Comparative Analysis of Psychometric Prediction System

Virti Shah
UG Scholar/CE
Indus University, Ahmedabad, India
Virti313@icloud.com

Shrey Modi
UG Scholar/IT
Charotar university of science and Technology
Modishrey007@gmail.com

Abstract— The quest to know more regarding the inner workings of the human brain and how that affects the course of actions we take has been unabated by recent technological developments in the integrated fields of psychology and computation. The eclectic collection of people participating actively on social media has led data scientists and machine learning engineers to analyse and infer from the gratuitous data available in profound quantities. There have been multiple approaches germane to the subject in hand using deep learning. The most prevalent one is the Big Five or OCEAN model which scrutinizes the dispositions or the acquired traits and then divides them into 5 broad colloquial categories namely: 'Open to Experience', 'conscientiousness', 'extraversion', 'agreeableness' and 'neuroticism'. Furthermore, Convolutional neural networking architecture has been used to hypothesize personality prediction in numerous ways, including handwriting analysis and hate-speech analysis. The motivation of the paper was to not use the convolutional neural network and try the other learning models which can give a good output and do a comparison which would indeed help to differentiate while choosing the algorithm for the other users, so we have compared two

Keywords— psychometric prediction system, model comparison, clustering, regression, machine learning

I. Introduction

Recent developments in the correlation between linguistics and the psyche have paved the way to tremendous growth in the industries aiming for prescience, like machine learning and Deep learning. They peruse through ample amounts of data and put them through rigorous training models in order to discover trends and conjecture developments in the said trends with an astonishing high value of accuracy. This novel application of psychology in computation has left the corpo- rate world with a beau coup of ways to recruit candidates more effectively by categorizing them into categories based on aptitude tests [2]. Feature selection algorithms have also been previously used to eventually minimize the error rate by finding the appropriate feature subset. A comparative analysis unequivocally shows us a comprehensive side-by- side comparison of five feature selection algorithms, namely the Pearson correlation coefficient (PCC), correlation-based feature subset (CFS), information gain (IG), symmetric uncer- tainty (SU) evaluator, and chi-squared (CHI) method [3]. In this paper, we use a regression model and perform clustering on the same dataset to concur which method brings a better solution which can be used to help the community. Accurate psychometric analysis can cater the needs of the recruitment industry and organizations which can construct methods to increase efficiency of existing employees or students.

II. LITERATURE OVERVIEW

D.W. Fiske in 1949 was the one who laid the foundation for the OCEAN or CANOE personality traits theory. OCEAN is the acronym of Openness, Conscientiousness, Extraversion or commonly referred to as Extroversion, Agreeableness and Neuroticism. The system broadly scores between the polar ends of the possibility of a given trait. In order to simplify the understanding regarding the functioning of the scoring system, we can say that for the trait Extraversion, 0 would be for someone who is entirely introverted and would rather spend time in solitude and 1 would be for someone who is entirely Extroverted and would be congenial by nature. Individuals who score in the trait of Openness tend to question more about the workings of the universe in a larger or confined sense. They are imaginative as opposed to individuals who score lower in this trait since they find it difficult to adjust to relatively new concepts and defy traditions. Individuals who score high in the Conscientiousness personality trait are the ones who believe in sharpening the axe before cutting the tree. They enjoy planning ahead of time and get done with tasks which require their utmost attention. Whereas, individuals who score low in this trait fail to complete important tasks and are overall grimy. Individuals who score high in the Agreeableness personality trait are empathetic and enjoy aiding people in need and find satisfaction in it as

opposed to individuals who score low in this trait as they care little about their acquaintance's feelings and manipulate them to get their way with them. Lastly, individuals who score high in neuroticism tend to have a relatively difficult time bouncing back after encountering stressful situations. The big five is not native to a particular region or race, and is rather universal. Scientists doubt that it may perhaps have a biological origin.\

2021 Smart Technologies, Communication and Robotics (STCR) | 978-1-6654-1806-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/STCR51658.2021.9588950

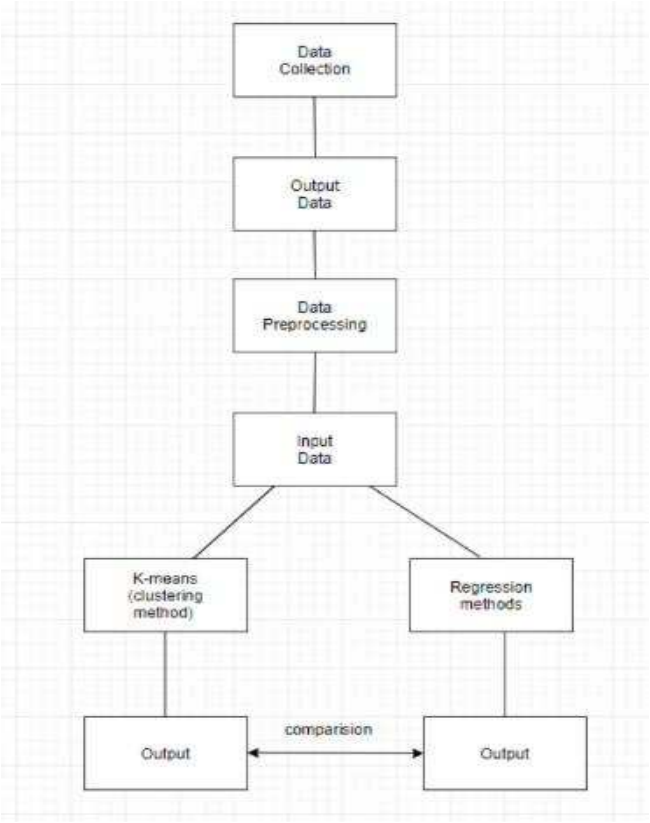

Fig. 1. Flowchart

III. PROPOSED METHOD

The above flowchart is the best explanation of the entire process performed. So the basic or the first step for any problem to appear or improve is the data or the dataset which gives a medium on which we can do our research implementation. So we will collect the data from various sources and properly preprocess it, once the data is there we can move forward by selecting a proper method to compute the data and get the results . So we have used two methods to get a comparative result, first the kmeans method and secondly the regression methods such as Logistic regression and SVC

IV. MODELS ARCHITECTURE

A. *Clustering*

The figure 2[clustering model flowchart] explains the entire kmeans clustering architecture, so firstly we would select the number of clusters k , then we assume the centroid of the cluster , Any random item can be used as the initial centroid, or the first k objects in a sequence can also be used. So the algorithm works into three steps , firstly we determine the coordinate of the centroid after that We'll figure out how

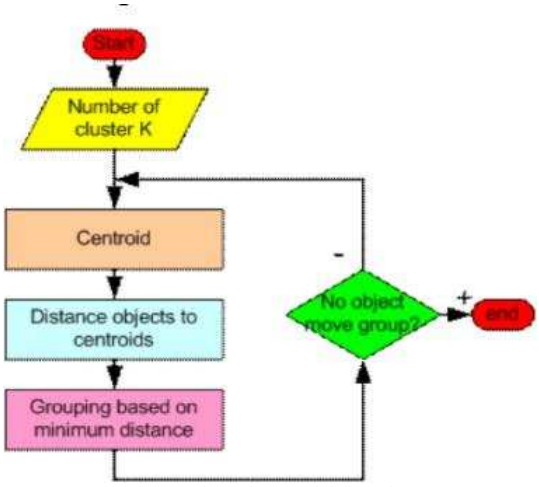

Fig. 2. Clustering model flowchart

B. *Support vector machine*

To implement nonlinear class borders, Support Vector Ma- chines use a linear model. To separate the target classes, support vectors (lines or hyperplanes) are created. To handle a nonlinear problem, the model uses a mapping function to apply numerous transformations to the data and then trains an alinear SVM model to classify the data in a higher-dimensional feature space.The data are mapped onto a high-dimensional feature space using kernel functions, allowing for linear clas- sification.Because the kernel determines the high-dimensional space where the samples will be categorised, choosing the right kernel function is critical for SVM performance. Gaussian RBF is the most widely utilised kernel function in intelligent fault diagnostics.

C. *Logistic Regression*

The method of modelling the probability of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary outcome, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model situations with more than two discrete outcomes. Logistic regression is a valuable approach of analysis.

V. METHODOLOGY

The comparative results which have been obtained through the research and the implementation using various models have followed a certain technique for achieving it. The techniques are achieved by firstly by making a dataset which is a major part of the ecosystem we are trying to build, after that the preprocessing of the dataset will take place which will lead to the model selection which in our case is a comparison between the two model methods and finally the results of the comparison.

A. *Data Preprocessing*

Data is a very important part of any ecosystem which a user is trying to make, and it plays an important role in the prediction system as the entire system is dependent on the Data and the parts of the data which are very essential for the extraction of the features, and it's detection. The collection of data is a major tedious task for anyone , we have collected our data from various users whose age are ranging from 16years to 28 years. The data is collected through a survey questionnaire which included various fields which helped us in getting the data. The fields were Gender , Age , Openness , Neuroticism, Extraversion , Agreeableness and Conscientiousness. So the user has to rate themselves on the scale of 1-10 on each of the mentioned trait's questions through which we could get a proper data systematically without any errors. The fields were all mandatory so to avoid the case of missing data which made our data accurate rather than dealing with this situation. Figure 4 shows the data overview of the data collected

	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

Fig. 4. Dataset Overview

B. Feature Extraction

Feature extraction is an important task as it basically does is identifying a particular pattern and common themes among a large amount of data. The existence and activity of other persons have an impact on persons behaviour. These interactions can have an effect on how new information or behaviours are passed down through the groupings. Understanding how such behaviours emerge and spread has a wide range of possible uses . There are various feature extraction techniques such as PCA(Principal Component Analysis) , LDA(Linear Discriminant Analysis). We would be using both of them as we are going to compare both regressional model and the clustering model. So LDA is used for the supervised models while PCA is used for unsupervised models, so the main reason for using the respective extraction methods is due to the above reason. So PCA is a method in which we obtain important variables from a large set of the variables available in the dataset, it finds the direction of the maximum variation. While LDA should be used mainly for the supervised algorithms, LDA works similarly to the PCA but in LDA we require labels unlike for the PCA.

C. Building the model

The basic architecture of the model which we are going to use is discussed above in the model architecture. Learning algorithms are divided into two parts supervised and unsuper- vised learning and in this We would be using supervised as well as unsupervised machine learning models. In supervised, we have used SVM (support vector machine) and Logistic regression, while in unsupervised machine learning we are using clustering. So first in the regression algorithms we would be dividing the dataset into two parts, the training dataset which would be used to train the regression model and the testing dataset on which we would test the accuracy of the model. The probability of the default class which are the features in our case is modelled using logistic regression and by finding the hyperplane and minimizing the range between the anticipated and observed values, SVM tries to reduce error. So we have performed the regression algorithm together and tried to achieve the accuracy from the algorithms applied. While in the unsupervised learning case we have used Kmeans for the clustering which would try to cluster the similar labels/features by choosing k as 5 just for the testing purpose and 5 is preferable for the basic testing purpose, so as the results would be getting the labels distance to the nearest centroid of a cluster by which we would be able to predict the personality of a person.

D. Testing the model

So firstly in the supervised learning section as we used the train test split to have two dataset one for the training and one for the testing, so we would be creating a confusion matrix(Fig 5) which would be able to give the exact correlation between the different labels used and how good our model is performing and where it is facing a problem.

In the unsupervised machine learning model we would be creating clusters of the training dataset which would create an idea if the clustering is getting done properly or not , we are taking the PCA features and creating a graph where the clusters would be there will give a clear idea about the labels as shown in the Fig 6.

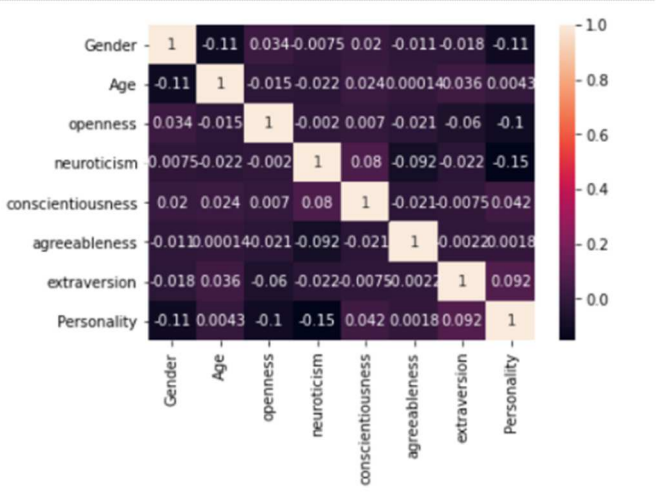

Fig. 5. Confusion Matrix

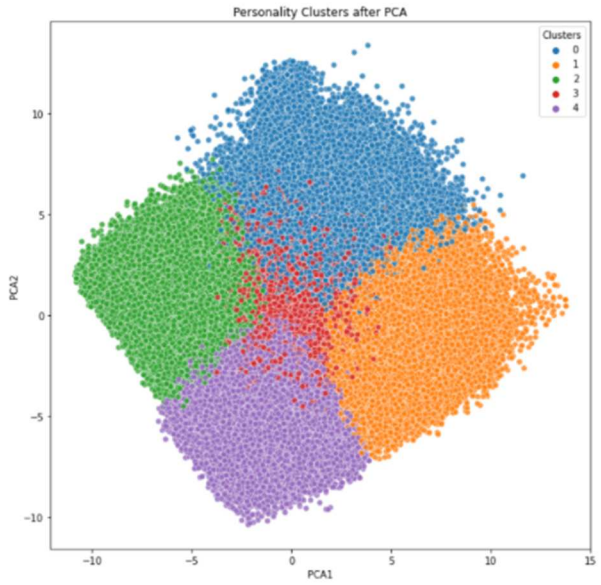

Fig. 6. Clustering Visualization

VI. EXPERIMENTAL RESULTS

So now the comparison case comes into play where we would have to match the results of both of the models, so for that we have to find the accuracy of both of the methods which will be a main criterion for the comparison along with the other factors. We would be getting a particular accuracy for the regression models, but in clustering we would have to find the accuracy as we just have the distance from the centroid of a particular cluster. So After comparing the results of both of the cases we got to know that clustering plays a better role in determining the personality as it gives a proper cluster in which the personality of a person belongs while in Regression we are getting an accuracy, but the accuracy varies between 60-65 on an average which is not a good accuracy to predict the type of personality, so we got to know that the clustering performs better than the SVM and Logistic regression. In the figure 7 we are predicting the cluster of a unknown data which was not fed into the dataset, and we got that the personality of that person belongs to cluster 2 which is Agreeable personality.

```
my_personality = k_fit.predict(my_data)
print('My Personality Cluster: ', my_personality)
```

```
My Personality Cluster:  [2]
```

Fig. 7. Clustering result

VII. CONCLUSION

After performing clustering and regression on the same database, we can infer that clustering is a better approach than regression in the said case Because all we're doing is computing the distances between points and group centres, clustering(Kmeans) has the benefit of being quite quick. As a result, it has a linear complexity which seems faster than SVM and regression. And grouping would help us to predict what personality a person falls in which is more accurate than the other two. The Big Five psychometric prediction system can be used in the corporate world to fulfil the purpose of recruitment of candidates whose personality profile can coincide with the desired one. In doing so, it will increase the output of the human resources department. The recruitment as well as the employee sector will benefit from the paper. The novel world requires a greater output in a shorter period of time. This system will significantly aid them because the results define that one system provides better outputs than the other.

VIII. FUTURE WORK

Furthermore, We can use this dataset to perform and compare the results of various models using different techniques like transfer learning, reinforcement learning, convolutional neural networks. We have a comparative small database which can be used as a platform for different operations from which numerous conclusions can be derived in order to determine the superior method.

IX. REFERENCES

[1]V. Ong et al., "Personality prediction based on Twitter information in Bahasa Indonesia," 2017 Federated Conference on Computer Sci- ence and Information Systems (FedCSIS), 2017, pp. 367-372, doi: 10.15439/2017F359

[2]https://www.academia.edu/download/59934116/IRJET-V6I234320190704-20746-16py1b3.pdf

[3]A. A. Marouf, M. K. Hasan and H. Mahmud, "Comparative Analysis of Feature Selection Algorithms for Computational Personality Prediction From Social Media," in IEEE Transactions on Computa- tional Social Systems, vol. 7, no. 3, pp. 587-599, June 2020, doi: 10.1109/TCSS.2020.2966910.

[4]E. Tas¸ and M. E. Kamas¸ak, "Prediction Of Personality Traits From Videos By Using Machine Learning Algorithms," 2019 4th International Conference on Computer Science and Engineering (UBMK), 2019, pp. 778-782, doi: 10.1109/UBMK.2019.8907179.

[5]P. S. Dandannavar, S. R. Mangalwede and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," 2018 International Confer- ence on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 62-65, doi: 10.1109/CTEMS.2018.8769304.

[6]K. S o¨nmezo¨z, O¨ . Ug˘ur and B. Diri, "MBTI Personality Prediction With Machine Learning," 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020, pp. 1-4, doi: 10.1109/SIU49456.2020.9302239

[7]A. V. Kunte and S. Panicker, "Using textual data for Personality Prediction:A Machine Learning Approach," 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 2019, pp. 529-533, doi: 10.1109/ISCON47742.2019.9036220.

[9]N. R. Ngatirin, Z. Zainol and T. L. Chee Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 2016, pp. 435-440, doi: 10.1109/ICCSCE.2016.7893613

[10]M. A. Moreno-Armend a´riz, C. A. Duchanoy Mart´ınez, H. Calvo and M. Moreno-Sotelo, "Estimation of Personality Traits From Portrait Pictures Using the Five-Factor Model," in IEEE Access, vol. 8, pp. 201649- 201665, 2020, doi: 10.1109/ACCESS.2020.3034639

[11]M. A. Iqbal and A. Shah, ”A Novel RE Teams Selection Process For User-Centric Requirements Elicitation Frameworks Based On Big-Five Personality Assessment Model,” 2020 IEEE 15th International Confer- ence on Industrial and Information Systems (ICIIS), 2020, pp. 522-527, doi: 10.1109/ICIIS51140.2020.9342649.
[12]J. Fu, J. Chang, Y. Huang and H. Chao, ”A Support Vector Regression-Based Prediction of Students’ School Performance,” 2012 International Symposium on Computer, Consumer and Control, 2012, pp. 84-87, doi: 10.1109/IS3C.2012.31.
[13]M. A. Moreno-Armend a´riz, C. A. Duchanoy Mart´ınez, H. Calvo and M. Moreno-Sotelo, ”Estimation of Personality Traits From Portrait Pictures Using the Five-Factor Model,” in IEEE Access, vol. 8, pp. 201649- 201665, 2020, doi: 10.1109/ACCESS.2020.3034639.
[14]A. A. Kındıro g˘lu, L. Akarun and O. Aran, ”Vision based personality analysis using transfer learning methods,” 2014 22nd Signal Processing and Communications Applications Conference (SIU), 2014, pp. 2058- 2061, doi: 10.1109/SIU.2014.6830665.
[15]]L. Teijeiro-Mosquera, J. Biel, J. L. Alba-Castro and D. Gatica-Perez, ”What Your Face Vlogs About: Expressions of Emotion and Big- Five Traits Impressions in YouTube,” in IEEE Transactions on Affec- tive Computing, vol. 6, no. 2, pp. 193-205, 1 April-June 2015, doi: 10.1109/TAFFC.2014.2370044.
[16]S. Katiyar, H. Walia and S. Kumar, ”Personality Classifica- tion System using Data Mining,” 2020 8th International Con- ference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 1020-1023, doi: 10.1109/ICRITO48877.2020.9197803.
[17]G. V. Rohit, K. R. Bharadwaj, R. Hemanth, B. Pruthvi and M. V. Manoj Kumar, ”Machine intelligence based personality prediction using social profile data,” 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1003-1008, doi: 10.1109/ICSSIT48917.2020.9214175
[18]A. A. Kındıro g˘lu, L. Akarun and O. Aran, ”Vision based personality analysis using transfer learning methods,” 2014 22nd Signal Processing and Communications Applications Conference (SIU), 2014, pp. 2058- 2061, doi: 10.1109/SIU.2014.6830665.
[19]Modi, Shrey, and Mohammed Husain Bohara. ”Facial Emotion Recognition using Convolution Neural Network.” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021
[20]Sai Rakshith Potluri, V Sridhar, Shrisha Rao, Effects of data localization on digital trade: An agent-based modeling approach,Telecommunications Policy,Volume 44, Issue 9, 2020,102022,ISSN 0308-5961