In recent years, machine learning algorithms have been developed to predict personality traits from text data, such as social media posts, emails, and text messages. These algorithms use natural language processing (NLP) techniques to analyze the language used in the text and make predictions about the author's personality traits.

The process of predicting personality traits using machine learning typically involves the following steps:

Data Collection: A large dataset of text data is collected from various sources, such as social media platforms, emails, and text messages.

Data Preprocessing: he collected text data is preprocessed to remove noise, punctuation, and stop words.

Feature Extraction: Relevant features are extracted from the preprocessed text data, such as word frequency, sentiment, and writing style.

Model Training: A machine learning model is trained using the extracted features and the corresponding personality trait scores. The model is trained to learn the relationship between the text data and personality trait scores.

Model Evaluation: The performance of the trained model is evaluated using metrics such as accuracy, precision, recall, and F1-score. The model's ability to predict the correct personality trait scores for new, unseen text data is measured against a held- out dataset.

Model Deployment: Once the model has been trained and evaluated, it can be deployed for real-world applications such as personality assessment, marketing, and customer segmentation.

Predicting personality traits using machine learning has numerous potential applications, such as identifying job candidates who are a good fit for a particular role, creating personalized marketing messages based on customers' personality traits, and developing targeted interventions to improve mental health.

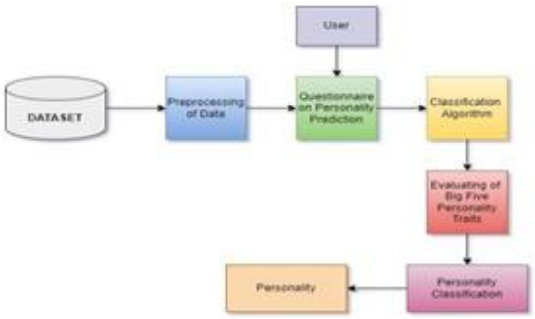

Fig2 Image of Working of Model

For This Purpose, we have built a model. The above system architecture provides an overview of how the system works. The system's operation begins with the acquisition of data from the database, after which we partition the data into training and testing data and pick the characteristics. The relevant data is then pre-processed such that duplicate data and incorrect data are removed. First, the user must log in and then complete the personality test, which consists of 10 questions, with each questions consists of 5 options. The data will be collected in an excel sheet. Based on the responses, the algorithms are applied, and the model is trained using the training data. We will look at the top five personality qualities and then categorize them. Accuracy is measured by testing the system Using testing data. So, after that personality is predicted.

VIII. IMPLEMENTATION OF PROPOSED MODEL

1. Training and Testing

The training and testing phases are very important steps in building and evaluating the performance of the model. Training and testing are essential steps in personality prediction, which is the process of determining the personality (extraverted, serious, dependable, lively, responsible) of a given data.

In personality prediction, the training and testing phases are two important stages in the machine learning pipeline. The training phase involves using a dataset of individuals' traits and behaviors to train a machine learning model to accurately predict personality traits. The testing phase is used to evaluate the performance of the trained model on new, unseen data.

During the training phase, the machine learning model is trained on a subset of the available data, which is called the training set. The model is adjusted and refined using the training data until it is able to accurately predict personality traits in the training set. This is done by adjusting the model's parameters to minimize the difference between the predicted personality trait scores and the actual scores from the training set.

After the model is trained, it is tested on a separate subset of the data, called the testing set. The testing set contains individuals who were not included in the training set, and therefore the model has not seen their data before. The model's performance is evaluated by comparing the predicted personality trait scores to the actual scores in the testing set. This helps to determine how well the model can generalize to new data.

It is important to note that the testing set should be completely separate from the training set, and should not be used in any way during the training phase. This is to ensure that the model is not overfitting to the training data, which can result in poor performance on new, unseen data.

The performance of the model on the testing set can be measured using various metrics, such as accuracy, precision, recall, and F1 score. These metrics provide information about the model's ability to correctly identify individuals with certain personality traits.

Overall, the training and testing phases are essential steps in building an accurate personality prediction model. By properly training and evaluating the model, researchers can develop a model that can accurately predict personality traits in new individuals based on their behavior or responses to personality assessments.

The result of training dataset is below:

	Gender	Age	openness	neuroticism	conscientiousness	agreeableness	extraversion
0	Male	17	7	4	7	3	2
1	Male	19	4	5	4	6	6
2	Female	18	7	6	4	5	5
3	Female	22	5	6	7	4	3
4	Female	19	7	4	6	5	4
...	...	...	...	...	...	...	...
704	Female	20	4	5	6	6	1
705	Male	18	6	3	1	5	5
706	Male	22	5	2	3	6	1
707	Male	19	5	6	5	7	5
708	Female	18	6	5	7	6	5

709 rows × 8 columns

Fig3 Image of Training Dataset Output

After completing the training process next we have moved to testing the dataset the result of this is shown below:

	Gender	Age	openness	neuroticism	conscientiousness	agreeableness	extraversion
0	Female	20	7	9	9	5	5
1	Male	17	5	4	5	2	4
2	Female	25	5	5	7	2	4
3	Female	18	6	2	7	4	7
4	Female	19	2	4	7	1	3
...	...	...	...	...	...	...	...
310	Female	19	6	5	6	4	3
311	Male	18	2	5	8	3	7
312	Male	18	7	5	6	2	7
313	Male	23	6	7	5	4	3
314	Female	18	5	7	3	5	6

315 rows × 8 columns

Fig4Image of Testing Dataset Output

2. Accuracy

The performance and accurate prediction of the model. Accuracy in personality prediction refers to how well a model can predict an individual's personality traits based on input variables. Accuracy is an important metric to consider when evaluating personality prediction models, but it should be considered alongside other metrics and practical considerations to fully understand the model's usefulness and limitations.

In personality prediction, a confusion matrix can also be used to evaluate the performance of a machine learning model. However, instead of predicting binary classes (e.g., positive/negative), the model may predict multiple personality traits, each of which can have multiple levels or categories.

In this case, the confusion matrix will be a multi-class confusion matrix that compares the predicted personality trait labels with the actual personality trait labels for each data point in the test set. The matrix will be organized into a square

matrix, where each row and column corresponds to a specific personality trait, and the cells represent the number of correct and incorrect predictions for each combination of actual and predicted labels.

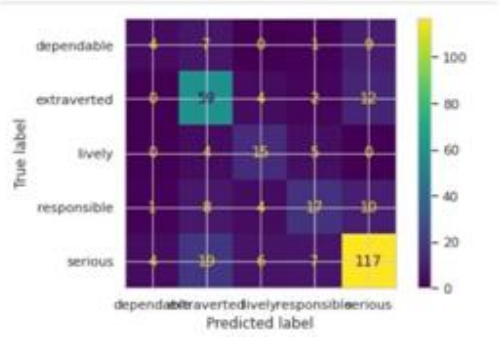

Fig5 Image of Confusion Matrix

The training and testing procedures have been finished. After that, we developed the model with the logistic regression approach and used the dataset to predict the personality. Once the model has been estimated, it can be used to predict the personality types of new individuals based on their predictor variables. The model can also be used to identify the most important predictor variables for personality prediction, which can provide insights into the factors that contribute to personality differences.

The output of the proposed system is shown below:

```
Personality for Person No 0 is dependable
Personality for Person No 1 is serious
Personality for Person No 2 is extraverted
Personality for Person No 3 is serious
Personality for Person No 4 is responsible
Personality for Person No 5 is serious
Personality for Person No 6 is serious
Personality for Person No 7 is serious
Personality for Person No 8 is serious
Personality for Person No 9 is serious
Personality for Person No 10 is serious
Personality for Person No 11 is responsible
Personality for Person No 12 is responsible
Personality for Person No 13 is serious
Personality for Person No 14 is lively
Personality for Person No 15 is extraverted
Personality for Person No 16 is serious
Personality for Person No 17 is serious
Personality for Person No 18 is serious
```

We tried the other two algorithms as well, but we didn't attain the same degree of accuracy. Only logistic regression outperformed the bagging classifier and decision tree in terms of prediction accuracy.

The logistic regression accuracy level is 67%, the bagging classifier accuracy level is 31%, and the decision tree accuracy level is 25%. As a result, we picked the model with the best forecast accuracy.

IX. CONCLUSIONS

In conclusion, logistic regression is a useful statistical method for predicting personality traits based on various input variables. By analyzing a large dataset of personality data, it is possible to identify significant predictors of personality traits and develop a reliable logistic regression model. The model can be used to predict personality traits for new individuals based on their input variables.

However, it is important to note that logistic regression models are not perfect and can have limitations, such as overfitting or underfitting the data, and may not generalize well to new data. It is also important to consider ethical implications when using such models for personality prediction, as they can potentially be used to unfairly discriminate against certain individuals or groups.

Overall, logistic regression can be a valuable tool for predicting personality traits, but it should be used with caution and in conjunction with other methods to ensure the most accurate and ethical predictions possible.

REFERENCES

[1] Fazel Keshtkar, Candice Burkett, Haiying Li, and Arthur C. Graesser, Using Data Mining Techniques to Detect the Personality of Players in an Educational Game

[2] R. Wald, T. M. Khoshgoftaar, A. Napolitano Using Twitter Content to Predict Psychopathy

[3] YagoSaez, Carlos Navarro, Asuncion Mochon and Pedro Isasi, A system for personality and happiness detection.

[4] Aleksandar Kartelj, Vladimir Filipović, Veljko Milutinović, Novel approaches to automated personality classification: Ideas and their potentials.

[5] Golbeck, J., Robles, C., and Turner, K. 2011a. Predicting Personality with social media. In Proc of the 2011 annual conference extended abstracts on Human factors in computing systems.

[6] DURGESH K. SRIVASTAVA, LEKHA BHAMBHU, “DATA Classification using Support Vector Machine,” Journal of Theoretical and Applied Information Technology

[7] YILUN WANG, “Understanding Personality through social media, “International of computer Science stand ford University.

[8] MANASI OMBHASE, PRAJAKTA GOGATE, TEJAS APTIL, KARAN NAIR, PROF.GAYATRI HEGDE, "Automated Personality classification using Data Mining Techniques," International Conference on Data and Software Testing.

[9] MAHESH KINI, SAROJA DEVI, PRASHANT G DESAI, NIRANJAN CHIPLUNKAR," Text Mining approach to classify Technical Research Document using naive Bayes", International Journal of Advanced Research in computer and communication engineering.

[10] Cantandir, I. Fernandez-Tobia] z, A. Belllogin, "Relating personality types with user preferences in multiple entertainment domains," EMPIRE 1st Workshop on Emotions and Personality in Personalized Services, 2013. P.T. Costa,R.R. McCrae (1992). Revised NEO personality inventory (NEO-PI-R) and NEO five-factor Inventory (NEOFFI), Psychological Assessment Resources.