e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

PERSONALITY PREDICTION USING SUPERVISED AND UNSUPERVISED ALGORITHMS

Shivnath Mandal[*1], Manish Yadav[*2], Rahul Maurya[*3], Kartikey Pathak[*4]

[*1,2,3,4]Department Of Information Technology, Thakur College Of Science And Commerce Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India.

ABSTRACT

The project uses Classification Algorithm like K-Mean Clustering, Hierarchical clustering Logistic Regression and Support Vector Machine to mine user characteristics data and learn from the patterns. This project comes where large data of personal behavior is used. This project is useful in identifying personality behavior of a person. Both the algorithms are implemented and simulated using python and the scikit-learn package. After implementation and simulation, it was observed that the performance of the logistic regression is more compared to the Support vector machine in the case of logistic accuracy was 87.11% whereas Support vector machine has shown 72.49%. Therefore, the research proposed that Logistic is better than support vector machine to predict the academic performance of the students.

Keywords: Logistic Regression And Support Vector Machine Accuracy Of Prediction, Personality Test.

I. INTRODUCTION

The project is based on identifying the personality of an individual using machine learning algorithms and big 5 models. The personality of a human plays a major role in the success of an organization. Nowadays, many organizations have also started selecting the candidates based on their personality as this increase the success rate of the work because the person is working in what he is good other than what he is not good to do. The Big Five model is also known as the Five-Factor Model (FFM) and OCEAN model was developed in the early 1980s according to many psychological theories. When the statistical analysis is applied to personality survey data, some words which describes the personality of a person and these words give a idea of the overall character or personality of the person accurately. In this project, two Supervised Algorithm i.e., Logistic Regression and Support Vector Machine is used to predict personality of a person.

II. LITERATURE REVIEW

Various researchers have contributed to the task of personality prediction. Some of them have taken psychological tests into account for deciding labels of personality while others have used machine learning algorithms like Naïve Bayes for prediction and most of them uses Facebook or twitter data to predict the personality. In the above Research Paper, Experts have done analysis to predict the personality through the social media site in which they have used Myers Briggs test for prediction So, all of them uses social media data or cv based data for prediction. Due to which we get an idea that how exactly this analysis was proceeded. So, we have decided to work on Personality based Question (MCQs).

III. LOGISTIC REGRESSION

Logistic regression is the most popular Machine Learning algorithms. It is part of Supervised Learning technique, It is mostly used for classifying object into different categories. For example, apple and tomato so apple is fruit and tomato is a vegetable.

IV. SUPPORT VECTOR MACHINE

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

V. K-MEANS CLUSTERING

K-Means Clustering algorithm is an Unsupervised Learning algorithm, which creates the groups of the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

VI. HIERARCHICAL CLUSTERING

Hierarchical clustering is another unsupervised learning algorithm that is used to create group together the unlabeled data points having similar characteristics and similar data. Hierarchical clustering algorithms falls into following two categories.

Agglomerative hierarchical algorithms

Divisive hierarchical algorithms

VII. METHODOLOGY FOR SUPERVISED METHOD

The methodology is the important part of any research-related work. The methods use to gain the result are shown in the methodology. Here the whole research implementation is done using python. There are different steps involved to get the entire research work done which is as follows:

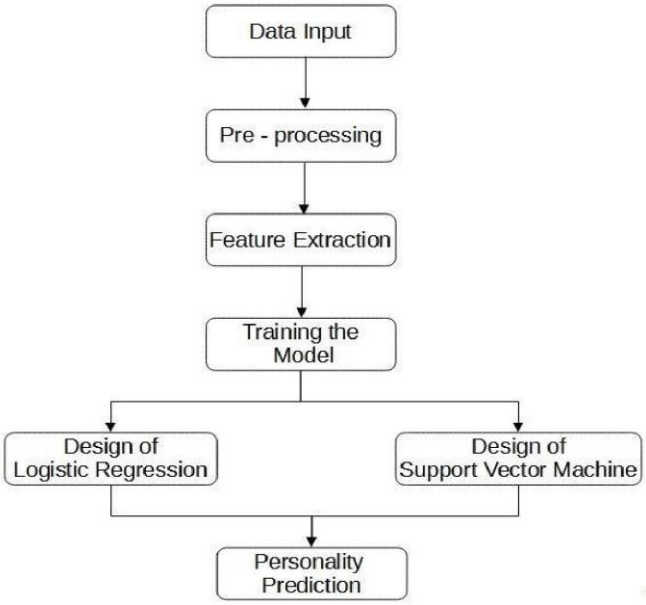

Fig. 1. Flowchart for supervised method

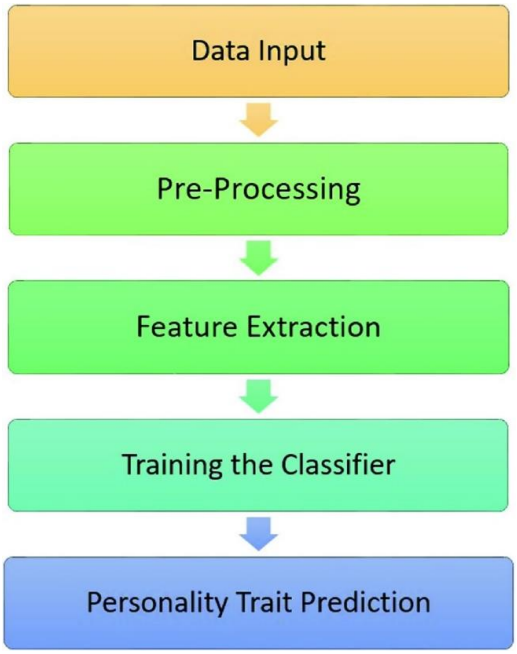

Fig. 2. Flowchart for unsupervised method

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

7.1. Acquire Personality Dataset

The kaggle machine learning is a collection of datasets, data generators which are used by machine learning community for analysis purpose. The personality prediction dataset is acquired from the kaggle website. This dataset was collected (2016-2018) through an interactive on-line personality test. The personality test was constructed from the IPIP. The personality prediction dataset can be downloaded in zip file format just by clicking on the link available. The personality prediction file consists of two subject CSV files (test.csv & train.csv). The test.csv file has 0 missing values, 7 attributes, and final label output. Also, the dataset has multivariate characteristics. Here, data-preprocessing is done for checking inconsistent behaviors or trend.

7.2. Data preprocessing

After, Data acquisition the first step is to clean and preprocess the data. The Dataset available has numerical type features. The target value is a five-level personality consisting of serious, lively, responsible, dependable & extraverted. The preprocessed dataset is further split into training and testing datasets. This is achieved by passing feature value, target value, test size to the train-test split method of the scikit-learn package. After splitting of data, the training data is sent to the following Logistic regression & SVM design is used for training the artificial neural networks then test data is used to predict the accuracy of the trained network model.

7.3. Feature Extraction

The following items were presented on one page and each was rated on a five point scale using radio buttons. The order on page was EXT1, AGR1, CSN1, EST1, OPN1, EXT2, etc. The scale was labeled 1=Disagree, 3=Neutral, 5=Agree

7.4. Training the Model

Train/Test is a method to measure that model is accurate to data set. It is called Train/Test because the data set is split into two sets: a training set and a testing set. 20% for testing, and 80% for training. The training set is used to train the data. In this model linear model is used to train the data. Logistic Regression & SVM from sklearn Package.

7.5. Personality Prediction Output

After the training of the designed neural network, the testing of Logistic Regression & SVM is performed using Cohen kappa score & Accuracy Score.

VIII. RESULT

The result is a very important component in any research paper. The analysis of the whole system is done with a result calculation. there are different efficiency and accuracy measurement methods present. Here, the confusion matrix or error matrix and Cohen's Kappa coefficient are used. The confusion matrix is used for summarizing the performance of classification problems. It provides the error that is occurring. Cohen's Kappa coefficient is a statistical method that measures inter-annotator agreement. Its values range between 0 and 1. The more the value closer to 1 greater the efficiency. After the calculation of the confusion matrix and Kappa coefficient, the accuracy of the Logistic Regression is 87.11% with a 0.80 kappa value from Table 1. On the other hand, the accuracy of the Support Vector Machine is 72.49% with a 0.55 kappa coefficient value from Table 2. Thus, it can be seen that Logistic Regression provides greater accuracy than Support Vector Machine.

On other hand with respect to our Unsupervised Algorithm i.e., K-mean Clustering we used silhouette score. Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means cluster are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

So, we got score about: 0.056

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:08/August-2022 Impact Factor- 6.752 www.irjmets.com

Table 1. Accuracy Assessment Of Logistic Regression

Sr. No	Parameter	Value
1	Accuracy	87.11%
2	Kappa Value	0.80

Table 2. Accuracy Assessment Of Support Vector Machine

Sr. No	Parameter	Value
1	Accuracy	72.49 %
2	Kappa Value	0.55

IX. CONCLUSION

Personality is important in a person's life. Thus, having prior knowledge of person's personality based on their behavior, demographic and social attributes is very essential. The two models of Supervised Machine Learning Algorithm i.e. Logistics Regression and Support Vector Machine are implemented successfully for the prediction of person's personality in personalities i.e. (Dependable, Serious, Responsible, Lively and Extraverted). From research, it can be concluded that personality's attributes such as question on all the five-behavior analysis etc., are important for prediction. Also, the creation of new attributes helps in better prediction. The above-mentioned is implemented using Logistics Regression and Support vector machine and its analysis is done with confusion matrix and kappa coefficient as shown in the result section. It shows that Logistics Regression accuracy is 87.11% and 0.80 kappa value which is greater than Support vector machine accuracy (72.49%) and kappa value (0.55). Thus, prediction and correct classification of person's performance can be achieved with Logistic Regression.

Various inferences can be made from the correlation matrix, the relation between every feature with personalities, heat-map, chart diagrams available from the data visualization and data exploration. Data visualization and exploration help a lot to understand the data and its significance with others.

X. REFERENCES

[1] https://www.kaggle.com/khotijahs1/big-five-personality-test-clustering.

[2] Personality prediction based on Twitter information in Bahasa Indonesia.

[3] Personality predictions based on user behavior on the Facebook social media platform Michael M Tadesse, Hong Fei Lin, Bo Xu, Liang Yang

[4] Dataset used from https://www.kaggle.com/ and https://archive.ics.uci.edu/

[5] https://www.academia.edu/Documents/in/.

[6] Algorithm study from https://www.javatpoint.com/.