IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 2 (May. – June. 2024), PP 28-32
www.iosrjournals.org

A Novel Approach To Personality Prediction Using The Ocean Model To Streamline The Hiring Process

Mrunal P. Joshi And Nilesh Alone
Department Of Computer Engineering, Ges's R.H. Sapat College Of Engineering,Nashik, Maharashtra

Abstract
The concept of using ocean models to predict person- ality traits based on machine learning (ML) and artificial intelli- gence (AI) is an emerging field in the intersection of psychology, computer science, and physics. The Ocean Model theory suggests that personality traits can be viewed as interconnected waves or patterns that ebb and flow over time, much like the ocean's tides. By analyzing these patterns and fluctuations using ML and AI algorithms, it may be possible to predict an individual's personality traits with a high degree of accuracy. This paper presents an overview of the Ocean Model theory, and its potential applications in personality assessment having two unique modules with a comparative study of algorithms, which can be used for streamlining the recruitment process.

Index Terms—*Machine learning, K Means Clustering, Logistic Regression.*

Date Of Submission: 12-05-2024 Date Of Acceptance: 22-05-2024

I. INTRODUCTION

This section briefs about the OCEAN model and states the problem definition. It also specifies the objectives of the study and mentions the contributions by the authors.

Background

Personality prediction based on the OCEAN model using machine learning represents a promising area of research and application. It has the potential to enhance our understanding of human behavior and improve decision-making processes specifically in the recruitment process and broader perspective it can be used for various domains

Problem Statement

To streamline the recruitment process through a machine- learning approach based on the Ocean model.

Objectives of the study

This project aims to create a machine learning and ar- tificial intelligence model that can : • Develop a machine learning algorithm that can accurately predict personality traits based on the Ocean model. • Train the algorithm on a large dataset of personality assessments to improve the accuracy and reliability of predictions. • Built a two-way system to predict the personality traits of an individual that will make the recruitment process more accurate and effective. • Provide personalized recommendations or interventions based on the predicted personality traits to help individuals improve self- awareness and personal development.

Contributions

This project aims to create a machine learning and artificial intelligence model that can predict an individual's personality traits using the OCEAN model. There will be two separate modules for the Candidate and Human resource person (HR), The Candidate module is implemented through various com- parative studies of various machine algorithms and used to develop an accurate prediction project.

II. RELATED WORKS

In [1] Allan Robey et al. successfully developed Per- sonality Prediction System through CV Analysis. In April 2022, Mr.Karthikeyan et al. in [2] presented Self-Prediction Using Machine Learning. This project helps write personality tests and diagnose people's personalities. By classifying the behavior, the type of behavior can be looked at and the behavior can be improved according to the results. In Septem- ber 2021, Atharva Kulkarni et al. [3] proposed the use of machine learning for personal prediction through CV analysis. The experimental model examines different machine learning methods to evaluate the quality of behavior by

analyzing CV using NLP technology. Mr. Karghatgi et al. proposed a neural network method based on the Big Five test aims to predict a person's behavior based on tweets posted on Twitter by extracting the meta content of the tweets. It is used to evaluate a person's behavior. The authors followed four steps: data collection from tweets, preprocessing, transformation, and classification. Although neural networks have been used to predict individuals, there are some limitations such as pre- venting misinformation, automatically analyzing tweets, and relying solely on Twitter to predict an individual's behavior is not enough, only predicting user behavior and patterns.

[4] J.Zubeda et al. are working on a project that uses word processing and machine learning for sorting. The system sorts resumes in all formats according to the company's criteria. The authors of the proposal also take into account the competitor's Github and Linkedin information to better understand the competitor, making it easier for the company to find competi- tors based on skills, capital structure and, most importantly, character. [5] [6] Dr.Md Tanzim Rıza and Dr. Sakib Zaman analyzed the personal data history using language processing and machine learning before converting the document back to HTML, then reverse-engineered the HTML code and then completed the segmentation finishing and appropriate feature extraction. The model extracts information from resumes and segments them based on importance. They used multiple logistic regression to analyze the regression analysis. However the dataset size is very small here. Another study proposed by A Mirhosseini et al.[7] builds on previous studies classify- ing MBTI types. Researchers use XGBoost to make MBTI personality predictions based on information obtained from social media. Before working on classification, the process first cleans and pre-treats the raw material, it starts with the removal of a word (URLs and restricted words) followed by lemmatization, for example using NLTK. The next step is to vectorize the processed text by weighting all relevant text elements using TF-IDF and finally complete the classification function for prediction. The results show that XGBoost's is accuracy is 78.17 percent, N accuracy is 86.06 percent, FT accuracy is 71.78 percent. In [8] Sudhir Bagade et al. said that character plays an important role in the life of an individual and the development of any organization. We created an online application that analyzes candidates' personalities based on their resume or CV. The system uses the TFIDF algorithm to select suitable candidates. VVCETCSE system predicts behav- ior based on ranking strategy. It will provide skills, experience, and other aspects of the resume. Candidates also took an aptitude test and answered personal questions. They also get the results with representative images.[9] Gangandeep Kaur et al. developed a machine learning technique called logistic regression. The system uses psychometric analysis to assess applicants' emotional intelligence and uses the OCEAN model to predict personality. Candidates' information is protected by a password encryption algorithm and the password is known only to necessary personnel. Candidates can find out whether they have been selected for the interview via the control panel and SMS. [10] Afroja Khatun Monalisa et al. in [11] made a model using a random forest algorithm, support vector machine and weighted majority voting algorithm. First of all, resumes are sent to the system and candidates are examined in line with the manager's requirements. Selected candidates are subjected to personality and skill tests that they must answer and then receive their grades. Candidates are selected based on grades and department needs.

III. METHODOLOGY

As shown in Fig no.1 Here we have gathered a diverse and representative dataset from kaggle.com that includes in- formation about individual's preferences, age, gender and various questions and answer with other relevant attributes for candidate module and for Human resource model(HR), Dif- ferent dataset from Kaggle.com is used which include relevent information with the self-reported personality scores based on standardized psychological tests. Initially the cleaning and pre- processing the dataset to handle missing values, outliers, and inconsistent data. This step involves feature engineering, trans- forming categorical variables, and scaling numerical variables. Next the most relevant features that contribute to predicting personality traits based on the OCEAN model is identified. This step also involves techniques such as correlation analysis,

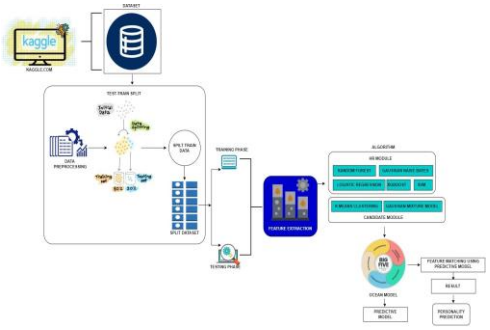

Fig. 1. Representation Of The Methodology.

Feature importance ranking, or dimensionality reduction.Then k-means clustering and Gausian mixture machine learning algorithms are impied to train a prediction model for candidate module and Support Vector Machine Algorithm, Logistic Regression, XGBoost, Random Forest, Gaussian Naive Bayes' are used for HR module. ,

Then 80 percent data is used for training and 20 percent is used for testing purposes, using the cleaned and pre- processed dataset. Next the performance of the trained model is examined using appropriate evaluation metrics namely accu- racy, precision, recall, or F1-score. This step involve splitting the dataset into training and testing sets and comparing the predicted personality scores with the actual scores. Fine-tune the model parameters and hyper parameters to improve its performance. Once the model shows satisfactory performance, it can be deployed as an application or integrated into an existing system. The application will take input data about an individual and provide predicted personality trait scores based on the OCEAN model.

This project aims to develop an accurate and reliable ma- chine learning and artificial intelligence model for predicting personality traits based on the OCEAN model.

IV. RESULTS

This section discusses the experimental setup for the model and also highlights the result analysis for the accuracy, con- fusion matrix, and cluster formation.

Experimental Setup

The machine utilized for our models was equipped with a Windows 11 operating system. It featured an Intel i3 CPU running at 2.10GHz, offering 8 cores. The system boasted 4 GB of RAM and a 512 GB SSD for storage. To ac- celerate computations, it incorporated an Nvidia GPU. The models were developed using Python 3.9 and relied on var- ious libraries, including sys, Keras 2.2.9, TensorFlow 2.12.0, Matplotlib 3.0.11, NumPy 1.16.2, and tensorflow gpu 2.7.0. Leveraging the power of the GPU, the system aimed to deliver improved performance and faster results for our models. It uses the dataset from the Kaggle.

Analysis of Results

Candidate Module

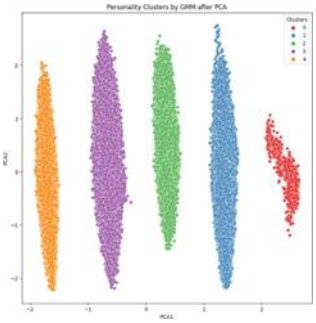

Fig. 2. Cluster formation in GMM.

In above Fig no.2 Cluster formation in Gaussian Mixture Model (GMM) involves the identification and characterization of clusters within a dataset based on the underlying assumption that the data is generated from a mixture of several Gaussian distributions. The process starts with an initial guess of the parameters for the Gaussian distributions. These parameters typically include the means, covariances, and mixing coeffi- cients (weights) for each Gaussian component. cluster forma- tion in GMM involves estimating the parameters of Gaussian distributions and assigning data points to clusters based on the likelihood of their generation. It's an iterative process aimed at finding the optimal representation of the data as a mixture of Gaussian components.

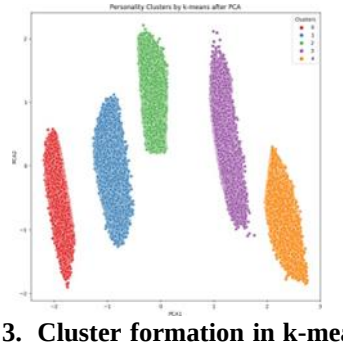

Fig. 3. Cluster formation in k-means.

As shown in the above Fig no.3 Cluster formation in K- Means clustering involves the grouping of data points into clusters based on their proximity to the centroids of those clusters.K-Means starts with an initial guess for the centroids of the clusters. These centroids can be randomly chosen from the data points or placed strategically using heuristics.Each data point is assigned to the nearest centroid, forming clusters based on Euclidean distance in the feature space. The data points are assigned to the cluster whose centroid is closest. K-Means clustering is an iterative process that aims to minimize the within-cluster variance by iteratively updating cluster as- signments and centroids. It results in the formation of clusters that group similar data points together, making it a widely used technique for unsupervised clustering tasks.

	precision	recall	f-score	support
0	0.43	0.14	0.21	21
1	0.68	0.91	0.78	77
2	0.64	0.75	0.69	24
3	0.72	0.65	0.68	40
4	0.85	0.78	0.82	153
accuracy			0.75	315
micro avg	0.66	0.65	0.64	315
weighted avg	0.75	0.75	0.74	315

TABLE I

CLASSIFICATION REPORT FOR LOGISTIC REGRESSION.

Comparing and observing the question answer in rational way the result of the GMM and K-Means algorithm, K-means results are more accurate and rational.

HR module

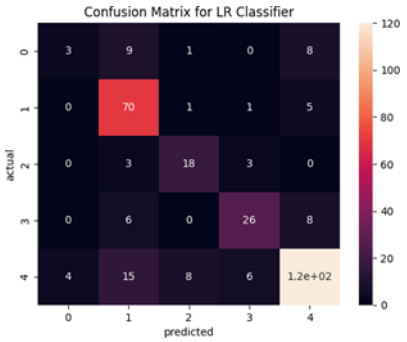

Fig. 4. Confusion Matrix for Logistic Regression.

A confusion matrix for logistic regression (LR) algorithm as shown in fig no. 4 provides a detailed breakdown of the model's performance in a binary classification task. Instances that are actually positive (belonging to the positive class) and are correctly classified as positive by the LR model. Instances that are actually negative (belonging to the negative class) and are correctly classified as negative by the LR model. Instances that are actually negative but are incorrectly classified as positive by the LR model. Also known as Type I error or false alarm. Instances that are actually positive but are incorrectly classified as negative by the LR model. Also known as Type II error or miss. By analyzing the confusion matrix and associated metrics, one can gain insights into the LR model's performance, identify areas of improvement, and make informed decisions regarding its deployment.

As per table no. 1, classification report for logistic regres- sion (LR) provides a comprehensive summary of the perfor- mance of the LR model in a binary or multi class classification task.Precision measures the accuracy of positive predictions made by the LR model.Accuracy is increased and is 0.75 . It is the ratio of true positives to the sum of true positives and false positives. Recall measures the model's ability to correctly identify positive instances out of all actual positive instances. It is the ratio of true positives to the sum of true positives and false negatives. The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall. Support is the number of actual occurrences of each class in the dataset. It indicates how many instances belong to each class. Accuracy is the proportion of correctly classified instances out of the total number of instances. While not included in the classification report per se, it is a crucial metric often reported alongside other metrics.

Traits based on patterns and correlations. Using two different modules for candidate and human resource (hr), it will help the rational recruitment process, also the project concludes that the k-means clustering is the most suitable algorithm for the candidate module (unsupervised algorithm) and Logistic regression is most accurate and efficient algorithm for HR module (supervised algorithm). This technology can be used in

fields like psychology, human resources, and marketing to make personalized recommendations and optimize advertising campaigns. However, it has limitations and ethical concerns, such as potential biases and privacy concerns. Despite these, the OCEAN model is a promising field with numerous appli- cations.

While the concept is still in its experimental phase, it can serve as a powerful tool for personal analysis and easy for the recruitment process.

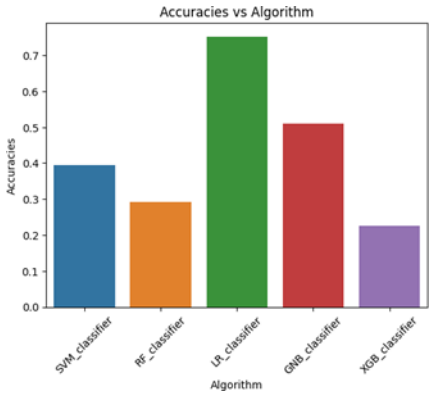

Fig. 5. Chart of Accuracy vs algorithm.

Fig 6. Accuracy vs Algorithms graph is a visualization used to compare the performance of different machine learn- ing algorithms based on their accuracy metrics.This axis represents the different machine learning algorithms being compared.Each algorithm is usually listed along the x-axis, often with labels or names for easy identification.The y- axis represents the accuracy of the algorithms.Accuracy is a com- mon evaluation metric used in classification tasks, measuring the proportion of correctly classified instances out of the total instances.It is usually expressed as a percentage, ranging from 0 percent(no correct classifications) to 100 percent (perfect classification). Each algorithm is associated with a data point on the graph. The data point represents the accuracy achieved by that algorithm on a specific dataset or datasets. The accuracy value determines the vertical position of the data point on the graph. Lines connecting the data points or bars representing the accuracy of each algorithm can be included to provide a clearer comparison between them. Lines may help in visualizing trends or patterns in the accuracy scores across different algorithms. Bars are often used in bar charts for a straightforward visual comparison. The graph usually has a title describing its purpose, such as Accuracy Comparison of Machine Learning Algorithms. Axes labels indicate what is being measured along each axis, such as Algorithms; for the x-axis and Accuracy for the y-axis. An Accuracy vs. Algorithms graph provides a visual representation of how different machine learning algorithms perform in terms of accuracy on a given dataset or across multiple datasets. It helps in selecting the most suitable algorithm for a particular task based on its performance.

V. CONCLUSIONS AND FUTURE SCOPE

The OCEAN is a widely used framework for personality prediction. Machine learning and artificial intelligence al- gorithms can analyze large data sets to predict personality

REFERENCES

[1] Allan Robey, Kashish Agarwal, Keval Joshi, Shalimali Joshi (2019).Per- sonality Prediction System through CV Analysis, IRJET vol 06 issue 02, e-ISSN-2395-0056, p-ISSN-2395-0072
[2] Devesh Agarwal, Mr. M. Karthikeyan(2022). PERSONALITY PREDIC- TION USING MACHINE LEARNING, IRJMETS Vol:04 /Issue:04/ April-2022, e-ISSN-2582-5208
[3] Atharva Kulkarni, TanujShankarwar, SiddharthThorat (2021). Personal- ity Prediction Via CV Analysis using Machine Learning, IJERT Sep 2021 Vol 10 Issue 09, ISSN – 2278-0181
[4] M. Kalghatgi, M Ramannavar, and Dr. N. S. Sidnal, "Neural Network approach to personality prediction based on the Big-Five Model" in IJIRAE, vol2 issue 8, August 2015, pp 56-63.
[5] J. Zubeda, M. Shaheen, G. Narsayya Godavari, and S. Naseem "Resume Ranking using NLP and Machine Learning", unpublished
[6] MdTanzim Reza, and Md. SakibZaman, "Analyzing CV/Resume using natural language processing and machine learning", unpublished.
[7] A mirhosseini, M.H.; Kazemian, H. Machine Learning Approach to Per- sonality Type Prediction Based on the Myers–Briggs Type Indicator®. Multimodal Technol. Interact. 2020, 4, 9.
[8] SudhirBagade, Jayashree Rout, PoojaYede, Personality Evaluation and CV Analysis using Machine Learning Algorithm
[9] VVCET - CSE, Personality Prediction System Through cv Analysis
[10] Gagandeep Kaur, ShrutiMaheshwari, Personality Prediction through Curriculam Vitae Analysis involving Password Encryption and Predic- tion Analysis
[11] Afroja Khatun Monalisa, Md. Omar Kaiser Mahin, Personallty Predic- tion System Through CV Analysis