A picture containing text

Description automatically generatedA picture containing text

Description automatically generated

A picture containing text

Description automatically generated

**Bilal Ahmad**

**FA19-BCS-136**

**Osama Arshad**

**FA19-BCS-066**

**Personality Prediction**

**Summary**

Predicting human behaviour and personality from the social media applications like Facebook, Twitter is achieving tremendous attention among researchers. Statistical information about the human thoughts is essential assets for research in predicting various human behaviour and personality. The current work mainly focuses on guessing user personality based on big five personality traits. An intelligent Sentence analysis model is built to extract personality features.

The corporate world today does not focus just on the skills a potential employee possesses but also their personality. Personality is what helps one be successful in professional as well as personal life. Hence, the recruiter must be aware of the personality traits a person has. With an exponential increase in job seekers, it is difficult to manually shortlist the best fit candidate for a suitable job by looking at the CV. Project focus on to examine different machine learning approaches for efficiently predicting personality through CV analysis using NLP techniques as well. Results show that the Random Forest algorithm achieved better accuracy when compared to other algorithms such as KNN, Logistic Regression, SVM and Naive Bayes.

The project is based on identifying the personality of an individual using machine learning algorithms and big 5 models which is known as the OCEAN model. When the statistical analysis is applied to personality survey data, some words used to describe the person of the overall character or personality of the person accurately.

**Open to Experience:** It involves various dimensions, like imagination, sensitivity, attentiveness, preference to variety, and curiosity.

**Conscientiousness:** This trait is used to describe the carefulness and diligence of the person. It is the quality that describes how organized and efficient a person is.

**Extraversion:** It is the trait that describes how the best candidates can interact with people that is how good are his/her social skills.

**Agreeableness:** It is a quality that analyses the individual behavior based on the generosity, sympathy, cooperativeness, and ability to adjust with people.

**Neuroticism:** This trait usually describes a person to have mood swings and has extreme expressive power.

**Methodology and working**

* **Choosing a dataset:**

Getting your data is the first step of model training. Dataset must be accurate and authentic. There can be multiple sources of data like sensors, flat files, database, data warehouse, social media websites etc. We get our project dataset from Kaggle. Reference: <https://www.kaggle.com/datasets/tunguz/big-five-personalitytest?datasetId=516764>

* **Cleaning the dataset:**

Pre-processing the dataset is the second step of machine learning. Removing noisy data, filling the null (empty) values, replacing the junk data, find unknown columns using formulas is important before training a better model.

For this purpose, we use different functions of python libraries like NumPy, Pandas etc. for cleaning the dataset.

* **Visualize the dataset:**

Understanding the dataset and choosing the method of Machine learning is important before training a model. Visualize the rows and columns into graphs, finding the trend, removing outliers and grouping the dataset columns for training is the part of defining the dataset.

As our problem is based on unsupervised learning, so we choose clustering.

* **Clustering the dataset:**

Before training the model, we need to find how many clusters should be there? Defining the number of clusters in important to find each distinct features of dataset. For this purpose, we use elbow method in order to find the best number of clusters which fits best with the dataset.

As we’re following OCEAN model, so there will be 5 main clusters for 5 different types of personalities.

* **Train the Model:**

Training the model is the main part of Machine learning. Train multiple algorithms on a dataset and check which will give the best the best accuracy.

There are many algorithms use for clustering like K-mean, partitioning etc. Provide the defined dataset, number of clusters to algorithm and it will train the model. Evaluate the model using confusion matrix, accuracy, ROC curve etc.

* **Visualize the Model:**

Visualizing of model is good for checking either the trained model will work well or not. Finding the outliers, errors are the objective of model visualization.

As our dataset contain rows in millions, so we used PCA model to reduce the dimensionality and distribute the correlated features linearly.

* **Test the Model:**

Testing the model is as important as training the model. After training a model, an unlabeled data will be provided, and trained model will predict the label of the data and give results.

If results are wrong, train the model with different algorithm again and again until the model provide a good accuracy and predictions.

**Results and Findings**

After evaluating the model, input the data for testing and find the results of prediction.

For example, Input the answer set of personality-based question and prediction your personality based on trained model with test data and get the score of your personality features based on predicted cluster.

**For example:**

* **Input the test data:**

data = pd.read\_excel('../dataset/test\_data.xlsx')

* **Test the model:**

personality = k\_fit.predict(data)

* **Results:**

| **extroversion** | **neurotic** | **agreeable** | **conscientious** | **open** | **cluster** |
| --- | --- | --- | --- | --- | --- |
| 3.2 | 2.4 | 3.4 | 3.3 | 3.4 | **4** |