

PERSONALITY PREDICTION SYSTEM
REPORT
A MINI PROJECT REPORT

Submitted by
KULWANT SINGH DHAMA
(18BCS039)
&
KANJIKA
(18BCS038)

Bachelors of Technology
IN
COMPUTER SCIENCE & ENGINEERING



SCHOOL OF COMPUTER SCIENCE & ENGINEERING
SHRI MATA VAISHNO DEVI UNIVERSITY
DECEMBER, 2020

SHRI MATA VAISHNO DEVI UNIVERSITY

CERTIFICATE

Certified that this project report “PERSONALITY PREDICTION SYSTEM” is the work of KULWANT SINGH DHAMA (Entry Number 18BCS039) and KANIKA (Entry Number 18BCS038), School of Computer Science Engineering who carried out the mini project work under my supervision.

DR. MANOJ KUMAR GUPTA

HEAD OF DEPARTMENT &
ASSOCIATE PROFESSOR SCHOOL
OF COMPUTER SCIENCE &
ENGINEERING

Submitted to the Viva-voce Examination held on

INTERNAL EXAMINER

EXTERNAL EXAMINER

ABSTRACT

Personality refers to individual differences in characteristic patterns of thinking, feeling and behaving. Currently, there isn't any focus on the mental health of a student in India. So, to make it easier for teachers/counsellors/parents to understand the psychological status of their children and spread awareness, we will be building an application that could serve the purpose. We are going to implement machine learning algorithms to calculate the Big-Five personality factors or OCEAN factors, to eliminate the discreteness present in the current calculation system of personality factors. This application can help teachers/counsellors/parents to provide detailed results on a student's personality instantaneously. This project will be helpful in India's education system, by making it more effective and more focused towards student's wellbeing.

ACKNOWLEDGEMENT

It is our advantage to acknowledge with a deep sense of gratitude to our project guide Dr. Manoj Kumar Gupta whose supervision, inspiration and valuable discussion have helped us to complete our project. The project could not have been implemented without constant inputs and encouragement from him.

We would also like to thank Ms. Aman Bali who guided us for all the psychology-related tasks. Last but not least, this acknowledgement would be incomplete without rendering our sincere gratitude to all those who have helped us in the completion of this project.

TABLE OF CONTENTS

CONTENTS	PAGE NUMBER
a. LIST OF FIGURES	(6)
b. ABBREVIATIONS	(7)
1. INTRODUCTION	
1.1. INTRODUCTION	(8)
1.2 PROBLEM MOTIVATION	(9)
1.3. PROBLEM STATEMENT	(9)
1.4 ORGANISATION OF REPORT	(9)
2. REVIEW OF LITERATURE	(10)
3. IMPLEMENTATION & ANALYSIS	
3.1 REQUIREMENTS	(15)
3.2 DATASET	(16)
3.3 WORKING & APPROACH	(16)
3.4 RESULTS	(20)
4. CONCLUSION	(25)
5. FUTURE GOALS	(26)
6. APPENDIX	
6.1. APPENDIX A	(27)
6.2. APPENDIX B	(28)
7. REFERENCES	(36)
8. PLAGIARISM REPORT	(37)

LIST OF FIGURES

FIGURE	TITLE	PAGE NUMBER
Fig1	Workflow diagram of application	18
Fig2	Workflow diagram of ML models	19
Fig3	Application's home page	21
Fig4	Application's questions page	22
Fig5	Application's result page	23
Fig6	Application's about page	24
Fig7	Accuracy score	25
Fig8	Preview of the dataset	30

ABBREVIATIONS & SYMBOLS

%	Percentage
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
ML	Machine Learning
AI	Artificial Intelligence

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Personality is characterized as the standard of conduct and the trademark perspective and sensation of an individual. In the psychological field, character is ordered in five qualities that are Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. These characteristics are the essential factors that recognize the character of an individual or a person. Openness characterizes that how much an individual is straightforward and prepared to discuss their sentiments to their companions and associates. It additionally shows how an individual grasps groundbreaking thoughts and curiosity. Conscientiousness is that how much an individual shows to mind and work altogether. The individual having high uprightness factor is normally the person who is coordinated and are proficient. Extraversion is the factor that shows whether the individual is a lot of social or likes to converse with individuals rather than favor being separated from everyone else or disconnected. Suitability is one of the five that shows the individual is warm and careful. People having a high factor of appropriateness shows that they are agreeable and coexist well with others. Neuroticism is the wide character characteristic as the propensity of an individual to be in uneasiness, dissatisfaction, desire, discouraged temperament and depression. This attribute in brain science is the higher-request character quality. In psychology and development, these characteristics are utilized to discover the character and nature of an individual and how and why they do things the manner in which they do. The varieties in individuals are a direct result of these elements.

1.2. PROBLEM MOTIVATION

Every individual has diverse character qualities to various people or soundness that shows conduct in various circumstances. Today, the majority of individuals are impulsive and are battling with the proper behaviour in troublesome circumstances. Character additionally can help guardians, educators and instructors to guide and shape conduct towards a reasonable character.

1.3. PROBLEM STATEMENT

The fundamental goal of this project is to construct an application that performs personality factor prediction using ML algorithms.

1.4 ORGANISATION OF REPORT

The report for the project is structured in the form of chapters, wherein each chapter aims to describe in detail a certain aspect of the project. The chapters are broadly divided into 4 parts (i) Introduction (ii) Implementation & Analysis (iii) Conclusion (iv) Future Goals.

CHAPTER 2

REVIEW OF LITERATURE

Characteristic hypotheses of character have since quite a while ago endeavoured to nail down precisely the number of character attributes exist. Prior speculations have proposed a different number of potential characteristics, including Gordon Allport's rundown of 4,000-character attributes, Raymond Cattell's 16-character variables, and Hans Eysenck's three-factor hypothesis.

However, numerous analysts felt that Cattell's hypothesis was too confounded and Eysenck's was too restricted in extension. Subsequently, the five-factor hypothesis arose to portray the fundamental qualities that fill in as the structure squares of character.

Note that every one of the five-character factors speaks to a reach between two limits. For instance, extraversion speaks to a continuum between extraordinary extraversion and outrageous introspection. In reality, the vast majority lie someplace in the middle of the two polar finishes of each measurement.

These five categories are usually described as follows.

- Openness to experience: in some cases, called mind or creative mind, this speaks to the readiness to attempt new things and consider new ideas. Qualities incorporate smarts, innovation and interest.
- Conscientiousness: the desire to be careful, diligent and to regulate immediate gratification with self-discipline. Conscientiousness include discipline, consistency, ambition and reliability.

- Extraversion: a state where an individual draws energy from others and looks for social associations or communication, instead of being distant from everyone else (introversion). Qualities incorporate being active, fiery and sure.
- Agreeableness: the proportion of how an individual interfaces with others, described by the level of sympathy and co-activity. Characteristics incorporate affability, graciousness and reliability.
- Neuroticism: an inclination towards antagonistic character qualities, enthusiastic insecurity and reckless reasoning. Characteristics incorporate cynicism, tension, uncertainty and dreadfulness.

The model was created between the 1950s and 1990s by a few unique analysts, coming full circle in a 6-2-1 system that surveys every character characteristic on two features and six sub-aspects. Altogether there are five attributes, 10 aspects and 30 sub-features that somebody can be surveyed on.

All speculations of character, the five-factor model is affected by nature and sustain both. Two investigations have discovered that the measure of fluctuation that credited to qualities of the Big Five attributes is 40-60%.

An examination was directed by Jang (1996) with 123 sets of indistinguishable twins and 127 sets of intimate twins. he assessed the heritability of scruples, suitability, neuroticism, receptiveness to experience, and extraversion to be 44%, 41%, 41%, 61%, and 53%, individually.

This finding was like the discoveries of another investigation, where the heritability of principles, suitability, neuroticism, receptiveness to experience and extraversion were assessed to be 49%, 48%, 49%, 48%, and half, separately.

The Big Five model came about because of the commitments of numerous free specialists. Gordon Allport and Henry Odbert first framed elite of 4,500 terms identifying with character attributes in 1936 (Vinney, 2018). Their work gave the establishment to different clinicians to start deciding the essential components of character.

In the 1940s, Raymond Cattell and his partners utilized factor examination (a factual technique) to limit Allport's rundown to sixteen attributes. Nonetheless, various clinicians analyzed Cattell's rundown and discovered that it very well may be additionally diminished to five characteristics. Among these therapists were Donald Fiske, Norman, Smith, Goldberg, and McCrae and Costa (Cherry, 2019).

Five essential components of character were pushed vigorously by Lewis Goldberg. McCrae and Costa developed his work, and they affirmed the model's legitimacy and given the model utilized today: reliability, pleasantness, neuroticism, receptiveness to experience, and extraversion. The model got known as the "Huge Five" and has seen gotten a lot of consideration. It has been investigated across numerous populaces and societies and keeps on being the most generally acknowledged hypothesis of character today.

Every one of the Big Five-character characteristics speaks to incredibly general classes which cover numerous character-related terms. Every characteristic incorporates a large number of different aspects. Extraversion classification contains marks like amiable, strong, vivacious, daring, eager, active.

During the early piece of the twentieth century, analysts got keen on arrangement how characters create and why they contrast between individuals. Somewhere in the range of 1923 and 1928, the American Psychological Association held various shows on the subjects of character a lot. In 1928, a unique issue of the *Journal of Abnormal and Social Psychology* was given to the character.

The main issue of *Character and Personality* turned out in 1932. This new diary needed to meld British and American information on individual contrasts between individuals with German investigations of individuals' characters. The absolute first issue attempted to figure out what the contrast among "character" and "character" was, in any case. The diary had acclaimed donors, similar to Alfred Adler and Carl Jung, and it included everything from contextual analyses to experimentation. In 1937, American analyst Gordon Allport distributed a book called *Personality: A Psychological Interpretation*. Allport needed to characterize and arrange character brain science. His book made a dream for how the investigation of character could

be set inside the sociologies.

Henry Murray was another significant figure in the field. An American therapist who extended the limits of character brain research and brain research through his utilization of experimentation, he showed the point to various alumni understudies who might grow the field considerably further in years to come. Murray was so notable as a specialist on the character that the U.S. government enrolled him to make a mental profile of Hitler. Murray built up the hypothesis that an individual's response to explicit components inside their current circumstance is the thing that shapes their character.

The advanced feeling of individual character is an aftereffect of the movements in culture beginning in the Renaissance, a basic component in innovation. Conversely, the Medieval European's self-appreciation was connected to an organization of social jobs: "the family unit, the family relationship organization, the society, the enterprise – these were the structure squares of personhood". Stephen Greenblatt notices, in relating the recuperation (1417) and profession of Lucretius' sonnet *De Rerum Natura*: "at the centre of the sonnet lay key standards of a cutting-edge comprehension of the world." "Subject to the family, the individual alone was nothing," Jacques Gélis notices. "The trademark characteristic of the cutting-edge man has two sections: one inner, the other outside; one managing his current circumstance, the other with his mentalities, qualities, and emotions." Rather than being connected to an organization of social jobs, the advanced man is to a great extent affected by the ecological factors, for example, "urbanization, training, mass correspondence, industrialization, and politicization."

Crystal gazing, the old Sastra got from Vedangas, additionally talks about character and season of birth. It isn't certain whether soothsaying is causative of a specific character characteristic or an assortment of attributes or whether it affects human character by any stretch of the imagination.

Since antiquated occasions, people have tried to clarify conduct by sorting characters into particular kinds. Character appraisals have been created in the course of recent hundreds of years to portray parts of an individual that stay stable all through

a lifetime: the person's character example of conduct, considerations, and sentiments. Character evaluations have been utilized to sort, group, and classify individuals. References to character appraisals have even advanced into books and motion pictures, for example, Harry Potter, who was sent into the Gryffindor House at the Hogwarts School by an "arranging cap" that could measure the demeanour of every understudy.

CHAPTER 3

IMPLEMENTATION & ANALYSIS

This chapter of the Report focusses on the implementation means taken by the author to reach to the results of the project. The minimum requirements and workflow along with a proposed methodology in detail has been explained in the later sections.

3.1 REQUIREMENTS

The following Software and Hardware Requirements are the minimum requirements before starting with this project.

3.1.1 Software:

- Programming Language: R, Shiny, R markdown.
- Software: RStudio
- Libraries: RCurl, caret, FNN, class, gmodels, splitstackshape, tidyverse, shiny, ggplot2, data.table, shinyWidgets, knitr, shinythemes.

3.1.2 EXPERIMENTAL SETUP:

- Operating System: Windows 10 (64-bit)
- Hardware Requirements: Processor i9-9880H
- RAM: 16 GB
- GPU: 4 GB
- Internal Storage: 1 TB SSD

3.2 DATASET

The dataset chosen contains the questionnaire and metadata from Kaggle, openpsychometrics.org.

This dataset includes attributes as explained: reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information), and links (also viewed/also bought graphs). The key which is used to evaluate the data and calculate the result for the respective columns is from the Goldberg, Lewis R. "The development of markers for the Big-Five factor structure." Psychological assessment 4.1 (1992): 26. Mentioned in Appendix A4

3.3 WORKING & APPROACH

This project is divided into two major components:

- Application
- ML model

3.3.1 WORKFLOW DIAGRAM:

3.3.1.1 Application

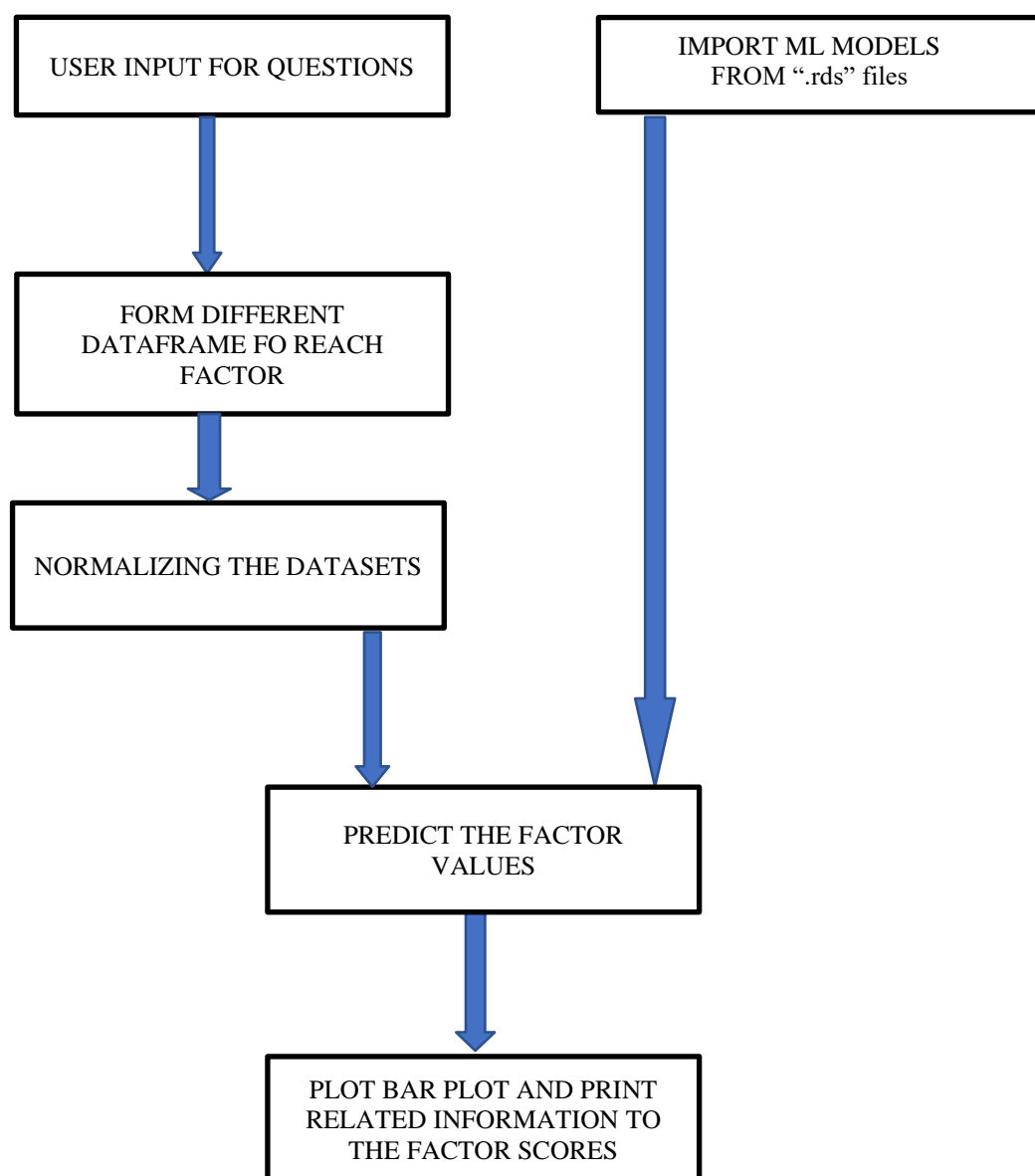


Fig1 (Workflow diagram of application)

3.3.1.2 ML Models

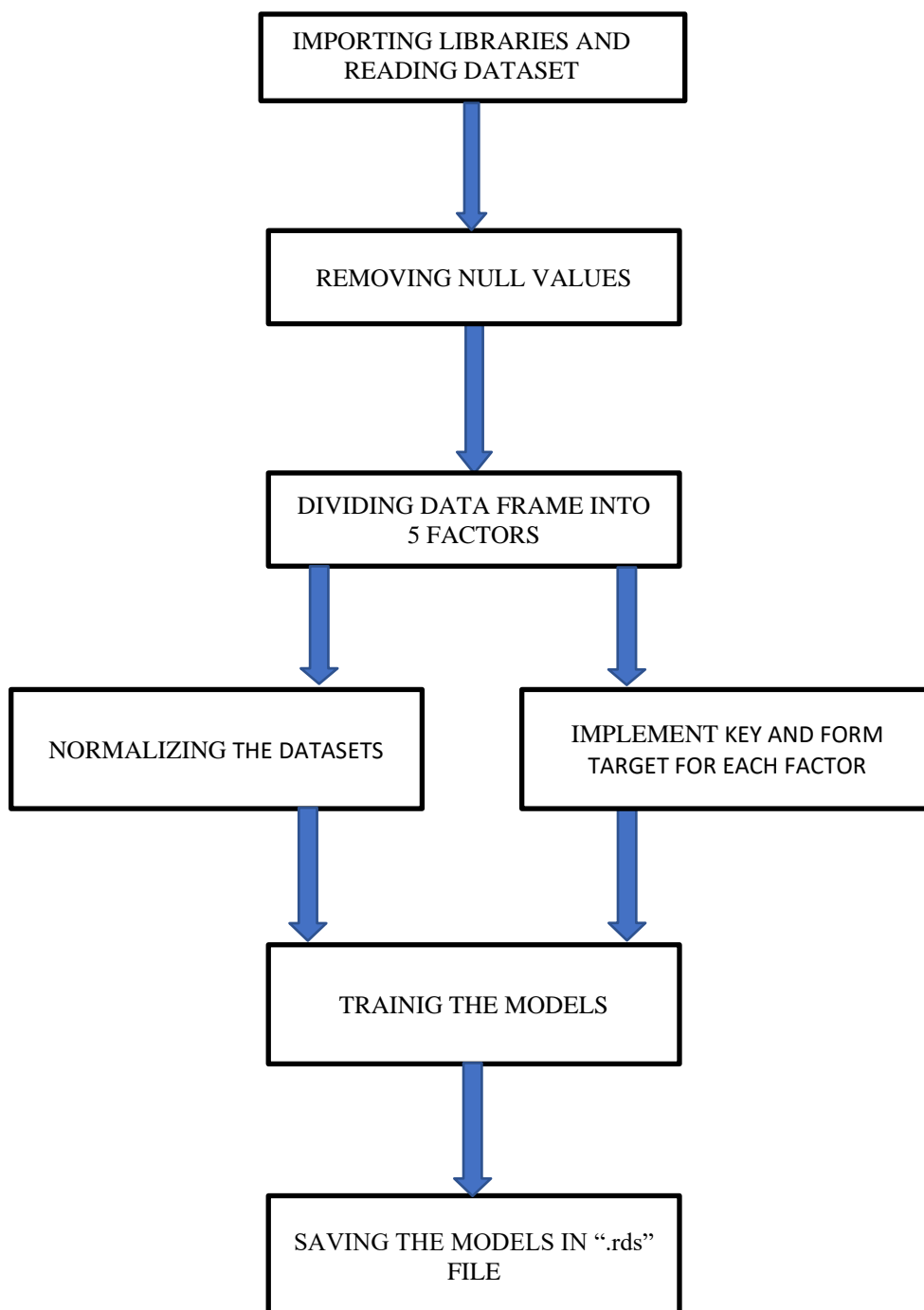


Fig2 (Workflow diagram of ML models)

3.3.2 PROPOSED METHODOLOGY:

3.3.2.1 ML Models:

1. Import required libraries and functions.
2. Read dataset.
3. Subset only first 50 columns from the dataset.
4. Remove rows containing NA, NaN, NULL.
5. Subset the dataset into 5 different factor datasets.
6. Select on the unique rows for each factor and remove the rest.
7. Make 2 copies of factor datasets:
 - a. Implement key
 - i. Divide the datasets into +ve and -ve according to the respective questions for each factor.
 - ii. Implement the scoring key for all the factors and form the target columns.
 - b. Normalize the factor datasets.
8. Train different model using Supervised Linear Regression algorithm for each factor, using the normalized datasets and the target column for the respective factors.
9. Save the models in the form of “.rds” file. (to be used in the application)

For more information, see Appendix B

3.3.2.2 Application

1. Home Page (page 1) – Introduction to the application and details of the OCEAN model that we are going to analyze their personality.
2. Questions (page 2) – List of 50 questions to be answered by the user to see the result.
3. Result (page 3) – Bar plot representing the score of reach factor and every factor score is explained in details below the bar plot.
4. About (page 4) – Contains reference links for the source of data, questions and models. There is also an explanation of the Big-Five factor theory.

3.4 Result

3.4.1 Application

3.4.1.1 Home Page

Personality Prediction System

[Home](#)
[Test](#)
[Result](#)
[About](#)

Welcome

Have you ever thought about who you are and why you do things the way you do ?

Here's the test that will reveal your personality !!!

This test takes about 8-10 minutes to finish.

There will be 50 questions consisting 5 options categorized as Strongly Disagree, Disagree, Neutral, Agree, Strongly agree. The test is based on the Five factor (OCEAN) model.

In psychology trait theory, the Big five personality traits, also known as the five-factor model and OCEAN or CANOE model, is a grouping for personality traits developed from the 1980s.

The five traits are Openess to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism, using these traits a person's most important characteristics are measured.

Openness

Openness to Experience, otherwise known as the 'creative' trait of The Big Five, describes the ability to be innovative and think outside of the box. People who are high in Openness tend to be comfortable with abstract ideas, and are drawn towards opportunities that allow you to flex your creativity. These individuals are usually optimistic with bold personalities, and have a thirst for a career that allows them to embrace these qualities in an exciting work environment. The need to keep learning and staying challenged is also important.

Conscientiousness

Conscientiousness personality types are known for their ability to remain calm under pressure, as well as being skilled in analysing complex situations and thinking up logical solutions. The Conscientiousness trait enjoys tackling new challenges that develop; utilising their pragmatic and logical mindset to achieve their goals.

Conscientiousness types have a great eye for change and innovation and are well-known for their high levels of thoughtfulness and superb attention to detail. Those who possess this personality trait enjoy working by a schedule in a work environment that accommodates their need to learn on the job and expand their knowledge with new skills.

Extraversion

People with a high level of extraversion are known for being assertive, persuasive and optimistic, with a high value for personal growth and the persistence to develop their skill set. They are social creatures who enjoy being the center of attention. They have hard-working natures and thrive in team environments.

Extroverts require a workplace that allows them to achieve their personal development goals. In contrast, people low in extroversion, (introverts), prefer to spend their time working on solo projects or with close friends, rather than being the life of the party.

These individuals are always on the lookout for new opportunities, and are very capable when working in leadership positions.

Agreeableness

Agreeable people are known for their kind, compassionate and friendly nature. These individuals require a career that allows them to support, attend to and care for others, as they long for a sense of purpose and are driven by their ability to make a difference to the world.

In an Agreeable work environment, there will be high levels of collaboration with plenty of opportunities that are challenging, yet very personally rewarding. People high in this trait are amazing at facilitating cooperation with warmth and tact.

Neuroticism

Neuroticism, is a trait that has many negative connotations attached as it relates to an individual's predisposition to demonstrating anxiety and negative feelings and emotions. People who score low on this scale are emotionally stable and confident.

Contrary to belief, healthy neurotics have an ability to combine high stress levels with high levels of attentiveness, which is a skill that most employers seek within a candidate. Individuals who identify with the Emotional trait are able to use the stress of a deadline or other work stressors to focus their mind and draw on their productivity, rather than allow these immense pressures to kill their drive.

Of course, this is all dependent on how that individual handles these stressors - high levels of neuroticism can cause stress in the workplace, which can then lead to both health and productivity issues.

So, Let's get started!

Start

Fig3 (Application's home page)

3.4.1.2 Questions Page

Personality Prediction System

Home

Test

Result

About

Question 1 - I get stressed out easily.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 2 - I feel little concern for others.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 3 - I am the most lively person at a party.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 4 - I am always prepared.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 5 - I have a good vocabulary.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

•

Question 45 - I spend time reflecting on things.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 46 - I often feel sad.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 47 - I make people feel at ease.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 48 - I am quiet around strangers.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 49 - I am very strict and specific about my work.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Question 50 - I am full of ideas.

☐ Strongly Disagree
☐ Disagree
☐ Neither Disagree nor Agree
☐ Agree
☐ Strongly Agree
☒ Prefer not to answer

Submit

Fig4 (Application's question page)

3.4.1.3 Result Page



Fig5 (Application's result page)

3.4.1.4 About Page

Personality Prediction System

Home

Test

Result

About

Big 5 Personality Traits

In psychological trait theory, the Big Five personality traits, also known as the five-factor model (FFM) and the OCEAN model, is a suggested taxonomy, or grouping, for personality traits, developed from the 1980s onwards. When factor analysis (a statistical technique) is applied to personality survey data, it reveals semantic associations: some words used to describe aspects of personality are often applied to the same person. For example, someone described as conscientious is more likely to be described as "always prepared" rather than "messy". These associations suggest five broad dimensions used in common language to describe the human personality and psyche.

- Openness**

This trait features characteristics such as imagination and insight. People who are high in this trait also tend to have a broad range of interests. They are curious about the world and other people and eager to learn new things and enjoy new experiences. People who are high in this trait tend to be more adventurous and creative. People low in this trait are often much more traditional and may struggle with abstract thinking.
- Conscientiousness**

Standard features of this dimension include high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines.
- Extraversion**

Extraversion (or extroversion) is characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness. People who are high in extraversion are outgoing and tend to gain energy in social situations. Being around other people helps them feel energized and excited. People who are low in extraversion (or introverted) tend to be more reserved and have less energy to expend in social settings. Social events can feel draining and introverts often require a period of solitude and quiet in order to "recharge."
- Agreeableness**

This personality dimension includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative while those low in this trait tend to be more competitive and sometimes even manipulative.
- Neuroticism**

Neuroticism is a trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in this trait tend to experience mood swings, anxiety, irritability, and sadness. Those low in this trait tend to be more stable and emotionally resilient.

Questions

Questions that are used in this test are taken from the official website of International Personality Item Pool (IPIP)

ML model

ML model used in the evaluation of the questions was a custom-build model, specifically designed for these specific questions.

This application was developed by Kulwant Singh Dhama, in 2020. For any query or suggestion, contact at dhamaks@gmail.com

Fig6 (Application's about page)

3.4.2 ML Models

The goal of this project was to determine the personality of an individual. Hence, accuracy will be the measure of evaluation, with it being defined as the proportion of the correctly determined reviews out of all the reviews that had been added to the test data. This model gives the highest accuracy of 97% at RMSE and R2 scales. The figure mentioned below represents the result obtained while performing with the algorithm and their accuracy.

##	Factors	RMSE	R2
## 1	Openness	0.2513535	0.9997130
## 2	Conscientiousness	0.2549866	0.9997806
## 3	Extraversion	0.2515616	0.9998501
## 4	Agreeableness	0.2509401	0.9997628
## 5	Neuroticism	0.2513456	0.9998294

Fig7 (Accuracy score)

4 CONCLUSION

The main objective of this project i.e., to predict the personality of an individual using Machine learning model which is tomorrow's future.

This was successfully performed on the IPIP dataset with a classification accuracy of 97 % approximately.

Personality Prediction was performed successfully using the Linear Regression Algorithm.

Thus, using this ML model the personality was predicted for an individual based on a test of 50 psychological questions and test is evaluated using LIKERT scale, so that one can understand the behavioural patterns and factors.

5 FUTURE GOALS AND APPLICATIONS

Since we were limited to use computing power, this application is limited to only predict the OCEAN factors, but with the use of correct AI/ML technique and more powerful computing resource, this application could be used to any the 16PF factors of a person, just by listening/reading text from a person. This project is combining the techniques of Machine Learning, Personality Prediction by using several algorithms it can be used in the future in predicting behavioural and thinking patterns of humans using more complicated ML/AI techniques. Its usefulness and applicability find not only in computer science/ applications but in the areas of research from psychology. It promises to revolutionize the research process. Personality is key to human. Every individual is conscious of their personal growth and professional success and how they behave and why they do things the way they do. Everyone wants to feel understood. This also helps in the growth of an individual shape as an adult and help in achieving their goals

6 APPENDIX

6.1 Appendix A

Important Links

[A1] The ML model is available at:

https://rpubs.com/Walker_2921/701660

[A2] This dataset used to train the models is available at:

https://openpsychometrics.org/_rawdata/

[A3] The scoring key for evaluation of data is available at:

https://sites.temple.edu/rtassessment/files/2018/10/Table_BFPT.pdf

[A4] The development of markers for the Big-Five factor structure by Goldberg, Lewis R.

<https://doi.apa.org/record/1992-25730-001?doi=1>

6.2 Appendix B

Since the project is designed with the motive to use Machine Learning, we are going to build a linear regression model and evaluate the dataset to predict the factor values and eliminate the discreteness present in the evaluation and libraries like RCurl, caret, FNN, class, models, etc.

We will conduct this analysis to implement the psychological analysis using the ML algorithms instead of the traditional scoring system used by psychiatrists.

We are going to build a Linear Regression model and evaluate the dataset to predict the factor values and eliminate the discreteness present in the evaluation.

The first step is to import the important libraries and some helpful functions.

```
library(caret)
library(RCurl)
library(FNN)
library(class)
library(gmodels)
library(splitstackshape)
library(tidyverse)
normalize <-function(x) {
  x <- x/5
  x <- format (x, digits = 2, nsmall = 1)
  return (x)
}
```

Now we will read the dataset:

```
data_set <- read.csv ("dataset.csv", header = TRUE)
```

Filter out first 50 columns and remove rows containing NaN, NA and NULL from the dataset.

```
i <- seq (1, 50)
data_set [, i] <- apply (data_set [, i], 2, function(x) as.numeric (as.character(x)))
data_set <- na.omit(data_set)
head(data_set)
```

The next step is to Preview the Dataset-

As you can observe that our dataset has 50 columns. Each column represents the responses by users for a specific question. 5 different column names specify which factor the respective question belongs to.

```
head(data_set)
```

```
##   OPN1 OPN2 OPN3 OPN4 OPN5 OPN6 OPN7 OPN8 OPN9 OPN10 CSN1 CSN2 CSN3 CSN4 CSN5
## 1    5    1    4    1    4    1    5    3    4    5    3    4    3    2    2
## 2    1    2    4    2    3    1    4    2    5    3    3    2    5    3    3
## 3    5    1    2    1    4    2    5    3    4    4    4    2    2    2    3
## 4    4    2    5    2    3    1    4    4    3    3    2    4    4    4    1
## 5    5    1    5    1    5    1    5    3    5    5    5    1    5    1    3
## 6    5    1    5    1    3    1    5    4    5    2    3    2    4    1    3
##   CSN6 CSN7 CSN8 CSN9 CSN10 EXT1 EXT2 EXT3 EXT4 EXT5 EXT6 EXT7 EXT8 EXT9 EXT10
## 1    4    4    2    4    4    4    1    5    2    5    1    5    2    4    1
## 2    1    3    3    5    3    3    5    3    4    3    3    2    5    1    5
## 3    3    4    2    4    2    2    3    4    4    3    2    1    3    2    5
## 4    2    2    3    1    4    2    2    2    3    4    2    2    4    1    4
## 5    1    5    1    5    5    3    3    3    3    5    3    3    5    3    4
## 6    2    4    3    4    3    3    3    4    2    4    2    2    3    3    4
##   AGR1 AGR2 AGR3 AGR4 AGR5 AGR6 AGR7 AGR8 AGR9 AGR10 NEU1 NEU2 NEU3 NEU4 NEU5
## 1    2    5    2    4    2    3    2    4    3    4    1    4    4    2    2
## 2    1    4    1    5    1    5    3    4    5    3    2    3    4    1    3
## 3    1    4    1    4    2    4    1    4    4    3    4    4    4    2    2
## 4    2    4    3    4    2    4    2    4    3    4    3    3    3    2    3
## 5    1    5    1    5    1    3    1    5    5    3    1    5    5    3    1
## 6    2    3    1    4    2    3    2    3    4    4    3    4    3    2    2
##   NEU6 NEU7 NEU8 NEU9 NEU10
## 1    2    2    2    3    2
## 2    1    2    1    3    1
## 3    2    2    2    1    3
## 4    2    2    2    4    3
## 5    1    1    1    3    2
## 6    1    2    1    2    2
```

Fig8 (Preview of the dataset)

Now, the Data Preprocessing will take place: -

To reduce computing cost and still generate a robust ML model, we will be forming 5 data frames from the original dataframe and then removing the duplicate responses for a specific factor.

This will eliminate the possibility of underfitting the model.

```
opn <- data_set [! duplicated (data_set [, 1:10]), 1:10]
csn <- data_set [! duplicated (data_set [, 11:20]), 11:20]
ext <- data_set [! duplicated (data_set [, 21:30]), 21:30]
agr <- data_set [! duplicated (data_set [, 31:40]), 31:40]
neu <- data_set [! duplicated (data_set [, 41:50]), 41:50]

n_opn <- nrow(opn)
n_csn <- nrow(csn)
n_ext <- nrow(ext)
n_agr <- nrow(agr)
n_neu <- nrow(neu)
```

Dividing each factor dataframe into two dataframes that is Positive and Negative dataframes for key evaluation.

Each question in the test is either considered +ve valued or -ve valued, and we are going to group all the +ve and -ve questions for each factor to evaluate them separately.

```
opn_p <- opn [, c (1, 3, 5, 7, 8, 9, 10)]
opn_n <- opn [, c (2, 4, 6)]
csn_p <- csn [, c (1, 3, 5, 7, 9, 10)]
csn_n <- csn [, c (2, 4, 6, 8)]
ext_p <- ext [, c (1, 3, 5, 7, 9)]
ext_n <- ext [, c (2, 4, 6, 8, 10)]
agr_p <- agr [, c (2, 4, 6, 8, 9, 10)]
agr_n <- agr [, c (1, 3, 5, 7)]
neu_p <- neu [, c (2, 4)]
neu_n <- neu [, c (1, 3, 5, 6, 7, 8, 9, 10)]
```

We are going to train 5 different models, one for each factor to achieve higher accuracy for each model.

But first we need to form the TARGET columns for each column by evaluating according to the scoring key of the model.

```
opn$result <- 0
csn$result <- 0
ext$result <- 0
agr$result <- 0
neu$result <- 0
```

```

for (i in seq (max (n_opn, n_csn, n_ext, n_agr, n_neu))) {

  if (i <= n_opn) {

    opn [i, "result"] <- sum (opn_p [i,]) - sum (opn_n [i,]) + 8

  }

  if (i <= n_csn) {

    csn [i, "result"] <- sum (csn_p [i,]) - sum (csn_n [i,]) + 14

  }

  if (i <= n_ext) {

    ext [i, "result"] <- sum (ext_p [i,]) - sum (ext_n [i,]) + 20

  }

  if (i <= n_agr) {

    agr [i, "result"] <- sum (agr_p [i,]) - sum (agr_n [i,]) + 14

  }

  if (i <= n_neu) {

    neu [i, "result"] <- sum (neu_p [i,]) - sum (neu_n [i,]) + 38

  }

}

```


Since the scoring is done on a scale of out of 40, we are converting the results out of 100. Forming the Final dataframes for each factor with their respective result values.

```
opn$result [opn$result > 40] <- 40  
opn$result [opn$result < 0] <- 0  
opn$result <- ceiling (opn$result * 2.5)
```

```
csn$result [csn$result > 40] <- 40  
csn$result [csn$result < 0] <- 0  
csn$result <- ceiling (csn$result * 2.5)
```

```
ext$result [ext$result > 40] <- 40  
ext$result [ext$result < 0] <- 0  
ext$result <- ceiling (ext$result * 2.5)
```

```
agr$result [agr$result > 40] <- 40  
agr$result [agr$result < 0] <- 0  
agr$result <- ceiling (agr$result * 2.5)
```

```
neu$result [neu$result > 40] <- 40  
neu$result [neu$result < 0] <- 0  
neu$result <- ceiling (neu$result * 2.5)
```

Normalize the feature columns in all dataframes

```
opn_a <- as.data.frame(lapply(opn[1:10], normalize))
opn_a$result <- opn$result
csn_a <- as.data.frame(lapply(csn[1:10], normalize))
csn_a$result <- csn$result
ext_a <- as.data.frame(lapply(ext[1:10], normalize))
ext_a$result <- ext$result
agr_a <- as.data.frame(lapply(agr[1:10], normalize))
agr_a$result <- agr$result
neu_a <- as.data.frame(lapply(neu[1:10], normalize))
neu_a$result <- neu$result
```

Train the different Linear Regression Models for each factor.

```
lr_model_opn <- lm(result ~ ., data = opn_a)
lr_model_csn <- lm(result ~ ., data = csn_a)
lr_model_ext <- lm(result ~ ., data = ext_a)
lr_model_agr <- lm(result ~ ., data = agr_a)
lr_model_neu <- lm(result ~ ., data = neu_a)
```

Finally, storing all the models in the form of “.rds” file for further use.

```
saveRDS(lr_model_opn, "models/lr_model_opn.rds")  
saveRDS(lr_model_csn, "models/lr_model_csn.rds")  
saveRDS(lr_model_ext, "models/lr_model_ext.rds")  
saveRDS(lr_model_agr, "models/lr_model_agr.rds")  
saveRDS(lr_model_neu, "models/lr_model_neu.rds")
```

Later these models are used in the application to predict the factor values for random users.

7 REFERENCES

- [1]<https://www.britannica.com/topic/personality>
- [2]https://openpsychometrics.org/_rawdata/
- [3]https://www.researchgate.net/publication/334239854_The_Importance_Role_of_Personality_Trait
- [4]https://sites.temple.edu/rtassessment/files/2018/10/Table_BFPT.pdf
- [5]<https://openpsychometrics.org/tests/IPIP-BFFM/>
- [6]<https://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures/Personality-BigFiveInventory.pdf>
- [7]<https://www.truity.com/test/big-five-personality-test>
- [8]https://en.wikipedia.org/wiki/Big_Five_personality_traits
- [9]<https://shiny.rstudio.com/>
- [10]<https://shiny.rstudio.com/tutorial/>
- [11]<https://shiny.rstudio.com/tutorial/written-tutorial/lesson1/>
- [12]<https://cran.r-project.org/>
- [13]<https://www.simplypsychology.org/big-five-personality.html>