PERSONALITY PREDICTION

A report submitted for the course of

Application Development Machine Learning Explore

III B. Tech I Semester

by

S. Hema Swaraj	**- 2111CS030041**
G. Kranthi Varma	**- 2111CS030048**
P. Nithin Kumar	**- 2111CS030069**
M. Raghavendra	**- 2111CS030083**

Under the esteemed guidance of

Ms. Rubeena Rab

(Assistant Professor)

Department of Computer Science & Engineering (Data Science)

Malla Reddy University

Maisammaguda, Dulapally,

Hyderabad, Telangana 500100

www.mallareddyuniversity.ac.in

2023-24

MALLA REDDY UNIVERSITY

(Telangana State Private Universities Act No.13 of 2020 and G.O.Ms.No.14, Higher Education (UE) Department)

DATA SCIENCE

CERTIFICATE

This is to certify that this bonafide record of the application development entitled Personality prediction submitted by S.HemaSwaraj (2111CS030041), Mr.G.Kranthi Varma (2111CS030048**),** Mr.P.NithinKumar (2111CS030069**)**, Mr.M.Raghavendra (2111CS030083**)** of III year I semister to the Malla Reddy University, Hyderabad. This bonafide record of work carried out by us under the guidance of our supervision. The contents of this report, in full or in parts, have not been submitted to any other Organization for the award of any Degree.

INTERNAL GUIDE:	**HEAD OF THE DEPARTMENT**
Ms.Rubeena Rab	**Dr.G.S Naveen Kumar**
(Assistant Professor)	**CSE(Data Science)**

Date:

External Examiner

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 determines how people interact with outside world. This project is helpful where we have data related to personal behaviour. This personal behaviour data can be useful for identifying person based on his/her personality traits. The personality characteristics will be already stored in database. Later when user enters his personality characteristics his personality is examined in database and system will detect the personality of user, It is based on Big Five Personality Traits Personality is one feature that determines how people interact with the outside world. 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 which analyses the personality of an applicant.

Keywords: personality traits, Machine learning Models, Behavioral Analysis, Classification Algorithms.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DISC	Dominance, Influence, Steadiness and Compliance.
MBTI	Myers-Briggs Type Indicator.

CHAPTER-1

INTRODUCTION

Personality is a key aspect of human life. More specifically personality is a branch of psychological study. Personality is constituted of elements like person's thoughts, feelings, behavior which continuously keeps on changing over time. Prediction of personality is an area of study where person gets categorized in a class according to his/her personality. There are number of psychological tests that yield different type of personality classes. Popular tests include DISC is another test of psychology that classifies personality in categories as Dominance, Influence, Steadiness and Compliance. [1]MBTI psychological test has 16 categories of personality. All these traditional methods of personality prediction use questionnaire for personality prediction. Filling a lengthy questionnaire is time consuming and tedious job, to overcome these lengthy methods BIG-FIVE PRSONALITY TRAITS came into existence. The personality classification in five categories as openness to experience, conscientiousness, extraversion, agreeableness and neuroticism.

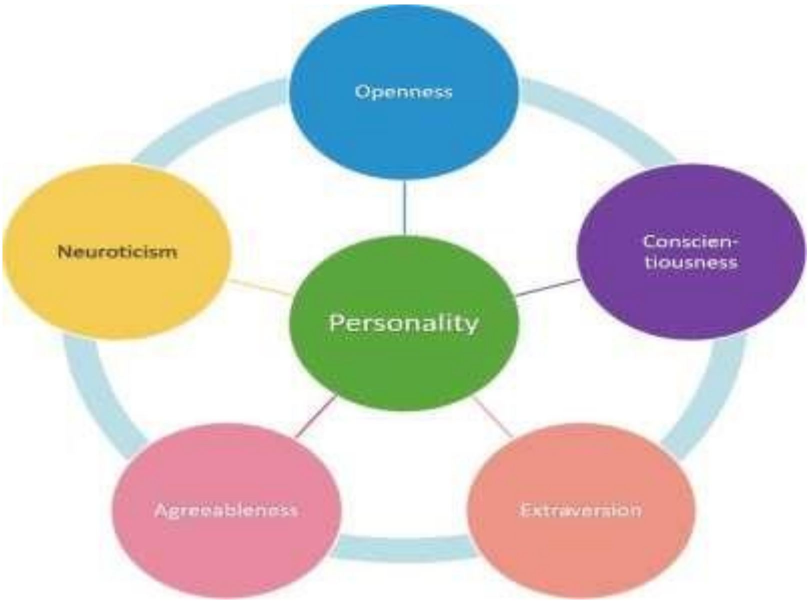


Fig-1.1: Big Five Personality Traits

The Big Five Personality traits are the five dimensions or the domains of personality that can be used to analyse or predict the personality of a user. The Big Five Personality Model is the most widely accepted and researched model for predicting the personality of a user. The Big Five Personality traits are found in a variety of people of different ages, locations and cultures. The Big Five Personality results are very accurate and predict the true personality of a user to a large extent.

The big five factors are:

1) Openness to Experience or Imagination Capability
2) Agreeableness
3) Extraversion
4) Neuroticism or Emotional Stability
5) Conscientiousness

- The Big Five personality traits extraversion (also often spelled extroversion), agreeableness ,openness , conscientiousness , and neuroticism.
- Each trait represents a continuum. Individuals can fall anywhere on the continuum for each trait.
- The Big Five remain relatively stable throughout most of one's lifetime.
- They are influenced significantly by both genes and the environment, with an estimated heritability of 50%.
- They are also known to predict certain important life outcomes such as education and health.

- **Openness to experience:** Also called as intellect or imagination, this personality trait represents the willingness to try new things and think out of the box. This trait includes insightfulness, originality, and curiosity.

- **Conscientiousness:** The desire to be careful, diligent and regulate immediate gratification with self discipline . This trait includes ambition, discipline, consistency and reliability.

- **Extroversion:** A state where an individual draw energy from others and seek social connections or interaction, as opposed to being alone. This trait includes being outgoing, energetic, and confident.

- **Neuroticism:** A tendency for negative personality traits, emotional instability, and self- destructive thinking. This trait includes pessimism, anxiety, insecurity, and fearfulness.The personality of individuals are predicted using data mining concepts. Before testing the dataset, it is pre-processed using different data mining concepts like handling missing values, data discretization ,normalisation etc. This pre-processed data can then be used to classify/predict user personality based on past classifications. The system analyses user characteristics and behaviors. System then predicts new user personality based on personality data model used to predict test dataset is "Logistic Regression" because Logistic regression is an effective model to predict output class labels for dependent categorical data.

DATA SET DESCRIPTION:

S.NO	ATTRIBUTE	TYPE	RANGE
1	Gender	Nominal	Male/Female
2	Age	Numeric	…..
3	Openness	Numeric	1-8
4	Neuroticism	Numeric	1-8
5	Conscientiousness	Numeric	1-8
6	Agreeableness	Numeric	1-8
7	extraversion	Numeric	1-8

Table1: Data Set Description.

1.1 Tools and libraries:

- **python**

In our project we used python ,It is the high level programming language ,It has a huge library. Python Programming Language is very well suited for Beginners, also for experienced programmers with other programming languages like C++ and Java.

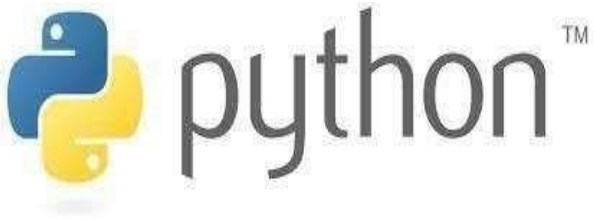


Fig-1.2: Python

- **Jupyter lab:**

Next we used jupyter lab because Jupyter Lab offers a general framework for interactive computing and data science in the browser, using Python, Julia, R, or one of many other languages we used jupyter lab to install the libraries which are used in our project.

Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality. Jupyterlab can open multiple ". ipynb" files inside a single browser tab. Whereas, Jupyter Notebook will create new tab to open new ". ipynb" files every time.

Fig-1.3: Jupyter lab

- **Jupyter Notebook**

The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Jupyter notebook can be the showcasing your work. You can see both the code and the results. Jupyter notebook has the ability to display plots that are the output of running code cells. The IPython kernel is designed to work seamlessly with the matplotlib plotting library to provide this functionality. Specific plotting library integration is a feature of the kernel.

- **NumPy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and you can use it freely. NumPy stands for Numerical Python. In Python we have lists that serve the purpose of arrays, but they are slow to process. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called `ndarray`, it provides a lot of supporting functions that make working with `ndarray` very easy. Arrays are very frequently used in data science, where speed and resources are very important. NumPy guarantees efficient calculations with arrays and matrices on high-level mathematical functions that operate on these arrays and matrices. The NumPy arrays takes significantly less amount of memory as compared to python lists. It also provides a mechanism of specifying the data types of the contents, which allows further optimization of the code.

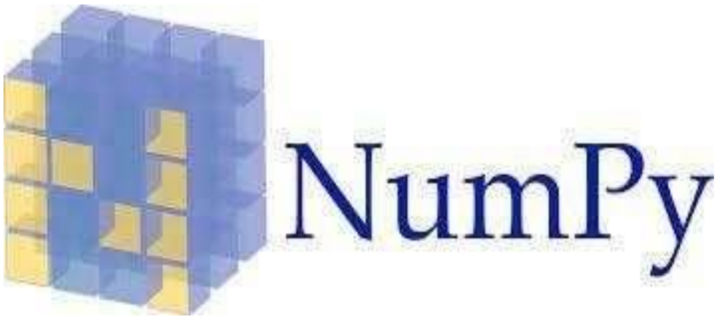


Fig-1.4: NumPy

- **PANDAS:**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Fig-1.5: Pandas

• Data Frame object for data manipulation with integrated indexing.

• Tools for reading and writing data between in-memory data structures and different file formats.

• Data alignment and integrated handling of missing data.

• Reshaping and pivoting of data sets.

• Dataset merging and joining

- **Scikit-learn:**

Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. Components of scikit-learn:

• Supervised learning algorithms.

• Cross-validation.

• Unsupervised learning algorithms.

• Various toy datasets.

• Feature extraction.

- **Sklearn. Preprocessing:**

The sklearn. preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. In general, learning algorithms benefit from standardization of the data set. From the sklearn preprocessing library we are importing Min Max Scalar function. MinMaxScaler preserves the shape of the original distribution. It doesn't meaningfully change the information embedded in the original data. Note that MinMaxScaler doesn't reduce the importance of outliers. The default range for the feature returned by MinMaxScaler is 0 to 1.

- **sklearn. Model_selection:**

Model_selection is a method for setting a blueprint to analyze data and then using it to measure new data. Selecting a proper model allows you to generate accurate results when making a prediction. To do that, you need to train your model by using a specific dataset. Then, you test the model against another dataset. From this package we have imported train_test_split function.

1.2 Algorithm:

Logistic regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

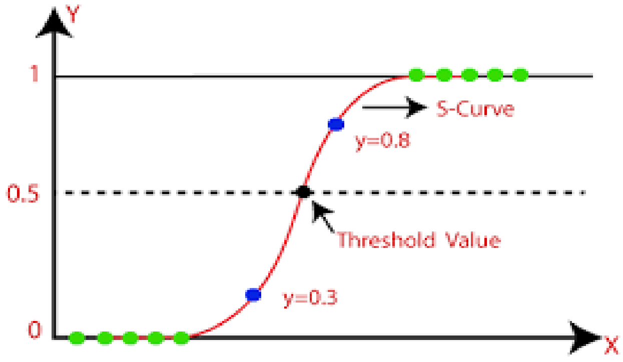


Fig-1.6LogisticRegression

CHAPTER-2

REVIEW OF RELEVENT LITERATURE

2.1. Educational Game (Detecting personality of players in educational game):

One of the goals of Educational Data Mining [1] is to develop the methods for student modelling based on educational data, such as; chat conversation, class discussion, etc. On the other hand, individual behaviour and personality play a major role in Intelligent Tutoring Systems (ITS) and Educational Data Mining (EDM). Predicting personality traits has been a subject of interest across various fields, including psychology, sociology, computer science, and even marketing. The literature on personality prediction spans a wide range of methodologies, from traditional psychological assessments to modern computational approaches leveraging big data and machine learning. Traditional psychological theories, such as the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism), have formed the basis for personality assessment for decades. Researchers like Costa and McCrae have extensively studied and validated these traits, which has influenced a multitude of studies aiming to predict personality based on self-reported surveys, observations, and behavioral assessments. Thus, to develop a user adaptable system, the student's behaviour that occurring during interaction has huge impact EDM and ITS. In this chapter, we introduce a novel data mining techniques and natural language processing approaches for automated detection student's personality and behaviour in an educational game (Land Science).where students act as interns in an urban planning firm and discuss in groups their ideas. In order to apply this framework, input excerpts must be classified into one of six possible personality classes. We applied this personality classification method using machine learning algorithms, such as: Naive Bayes, Support Vector Machine (SVM) and Decision Tree. There are different traditional and manual techniques for the detection of psychopathy, and these are listed as follows: the Hare Psychopathy Checklist, matrix linguistics analysis, Welsh Anxiety Scale, Psychopathy Checklist-Revised, Wide Range Achievement Test, Self-Report Psychopathy Test, Wonderlic Personnel Test, Levenson Self-Report Psychopathy Scale, and Dictionary of Affect in Language However, online users of social network websites express themselves using texts, images, public profiles.

2.2 Using Twitter Content to Predict Psychopathy:

An ever-growing number of users share their thoughts and experiences using the Twitter micro logging service. Although sometimes dismissed as containing too little content to convey significant information, these messages can be combined to build a larger picture of the user posting them. One particularly notable personality trait which can be discovered this way is psychopathy: the tendency for disregarding others and the rule of society. In this paper, we explore techniques to apply data mining towards the goal of identifying those who score in the top 1.4% of a well known psychopathy metric using information available from their Twitter accounts.

We apply a newly proposed form of ensemble learning, Select RUS Boost (which adds feature selection to our earlier imbalance aware ensemble in order to resolve high dimensionality), employ four classification learners, and use four feature selection techniques. The results show that when using the optimal choices of techniques, we are able to achieve an AUC value of 0.736. Furthermore, these results were only achieved when using the Select RUS Boost technique, demonstrating the importance of feature selection, data sampling, and ensemble learning. Overall, we show that data mining can be a valuable tool for law enforcement and others interested in identifying abnormal psychiatric states from Twitter data. R. Wald et. al. have used social media like twitter contents to identify human psychology. They said Twitter, a micro blogging site, is used by a number of users to share their experiences and thoughts about their day-to-day life. Although researchers have often discarded the method of predicting personality by analyzing the tweets because they are of the view that it contains very little content to predict significant information, but these tweets can be combined to make a larger picture of the user who is posting them. Select RUS Boost, a new form of ensemble learning has been used to predict psychopathy using Twitter, which uses four classification learners and four feature selection techniques.

2.3 An Examination of Online learning effectiveness using data mining

Nurbiha A Shukora have given the concept of Online learning which became highly popular because of technological advancement that made it possible to have discussions even from a distance. Most studies have reported that have been conducted report how effective online learning has helped students to improve their learning power while assessing the learning process Simultaneously. This kind of discussion can be possible only by applying data mining technique where we can access the different experiences of students which they filed online on basis of their log files. However , it is said that students should put more hard work to become an excellent online learner.

CHAPTER-3

METHODOLOGY

Data Collection

First step for prediction system is data collection and deciding about the training and testing dataset. In this project we have imported dataset from Kaggle website which includes 70% of training dataset and 30% of testing dataset. Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. [3]A researcher can evaluate their hypothesis on the basis of collected data. In most cases, data collection is the primary and most important step for research, irrespective of the field of research. The approach of data collection is different for different fields of study, depending on the required information.

Training Dataset

In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Here, you have the complete training dataset. You can extract features and train to fit a model and so on.

Testing Dataset

Here, once the model is obtained, you can predict using the model obtained on the training set. Some data may be used in a confirmatory way, typically to verify that a given set of input to a given function produces some expected result. Other data may be used in order to challenge the ability of the program to respond to unusual, extreme, exceptional, or unexpected input.

Attribute Selection

Attribute of dataset are property of dataset which are used for system and for personality many attributes are like heart gender of the person, age of the person ,Big five traits like Openness Neuroticism, Extraversion, Agreeableness, Consciousness(value 1 -10). The importance of feature selection can best be recognized when you are dealing with a dataset that contains a vast number of features. This type of dataset is often referred to as a high dimensional dataset. Now, with this high dimensionality, comes a lot of problems such as - this high dimensionality will significantly increase the training time of your machine learning model, it can make your model very complicated which in turn may lead to Overfitting.

Pre-Processing of Data

Pre-processing needed for achieving best result from the machine learning algorithms. In this, we gathered dataset and it was pre-processed before it is sent to training stage. Sampling is a very common method for selecting a subset of the dataset that we are analyzing . In most cases ,working with the complete dataset can turn out to be too expensive considering the memory .Using a sampling algorithm can help us reduce the size of the dataset to a point where we can use a better, but more expensive, machine learning algorithm. When we talk about data, we usually think of some large data sets with huge number of rows and columns. While that is a likely scenario, it is not always the case — data could be in so many different forms: Structured Tables ,Images, Audio files, Videos etc. Machines don't understand free text, image or video data as it is, they understand 1s and 0s. So we pre-process the data. The biggest step in data analytics is data preprocessing; it plays a vital role in reducing several problems. Through preprocessing partially filled columns are eliminated. the data exploration summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data, and other attributes.

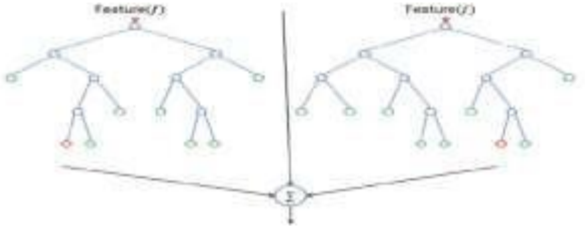


Fig-3.1: Decision tree graph

Prediction of Personality Classification

In this, system we used logistic regression supervised machine learning algorithm it is given best accuracy score for personality prediction. By applying all the modules finally the personality of a person is predicted and the final result will be the personality of a person by using the training and testing dataset the personality of a person is classified.

3.1 EXISTING METHODOLOGY

3.1.1 KNN

k- Nearest Neighbor model is one the simplest but most effective models. In this model, the class label of the test datasets on the basis of the class label of the neighboring training data elements. The similarity between two elements is measured using Euclidean Distance. It is also known as an Instance learning or Lazy model. The value of 'k' is calculated which actually the number of is nearest neighbors that have to be considered. A suitable value for 'k' should be chosen. An appropriate distance metric is also a requirement. Sometimes, the 'Minkowski' distance may be used. It is a generalization of the Euclidean and Manhattan distance. Mathematically, it is can be represented as

$$d\left(x^{(i)}, x^{(j)}\right) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p}$$

3.1.2 Decision Tree

This is one of the most widely used predictive modeling approaches. As per the name of the model, this is built in the form of a tree like structure. This model maybe used in case of a multi-dimensional analysis where there are multiple classes present. The past data also known as the past vector is used to create a model that can be used to predict the value of the output based on the input being provided. There are multiple nodes in a tree and each node corresponds to one or the other vector. The tree terminates at a leaf node where each such node represents a possible outcome or output.

3.1.3 PROPOSED ALGORITHM

Since the above stated algorithms have been already used in many such research that revolves around the personality prediction, We will be proposing a Ensemble algorithm in this section and try to achieve our main aim of attaining a favorable percentile score of accuracy of this proposed model. This Ensemble model is a combination of KNN algorithm and Logistic Regression Algorithm.

If a model possesses two qualities, it may be classified as an ensemble learning approach:

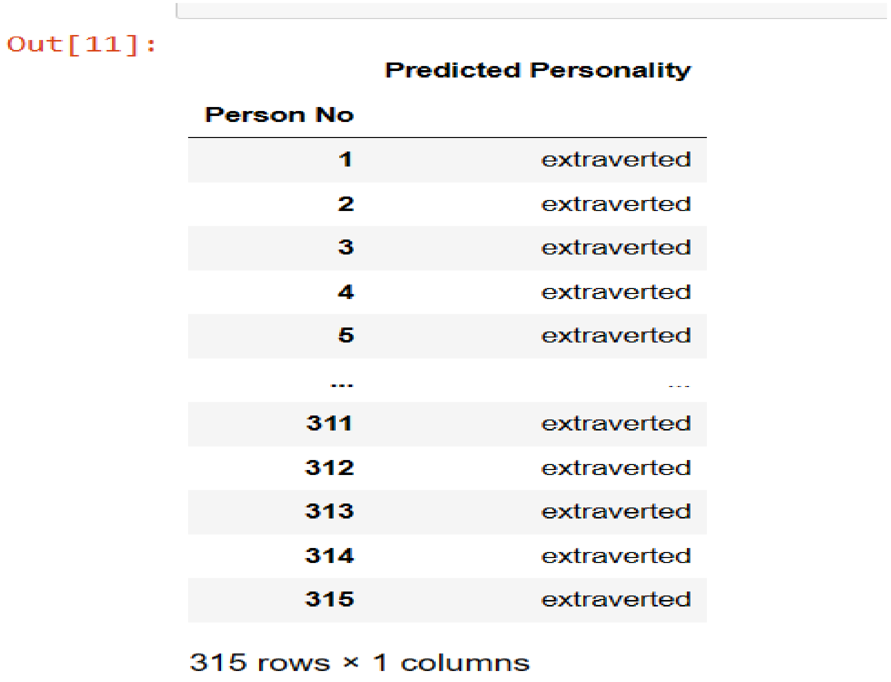
1. Comprising two or more Weaker models.

1. Prediction of two or more models are comprised.

The Final goal of Ensemble technique is to improve the result than the previous results of weaker models that are combined to form a hybrid approach. Although a less important objective can be to increase the model's stability. There are majorly two ways of achieving ensemble machine learning algorithm which are Bootstrap Aggregating and Boosting. We will be using both these methods in our system for getting the desired results.

CHAPTER-4

RESULTS AND DISCUSSION



```
Out[11]:
                     Predicted Personality
 Person No
          1                    extraverted
          2                    extraverted
          3                    extraverted
          4                    extraverted
          5                    extraverted
        ...                            ...
        311                    extraverted
        312                    extraverted
        313                    extraverted
        314                    extraverted
        315                    extraverted

315 rows × 1 columns
```

Fig- 4.1 Output

The model will predict the personality of a person basing on the rating given to the every personality trait of a person in a particular range (1-8), the system will decide the personality of a person based on highest rating given to the personality trait and decides the Personality factors in a traits.

CHAPTER-5

CONCLUSIONS AND FUTURE SCOPE OF STUDY

This project, we discuss about how the personality is identified using different classification algorithms. Here we study relationship between user and his/her personality. In this we used logistic regression because it gives best accuracy around 86.53% while compare to other algorithms that are used previously like naive Bayes, SVM , Logistic regression is fast and give accurate results compared to other algorithms. Personality system is used in E-commerce sites, in Competitive exams , Psycho metric tests , matrimonial sites , Government sectors like army, navy, Air force..Thus the personality is automatically classified by the system after user attempts the survey by the data set provided in the back end .Personality analysis and prediction is more recent times so further in future more personality traits can be added. Further any improvement can be done using the data set and algorithms to improve the accuracy. [4]The research examines the literature on the usage of social media framework as a behavioral feature study by looking at the link between users' personalities and their social network behaviour. In this research we already considered all the possibilities to improve the accuracy to the best. As a result, the suggested machine learning model was created to predict personality with a better degree of accuracy[2]. Many emotive and personally detailed materials have been revealed on social media sites. In addition, the suggested technique uses language translators to analyze every non-English content extraction. While analyzing the personality, this combination of KNN and Logistic Regression algorithms offered high prediction and classification results. The KNN method is probabilistic and outperforms all other machine learning algorithms in terms of prediction rate. Furthermore, Logistic Regression has a greater classification accuracy and a lower classification error rate. By leveraging improved accuracy and minimal relative error in the classification, the hybrid version of these machine learning algorithms give good prediction rate in the personality prediction paradigm. Human resources working in numerous areas of the ICT industry would profit from this in their recruiting process. Using the suggested technique, they can more accurately applicant

REFERENCES

[1]Machine Learning in Data Science using python-Dr.R.Nageshwar

Rao. https://youtube.com/watch?v=Cavebr_NNq8&feature=share

https://www.verywellmind.com/the-big-five-personality-dimensions-2795422

[2] J. Staiano, B. Lepri, N. Aharony, F. Pianesi, N. Sebe, and A. Pentland, "Friends don't lie: Inferring personality traits from social network structure," in Proc. ACM Conf. Ubiquitous Comput., 2012, pp. 321-330

[3]Paschen, J., Wilson, M., and Ferreira, J. (in press). Collaborative intelligence: How human. and artificial intelligence create value along the B2B funnel. Business Horizons.

[4]Herman, A., O'Boyle, E. (2012). The best and the rest: Revisiting the norm of normality in individual performance. Personnel Psychology, 65(1), 79e119.

[5]We Referd From IEEE Xplore and Kaggle.