



Faculty of Engineering and Technology  
Department of Electronics & Communication Engineering  
Jain Global Campus, Kanakapura Taluk  
Ramanagara District, Karnataka, India -562112

2017-2021

A Project Report on

**“Machine Learning Based Personality Classification Using K-Means Clustering Algorithm”**

Submitted in partial fulfilment for the award of the degree of

**BACHELOR OF TECHNOLOGY  
IN  
ELECTRONICS AND COMMUNICATION ENGINEERING**

Submitted by

**Sumukha A  
17BTREC108  
Sai Ritwik Reddy K  
17BTREC006**

Under the guidance of

**Mr. Sunil M P  
Assistant Professor**  
Department of Electronics & Communication Engineering  
Faculty of Engineering and Technology  
**JAIN** (Deemed-to-be University)



**Faculty of Engineering and Technology**  
**Department of Electronics and Communication Engineering**

Jain Global Campus, Kanakapura Taluk - 562112  
Ramanagara District, Karnataka, India

**2017-2021**

**A Project Report on**

**“MACHINE LEARNING BASED PERSONALITY  
CLASSIFICATION USING K-MEANS CLUSTERING  
ALGORITHM”**

**Submitted in partial fulfilment for the award of the degree of**

**BACHELOR OF TECHNOLOGY**  
**IN**  
**ELECTRONICS AND COMMUNICATION ENGINEERING**

**Submitted by**

**Sumukha A**  
**17BTREC108**

**Sai Ritwik Reddy K**  
**17BTREC006**

**Under the guidance of**

**Dr. Manjula TR**  
Assistant Professor  
Department of Electronics and Communication Engineering  
**Faculty of Engineering & Technology**  
**JAIN DEEMED-TO-BE UNIVERSITY**

**Faculty of Engineering & Technology**  
**Department of Electronics & Communication Engineering**

Jain Global campus  
Kanakapura Taluk - 562112  
Ramanagara District  
Karnataka, India

**CERTIFICATE**

This is to certify that the project work titled “**MACHINE LEARNING BASED PERSONALITY CLASSIFICATION USING K-MEANS CLUSTERING ALGORITHM**” is carried out by **Sumukha A (17BTREC108)**, **Sai Ritwik Reddy K (17BTREC006)** are bonafide students of Bachelor of Technology at the Faculty of Engineering & Technology, JAIN DEEMED-TO-BE UNIVERSITY, Bengaluru in partial fulfillment for the award of degree in Bachelor of Technology in Electronics and Communication Engineering, during the year **2020-2021**.

**Dr. Manjula TR**

Assistant Professor  
Dept. of ECE,  
Faculty of Engineering & Technology,  
JAIN DEEMED-TO-BE UNIVERSITY  
Date:

**Dr. R. Sukumar**

Head of the Department,  
Electronics and Communication,  
Faculty of Engineering & Technology,  
JAIN DEEMED-TO-BE UNIVERSITY  
Date:

**Dr. Hariprasad S.A**

Director,  
Faculty of Engineering & Technology,  
JAIN DEEMED-TO-BE UNIVERSITY  
Date:

Name of the Examiner

Signature of Examiner

1.

2

# DECLARATION

We, **Sumukha A (17BTREC108)**, **Sai Ritwik Reddy K (17BTREC006)**, are students of eighth semester B.Tech in **Electronics and Communication Engineering**, at Faculty of Engineering & Technology, **JAIN DEEMED-TO-BE UNIVERSITY**, hereby declare that the project titled “**Machine Learning Based Personality Classification Using Clustering Algorithm**” has been carried out by us and submitted in partial fulfilment for the award of degree in **Bachelor of Technology in Electronics and Communication Engineering** during the academic year **2020-2021**. Further, the matter presented in the project has not been submitted previously by anybody for the award of any degree or any diploma to any other University, to the best of our knowledge and faith.

Signature

Sumukha A  
17BTREC108



Sai Ritwik Reddy K  
17BTREC006



Place: Bengaluru  
Date : 23/05/2021

## ACKNOWLEDGEMENT

*It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project work.*

*First, we take this opportunity to express our sincere gratitude to **Faculty of Engineering & Technology, JAIN DEEMED-TO-BE UNIVERSITY** for providing us with a great opportunity to pursue our Bachelor's Degree in this institution.*

*In particular, we would like to thank **Dr. Hariprasad S.A, Director, Faculty of Engineering & Technology, JAIN DEEMED-TO-BE UNIVERSITY** for his constant encouragement and expert advice.*

*It is a matter of immense pleasure to express our sincere thanks to **Dr. R. Sukumar, Head of the department, Electronics and communication Engineering, JAIN DEEMED-TO-BE UNIVERSITY,** for providing right academic guidance that made our task possible.*

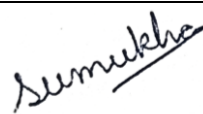
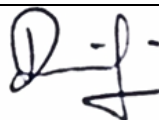
*We would like to thank our guide **Dr. Manjula TR Assistant Professor, Dept. of Electronics and Communication Engineering, JAIN DEEMED-TO-BE UNIVERSITY,** for sparing her valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.*

*We would like to thank our Project Coordinator **Mr. Sunil M P** and all the staff members of Electronics and Communication for their support.*

*We are also grateful to our family and friends who provided us with every requirement throughout the course.*

*We would like to thank one and all who directly or indirectly helped us in completing the Project work successfully.*

*Signature of Students*

Sumukha A	
Sai Ritwik Reddy K	

# ABSTRACT

The personality of an individual is a characteristic of person's attitude. A person's way of thinking and viewing one's surrounding environment in a certain way depends on the person's personality. It is important to understand one's personality because it would help one understand why they behave in certain situation in a certain way. Identifying a person's personality would be useful in assisting, choosing a right career or profession. When a company is estimating the success of a product, one of the factors to be considered is the personality of consumers. Further the personality has greater impact on social, personal and economic dimensions of life. Assessing the personality of group of people in an organization would enable to understand the work feel factor and enhance the productivity. This project presents machine learning method of classifying the personality pattern based on Big 5 model. The data used in this project is raw practical data that is directly derived from the system. This data is pre-processed on multiple stages so that it is suitable for training the model. The data is analysed and the optimal number of clusters possible for this data is found out, this optimal number of cluster is given to the unsupervised K-Means algorithm to generate clusters in order to recognize the patterns in the dataset and group individuals according to their personality pattern. This would enable recognizing the dominant personality of a country or assess personality patterns of an organization. This project can be further improved with more data and using dimensionality reduction techniques. It can be used by the recommendation system to recommend products without the consumer's order history. It can also be used by teachers to find the optimal methods of teaching by analysing the pupil's personalities.

# TABLE OF CONTENTS

List of Figures	v
Nomenclature used	v
<b>Chapter 1</b>	<b>01</b>
<b>1. INTRODUCTION</b>	<b>01</b>
1.1 Literature Survey	02
1.2 Limitations of the Current Work	02
1.3 Problem Definition	03
1.4 Objectives	03
1.5 Methodology	03
1.6 Software tools used	06
 <b>Chapter 2</b>	 <b>07</b>
<b>2. BASIC THEORY</b>	<b>07</b>
 <b>Chapter 3</b>	 <b>09</b>
<b>3. TOOL DESCRIPTION</b>	<b>09</b>
 <b>Chapter 4</b>	 <b>13</b>
<b>4. IMPLEMENTATION</b>	<b>13</b>
4.1 Software algorithm	13
4.1.1. Finding optimum number of clusters (Elbow Method)	13
4.1.2. K means Clustering	14
4.1.3. Cluster Visualizations	14
4.1.4. Data points visualization using PCA	16
4.1.5. Classifying personality of a new data point	17
 <b>Chapter 5</b>	 <b>19</b>
<b>5. RESULTS AND DISCUSSION</b>	<b>19</b>
 <b>CONCLUSIONS AND FUTURE SCOPE</b>	 <b>20</b>

<b>REFERENCES</b>	<b>viii</b>
<b>APPENDICES</b>	
<b>APPENDIX – I</b>	<b>ix</b>
<b>APPENDIX – II</b>	<b>xi</b>
<b>DETAILS OF PAPER PUBLICATION(ALONG WITH PAPER)</b>	<b>xiv</b>
<b>INFORMATION REGARDING STUDENTS</b>	<b>xxiii</b>
<b>BATCH PHOTOGRAPH ALONG WITH GUIDE</b>	<b>xxiv</b>



## LIST OF FIGURES

Fig. No.	Description of the figure	Page No.
1	Flow chart of personality classification using K Means clustering algorithm	11
2	Sample of data set layout	11
3	Working of clustering algorithm	15
4	Elbow method of optimal k value	18
5	Average personality pattern of each clusters	20
6	Cluster visualization using PCA	21
7	Obtained inertia value	23
8a	Average scores of each personality traits	23
8b	New participant's cluster given by K Means model	24

## NOMENCLATURE USED

EEG	electroencephalography
SVM	support vector machine
MBIT	Myers-Briggs Type Indicator
FRM	Fuzzy Relational Mapping
SSE	sum of squared distance
PCA	Principle Component Analysis

# Chapter 1

## 1. INTRODUCTION

A person's way of thinking and behavior in certain situations will contribute to one's personality. A person can be understood better by having a sense of awareness about one's personality. A lot of researches have been conducted in this area and researchers have come up with five core personality traits, they are called the Big Five personality traits. Each of five personality traits in the Big Five model represents two polar ends of a personality. Theoretically a person's personality can be anywhere in this spectrum of two polar extremes, but practically majority of the time a person's personality tend to remain somewhere in between the two extremes.

The Big Five personality traits are as follows:

1. Openness – Intellectual curiosity, Creative imagination
2. Conscientiousness – Organization, Responsibility.
3. Extroversion – Sociability, Outgoing.
4. Agreeableness – Trusting, Sympathetic
5. Neuroticism – Anxiety, Depression.

The personality reflected by high and low scorers in each personality trait is given below.

1. Openness: Individuals who score high in this personality trait are often more likely to be audacious and imaginative. People who score low in this trait are often people who like to follow the set path and would not like to break any set pattern, these people may struggle with abstract thinking.
2. Conscientiousness: Individuals who score high in this trait are often people who pays attentions to details and finishes important tasks right away. These people are very organized and procrastinates less. The low scorers are procrastinators and dislikes structure.
3. Extraversion: Individuals who score high in this trait tend to be more outgoing, active and are energized around social situations. These people say things before thinking, and often have a wide social circle. People who score low are the ones who prefer alone time to recharge, are very careful before speaking and have difficulty starting a conversation.
4. Agreeableness: Individuals who score high in this trait tend to have great deal of interest in other people, empathetic and assist other who are in need. People who score low are often self-centric, are not concerned about others and even try to manipulate others to get what they want.
5. Neuroticism: People who score high in this trait tend to experience anxiety, sadness, irritability and moodiness. People who score low tend to be calm, emotional. stable and balanced.

## **1.1. Literature Survey**

- This paper introduces a new method of recommendation system. It uses social media to predict personality of a person and use the personality to recommend the products in the absence of order history. [\[1\]](#)
- This paper uses the six-factor HEXACO personality model to rate the personality of the employees. It proposes a system where it aims to predict the personality of a candidate and hence the job fit by analyzing candidate's responses to open ended interview questions. [\[2\]](#)
- This paper uses electroencephalography (EEG) to classify personality dimensions. The classification was performed using support vector machine (SVM) to determine personality dimensions. [\[3\]](#)
- This paper focuses on the association between personality traits and facial attributes, it uses the dataset which consists of facial images and corresponding personality traits. The classifiers are trained on this labeled dataset and output is based on one-factor or two-factor personality classification [\[4\]](#).
- This paper proposes a study of personality based on a person's food preferences. It uses a dataset from a survey of 100 people based on a questionnaire about their food related behavior along with standard personality traits. [\[5\]](#)
- This paper focuses on individuals of a nation using dataset based on MBIT (Myers-Briggs Type Indicator) and determine whether personality affects various national indexes [\[6\]](#)
- This paper attempted to study the entire picture of personality's influence on work-life balance. The study used Big Five factors, and the data was analyzed FRM model (Fuzzy Relational Mapping)[\[7\]](#).

## **1.2. Limitations in current work**

The field of personality classification has been a hot topic for researchers since 1900s, various attempts have been made for classifying the model. In 1936 independent researches by Gordon Allport and Henry Odbert led to the foundation for other psychologists to start finding basic personality traits. The current personality tests take the responses from the participants and after the calculation of scores for each of the personality trait the result is displayed. The classification models are using this data to classify individuals based on their responses to the questionnaire alone, which makes it difficult to identify the dominant personality on

a geographical location basis. We can also check top participants who possess dominant personality trait in a category.

### 1.3. Problem definition

For a company to get the estimate on success of a product, one of the factors to be considered is that will the personality of people of a country match with the product they are about the launch. For recruiter, in order to hire the best talent, one of the factors to be considered is the candidate's personality. Considering the best out of all the applicants will be the major concern for the recruiter.

### 1.4. Objectives

This project aims to classify the participants based on their geographical location. It also focuses on classifying the new participants to one of the personality pattern groups. It shows the current number of participants from each country in the used dataset. It shows the top participants with highest scores in a particular personality trait.

### 1.5. Methodology

The personality classification using K means clustering algorithm is illustrated in the flow chart as shown in the Figure.1

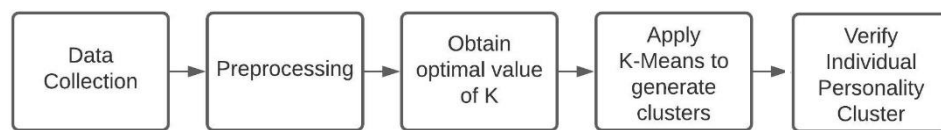


Fig.1. flow chart of personality classification using K Means clustering algorithm

#### Data set

The dataset was collected from Open Psychometrics during the year 2016-2018, at the start of the test, participants were informed that their responses would be registered and used for research, and at the end, they were asked to affirm their

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	dateload	screenw	screenh	introelapse	testelapse	endelapse	IPC	country
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0	6	1	GB
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0	11	1	MY
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0	7	1	GB
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0	7	1	GB
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	2016-03-03 02:02:57	1366.0	768.0	8.0	315.0	17	2	KE

Fig.2. Sample of data set layout

consent.

There are 10 questions for each personality trait, therefore there are 50 questions. Some of the questions are reverse marked, i.e. in extroversion one of the question is “I don't talk a lot”, here as we can observe, the question is addressing introversion, which is opposite of extroversion. As a result, when calculating the final score for this personality trait we must consider the reversed marks of the actual marks. To make the calculation simpler we have subtracted six from each of the reverse marked questions so that when we take the average of a personality trait, we get the magnitude a person's personality in that trait. There were reversed questions in each of the personality traits, all of those were identified and modified as mentioned above to make the further steps easier. The dataset also contains a column ('IPC') which indicates the number of participants from one IP address, for accuracy purpose we have selected all the participants whose 'IPC' was no more than 2. The dataset also provides participants approximate location based on IP address and also their country. We have used these data to derive insights based on geographical location. Few of the columns from the dataset has been shown in Fig (2)

### **Big five personality model**

The model used in the dataset for determining the personality is Big Five, many independent researchers contributed to the creation of the model. It has been studied in a variety of communities and cultures, it is the most widely accepted theory of personality today. The details of each personality trait has been given below.

(High and low scores exhibit the following, according to researchers [\[12\]](#))

1. Openness: It refers to a person's willingness to try new things, includes the ability to “think outside the box.”
  - a. High Score: Curious, Creative, Imaginative and Unconventional
  - b. Low Score: Dislikes Change, Predictable, Follower, Prefers routine and Traditional
2. Conscientiousness: It describes a person's ability to manage their impulses in order to participate in goal-oriented activities.
  - a. High Score: Directed, purpose, and self-disciplined and Dutifulness.
  - b. Low Score: Disorganized, Inept and irresponsible, Procrastinates and Indiscipline.
3. Extraversion: It reflects the willingness of a person to interact with their social environment.
  - a. High Score: Sociable, thrives on excitement and loves being the focus of attention. and outgoing.
  - b. Low Score: Prefers solitude, tired by an excessive amount of social contact, Reflective and dislikes being center of attention.
4. Agreeableness: It refers to people's attitudes toward other people's relationships.
  - a. High Score: Forgiveness, forthrightness, a desire to assist, and sympathy and empathetic.

- b. Low Score: Demanding, Stubborn, show-off, insults others and does not care about other's feelings.
- 5. Neuroticism: It describes the emotional stability of a person through how they view the world.
  - a. High Score: Anxious, Stress, Vulnerability, Mood swings, irritable and shy
  - b. Low Score: Calm, Confident, Resilient, Emotionally Stable and Balanced.

To find the factors that influence Big Five, Researchers have conducted a twin studies with 123 pairs of identical twins and 127 pairs of fraternal twins, and found that the amount of variance that can be attributed to genes is 40-60%. [\[13\]](#)

### **Algorithm for personality classification:**

#### Cleaning of data:

The real world datasets are full of noise and need to be preprocessed before feeding it to the algorithm. This dataset had total participants of 10,15,341. After removing missing values, redundant participants (more than two participants from same IP Address), the total number participants were 800543.

#### Reverse Marking:

The answers to each questions can be given in numbers in the range one to five, where one represents disagree and five represents agree. The reversed marked questions were subtracted from six to represent it in a proper manner.

#### Preprocessing:

In order to decrease the variance in the dataset, the features were normalized using “Min-Max Scaler”. By doing so, all features will be transformed into the range (0-1). The mathematical formulation is given below. Here,  $x$  represents a single feature.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

All the rescaling is done feature-wise, so it is independent.

#### Optimal number of clusters

The K-Means algorithm requires K value. For best results, we need to specify optimal value for K. This optimal value depends on dataset. In order to find the optimal value for K, we use the “Elbow Method”. It checks for range of values for K and returns the optimal value.

#### K-Means Algorithm

When K value is given, fits the data after a number of iterations, after each iteration we get more converged result. After the completion of the algorithm, we get the clusters. Data points belonging to same cluster have similar properties while, data points belonging to different clusters have different properties.

Inertia value: The quality of clusters formed by K-Means can be accessed using inertia value. This value indicates how far the points within a cluster are. The value

can be obtained from the algorithm after the completion. Smaller the inertia value, better is the cluster quality.

#### Verify individual personality cluster

Using the data points in the dataset, the clusters are obtained. We can also classify new data points to the existing clusters, and see where the new data point belongs.

## **1.6. Software tools used**

Software tools used

- Anaconda
- Jupyter Notebook
- Python

## Chapter 2

### 2. Basic Theory

#### 2.1. Clustering algorithm

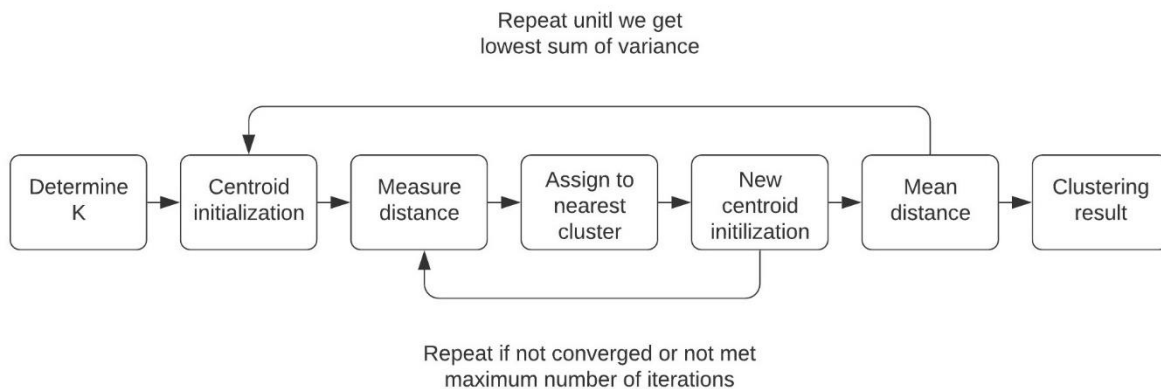


Fig.3. Working of clustering algorithm

The clustering algorithm focuses on grouping the data such that data belonging to same group are more similar than the data belonging to different groups. The algorithm chosen in this project is K-Means. As it is fast (less memory intensive) and simple.

The working of K-Means algorithm is as follows,

1. K (number of clusters/ groups) is initialized
2. Centroids are initialized randomly
3. Euclidian Distance between centroid and each data point is measured using the following expression.

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

4. Clusters are formed such that distance between data point and a centroid is minimum
5. Centroids are updated to the center of the cluster.
6. Mean distance is calculated.
7. Step 2 to step 6 is repeated until we get lowest sum of variance
8. Step 3 to step 5 are repeated if it is not converged or maximum number of iterations are met.



## 2.2. Elbow Method

The K-Means algorithm requires a K value (i.e. number of clusters). K value should be optimal; this depends on the dataset. To find the optimal value for K, we use elbow method.

The working of Elbow method is as follows,

1. A range of values is given, starting from two.
2. The model is fit for every value and sum of squared distance (SSE) is calculated for each of the values.
3. The optimal value is the point where the SSE starts to flatten out, forming an elbow.

## 2.3. Principle Component Analysis (PCA)

The data is of high dimension, in order to visualize this data, we reduce the dimension of the data and preserving as much information as possible at the same time. PCA is used for this purpose

The working of PCA is as follows,

1. Dataset is standardized, i.e. all the values in the dataset lie in the range of -1 to +1.
2. Covariance matrix is calculated using the below expression,

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

3. Eigenvalues and eigenvectors are calculated using the below expression,

$$Av - \lambda v = 0 ; (A - \lambda I)v = 0$$

where A is covariance matrix, v is vector and  $\lambda$  is scalar.

4. Eigenvalues and eigenvectors are sorted
5. K eigenvalues are picked and a matrix of eigenvectors are formed
6. The original matrix is transformed.

## Chapter 3

### 3. Tool Description

#### 1. Jupyter Notebook:

It is an open source application that allows writing code, equations, visualizations and text (in markdown format) in one place, making the readability, documentation and execution of code very simple and easy. It is one of the best applications for data scientists. The standard jupyter notebook document extension is .ipynb. It is a book record that is put away in the JSON design that contains the substance of the notebook. There might be numerous cells in a notebook and the substance of each can be python code, text or a video connection that has been changed over into strings of text and is accessible alongside the metadata of the notebook.

The kernel is a software that runs and translates the user's code. There are exclusive kernels for exclusive languages that Jupyter Notebook makes use of however for Python it extends the ipython kernel.

A notebook kernel is a “computational engine” that executes the code contained in a Notebook document. The ipython kernel, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels).

The kernel executes the code within the notebook and returns the output (if any) to the frontend interface. The state of a kernel relates to the complete record and now no longer simply individual cells. Anything carried out in a single cell could be available to be used within the subsequent cell as well.

There are 4 types of cells in a jupyter notebook:

- a. Code — This is the cell where we write our python code that will be computed by the ipython kernel and the output is displayed under the cell. Here is an example of a code cell
- b. Markdown — This is where you add the documentation by putting text formatted using Markdown. The output is displayed in place of the cell when it is run.
- c. Raw NBConvert — This is another tool to convert your jupyter notebook into another file format like PDF, HTML, etc.
- d. Heading — This is the same as writing a heading (Line starting with #) in Markdown. To add a headline to your notebook you can use this.

The latest product, JupyterLab incorporates Jupyter Notebook into an Integrated Development type Editor that can run in the browser. It can be thought of JupyterLab as an advanced version of Jupyter Notebook. JupyterLab allows the user to run terminals, text editors and code consoles in your browser in addition to Notebooks.

Various functions in the Kernel menu:

1. Interrupt: Used to stop the execution or running of a particular cell. This command is useful when you have reached a desired result within a specific number of epochs or in case you made an error, and you realize this while running the code.

2. Restart: Useful to restart the Kernel of the Notebook.
3. Restart and Clear Output: Used to restart the Kernel of the Notebook and reset all the cells that were run previously.
4. Restart and Run all: Used to restart the Kernel of the Notebook, and reset all the cells that were run previously, and finally re-run through all the cells of the Notebook.
5. Reconnect: Used to reconnect to a dead kernel which might occur at times due to a lack of memory.
6. Shutdown: Used to shutdown the current working Kernel of the Notebook.
7. Change kernel: Allows you to switch Kernels.

Programming with python has been given a brand new picture by these notebooks. Features provided by this notebook can be utilized and data science journey can be more enjoyable with those cells containing the step-by-step evaluation along with the documentation and visible insights.

## **2. Anaconda:**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

A python distribution for data science that aims to ease the package management by checking dependencies and environments before making a package installation. It comes with Anaconda Navigator which is the graphical user interface for managing other applications that comes with the package.

Together with a list of Python packages, tools like editors, Python distributions encompass the Python interpreter. Anaconda is certainly considered one among numerous Python distributions. Anaconda is a brand new distribution of the Python and R data science package. It turned into previously called Continuum Analytics.

Anaconda has more than one hundred new packages. The package manager is also an environment manager, a Python distribution, and a collection of open source packages and contains more than 1000 R and Python Data Science Packages.

Anaconda helps in simplified package management and deployment. Anaconda comes with a wide variety of tools to easily collect data from various sources using various machine learning and AI algorithms. It helps in getting an easily manageable environment setup which can deploy any project with the click of a single button.

Packages available in Anaconda

- Over 250 packages are automatically installed with Anaconda.
- Over 7,500 additional open-source packages (including R) can be individually installed from the Anaconda repository with the conda install command.
- Thousands of other packages are available from Anaconda.org.
- You can download other packages using the pip install command that is installed with Anaconda. Pip packages provide many of the features of conda packages and in some cases they can work together. However, the preference should be to install the conda package if it is available.
- You can also make your own custom packages using the conda build command, and you can share them with others by uploading them to Anaconda.org, PyPI, or other repositories.

### **3. Python:**

The Python programming language is older than many of its popular counterparts including R, Java, and even JavaScript. The concept behind it was first implemented in the 1980s. The then developer, Guido van Rossum came up with Python as a hobby project during Christmas. His reason for creating it was to help him work on Amoeba operating system in handling and user-interfacing.

A high level programming language for data science and machine learning projects mainly due to the extensive libraries, community support and ease of use. Python language can be used on any modern computer operating system. It can be used for processing text, numbers, images, scientific data and just about anything else you might save on a computer. It is used daily in the operations of the Google search engine, the video-sharing website YouTube, NASA and the New York Stock Exchange.

Python is an interpreted language. This means that it is not converted to computer-readable code before the program is run but at runtime. In the past, this type of language was called a scripting language, intimating its use was for trivial tasks. However, programming languages such as Python have forced a change in that nomenclature. Increasingly, large applications are written almost exclusively in Python.

It is a suitable language that bridges the gaps between business and developers. Subsequently, it takes less time to bring a Python program to market compared to other languages such as C#/Java. Additionally, there are a large number of python machine learning and analytical packages. A large number of communities and books are available to support Python developers. Nearly all types of applications, ranging from forecasting analytical to UI, can be implemented in Python.

Python works as follows:

- A Python virtual machine is created where the packages (libraries) are installed. Think of a virtual machine as a container.
- The python code is then written in .py files
- CPython compiles the Python code to bytecode. This bytecode is for the Python virtual machine.
- When you want to execute the bytecode then the code will be interpreted at runtime. The code will then be translated from the bytecode into the machine code. The bytecode is not dependent on the machine on which you are running the code. This makes Python machine-independent

Modules in python:

- Python is shipped with over 200 standard modules.
- A module is a component that groups similar functionality of your python solution.
- Any python code file can be packaged as a module and then it can be imported.
- Modules encourage componentized design in your solution.
- They provide the concept of namespaces to help you share data and services.
- Modules encourage code reusability and reduce variable name clashes.

Packages in python:

- Package is a directory of modules.
- If your Python solution offers a large set of functionalities that are grouped into module files, then you can create a package out of your modules to better distribute and manage your modules.
- Packages enable us to organize our modules better which helps us in resolving issues and finding modules easier.
- Third-party packages can be imported into your code such as pandas/sci-kit learn and tensor flow to name a few.
- A package can contain a large number of modules.

## Chapter 4

### 4. Implementation

#### 4.1. Software Algorithm

##### 4.1.1. Finding optimum number of clusters (Elbow Method):

```
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer

kmeans = KMeans()
visualizer = KElbowVisualizer(kmeans, k=(2,15))
visualizer.fit(df_sample)
visualizer.poof()
```

To implement the elbow method, “yellowbrick” package has been used, the range of values provided for “k” was from 2 to 15, a sample from the dataset was taken and the result was visualized.

In the fig (3), the range for ‘k’ was set from 2 to 14, and the elbow was observed at k=5(Blue line).

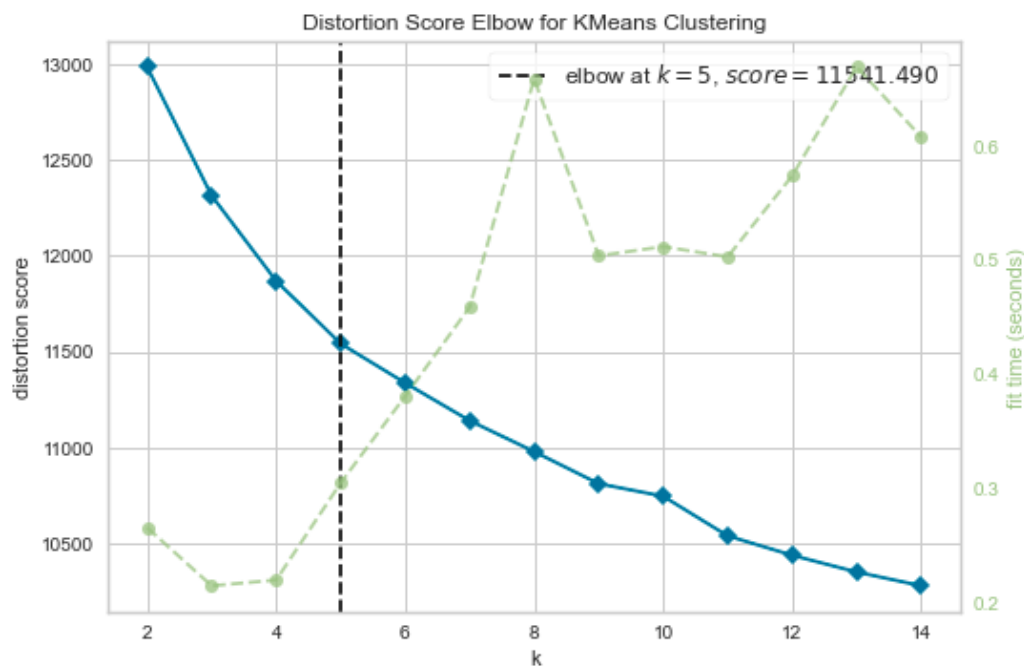


Fig.4. Elbow method of optimal k value

### 4.1.2. K means Clustering:

The clustering algorithm chosen is K-Means. As it is fast (less memory intensive), simple. As we were looking for reducing the data to centroids, this algorithm was chosen.

```
# Creating K-means Cluster Model
from sklearn.cluster import KMeans

# I define 5 clusters and fit my model
kmeans = KMeans(n_clusters=5, verbose=0)
k_fit = kmeans.fit(df)
```

The model was trained with the data. The optimal number of clusters are “5” as obtained in the previous step.

```
pd.options.display.max_columns = 10
predictions = k_fit.labels_
df['Clusters'] = predictions
df.head()
```

A new column was created, named “Clusters” and the corresponding classification for each participant is written to it.

```
data_sums = pd.DataFrame()
data_sums['extroversion'] = df[ext].sum(axis=1)/10
data_sums['neurotic'] = df[est].sum(axis=1)/10
data_sums['agreeable'] = df[agr].sum(axis=1)/10
data_sums['conscientious'] = df[csn].sum(axis=1)/10
data_sums['open'] = df[opn].sum(axis=1)/10
data_sums['clusters'] = predictions
data_sums.groupby('clusters').mean()
```

For better understandability of each personality traits the average of each personality trait’s questions was taken and saved in a new column with column name same as that of personality trait respectively.

### 4.1.3. Cluster Visualizations

```
dataclusters = data_sums.groupby('clusters').mean()
plt.figure(figsize=(10,3))
for i in range(0, 5):
    plt.subplot(1,5,i+1)
    plt.bar(dataclusters.columns, dataclusters.iloc[:, i], color='blue', alpha=0.2)
    plt.plot(dataclusters.columns, dataclusters.iloc[:, i], color='red')
    plt.title('Cluster ' + str(i))
    plt.xticks(rotation=45)
    plt.ylim(0,1)
```

The average personality trait pattern for each cluster is visualized. Bar graph and line graph have been overlapped so that the pattern is clearly visible. The visualizations have been given below.

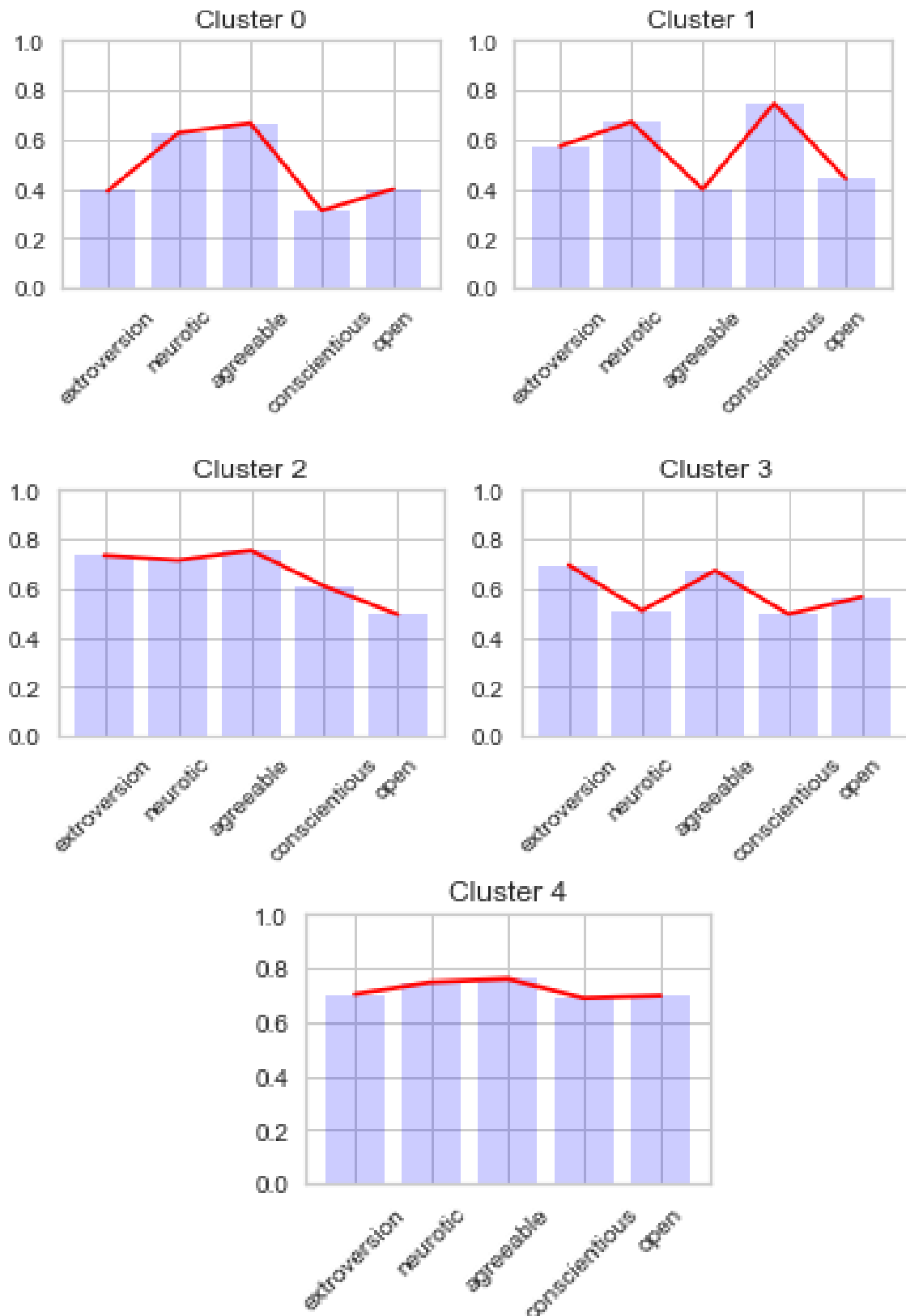


Fig.5. Average personality pattern of each clusters



For visualizing each clusters, the mean of all data points belonging to a cluster was considered. Using 'matplotlib' library the visualization of each cluster was achieved. Each of the clusters represented the pattern of personality trait of all the participants belonging to that cluster.

#### **4.1.4. Data points visualization using PCA**

PCA was used from sklearn and seaborn packages of python, the data points were visualized, for visualizing the data points as clusters, PCA (Principal Component Analysis) was used. PCA is used to create visualization of data that maximizes the variance of the projection coordinates while minimizing residual variance in the least squares sense. (for more information refer [docs](#))

```
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
pca_fit = pca.fit_transform(df)

df_pca = pd.DataFrame(data=pca_fit, columns=['PCA1', 'PCA2'])
df_pca['Clusters'] = predictions
df_pca.head()
```

The above code snippet shows PCA algorithm storing the “dimensionally reduced” result to the “Clusters” column. This column is used to visualize the data points as shown below.

##### Personalities based on geographical regions

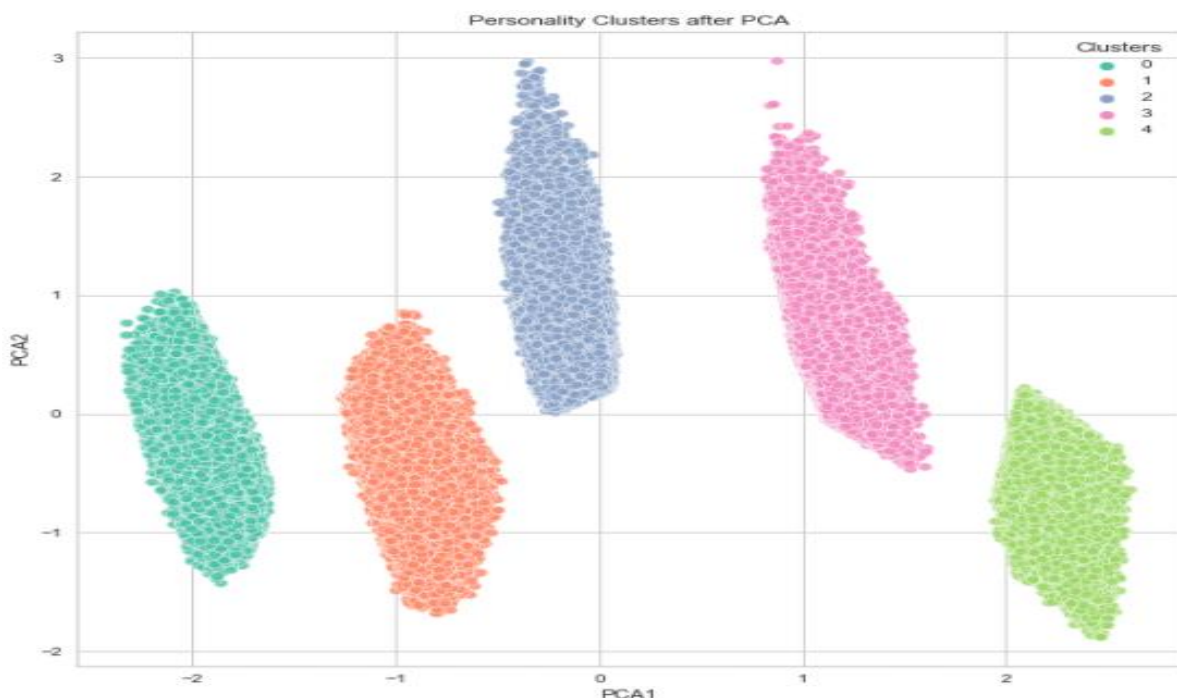


Fig.6. Cluster visualization using PCA

```
countries = df1['country'].unique()

for i in countries:
    c = df1[df1['country'].str.contains(i)]
    result = list(c.Clusters.mode())
    print(i, "    Clusters:", result)
```

The above code snippet shows that, for every participant belonging to a country “i”, the highest number of repetition of a cluster (mode operation), is the dominant personality of that country. A fraction of result of the above algorithm is given below.

ZA	Clusters: [0]
HK	Clusters: [1]
BR	Clusters: [2]
CH	Clusters: [4]
TH	Clusters: [3]
IT	Clusters: [2]
ES	Clusters: [1]
AE	Clusters: [1]
HR	Clusters: [2]
GR	Clusters: [2]
IE	Clusters: [1]
DE	Clusters: [3]
PT	Clusters: [2]
NL	Clusters: [4]

#### **4.1.5. Classifying personality of a new data point**

```
my_data = pd.read_excel(r"H:\MY projects\project\Project\my_personality_test.xls")
my_data
```

The above code snippet shows the loading of a file “my\_personality\_test.xls” containing answers to all the questions by the participant.

```
MYcol_list = list(my_data)
ext = MYcol_list[0:10]
est = MYcol_list[10:20]
agr = MYcol_list[20:30]
csn = MYcol_list[30:40]
opn = MYcol_list[40:50]

my_sums = pd.DataFrame()
my_sums['extroversion'] = my_data[ext].sum(axis=1)/10
my_sums['neurotic'] = my_data[est].sum(axis=1)/10
my_sums['agreeable'] = my_data[agr].sum(axis=1)/10
my_sums['conscientious'] = my_data[csn].sum(axis=1)/10
my_sums['open'] = my_data[opn].sum(axis=1)/10

my_sums['cluster'] = my_personality
my_sums
```

Sum of my question groups

	extroversion	neurotic	agreeable	conscientious	open	cluster
0	3.3	2.6	3.2	3.1	3.2	4

The above code snippet, shows the processing of the responses by the participant and the corresponding classification by the K-Means algorithm.

## Chapter 5

### 5. Results and Discussion

#### Country based dominant personalities:

As we have discussed in the dataset section that we have country column in our dataset which indicates the participant's country, we can use that to find the dominant personality of that country. First, we take all the participants from a country and consider their corresponding cluster, then we apply mode operation to get the dominant personality of the country.

From clustering the participants into groups, we are able to understand that the personality traits of a cluster, is the mean of personality traits of all participants belonging to that cluster. With the help of this clusters we can find the dominant personality of a county by simply considering the mode of cluster all the participants from that country.

```
In [15]: # Creating K-means Cluster Model
from sklearn.cluster import KMeans

# I define 5 clusters and fit my model
kmeans = KMeans(n_clusters=5, verbose=0)
k_fit = kmeans.fit(df)
kmeans.inertia_

Out[15]: 1879676.5329615066
```

Fig.7. Obtained inertia value

Due to the high dimensionality of the dataset with 50 columns i.e. 10 columns for each of five personality traits, the inertia obtained was quite high also called as 'curse of dimensionality'. Although steps were taken to reduce dimensionality the model which in turn reduced the inertia value, it was observed that prediction was highly skewed. In order to avoid that, the dimensionality reduction was discarded.

#### Clustering of new participants:

If a new participant answers the questions, the new responses can also be put in appropriate cluster. By doing this, the accuracy of country based dominant personalities can be increased. We can take the average of each personality groups and find the individual's dominant personality trait. New participant's data can be used to classify to a cluster.

```
In [34]: my_data = pd.read_excel(r"H:\MY projects\project\Project\my_personality_test.xls")
my_data

Out[34]:
```

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6
0	2	5	3	3	4	4	3	2	4	3	2	2	3	4	3	4

```

In [35]: my_personality = k_fit.predict(my_data)
print("My Personality Cluster: ", my_personality)

My Personality Cluster: [4]
```

Fig.8A. New participant's cluster given by K Means model

Sum of my question groups

Out[44]:

	extroversion	neurotic	agreeable	conscientious	open	cluster
0	3.3	2.6	3.2	3.1	3.2	4

Fig .8B. Average scores of each personality traits

The average of each of personality trait's responses are computed and stored in respective trait's names. The dominant personality trait of the new participant is extroversion (as the average is highest compared to other traits).

### **Conclusion and future scope**

**Conclusion:** In this project, we understood how clustering can be used to derive patterns in the dataset and classify the data points to corresponding clusters. We also understood how we can use these clusters to determine the approximate dominant personality of a country, also we saw how we can add new participants to our dataset to get more generalized model.

**Future Scope:** It can be used by the recommendation system to recommend products without the consumer's order history. It can be used by hiring managers to determine the behaviour of potential candidate and analyse if the candidate is right fit or not. It can also be used by the companies to get an insight whether a product succeeds in a country by analysing the dominant personality. It can be used by teachers to find the optimal methods of teaching by analysing the pupil's personalities

## REFERENCES

- [1] Sahraoui Dhelim, Huansheng Ning, Nyothiri Aung, Runhe Huang and Jianhua Ma, “Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery”, IEEE Transactions on Computational Social Systems, Volume: 8, Issue: 1, Feb. 2021.
- [2] Madhura Jayaratne and Buddhi Jayatilleke, “Predicting Personality Using Answers to Open-Ended Interview Questions”, IEEE Access, Volume: 8, June 2020.
- [3] Fadhilah Qalibi Annisa, Eko Supriyanto, Sahar Taheri, “Personality Dimensions Classification with EEG Analysis using Support Vector Machine”, 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Dec 2020.
- [4] Mingliang Xue; Xiaodong Duan; Yuangang Wang; Yuting Liu, “A Computational Personality Traits Analysis Based on Facial Geometric Features”, IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nov 2019.
- [5] Tasfia Hoque, Raqeebir Rab, Khushnoor Rafsan Jani Alam, Saif Hasan Khan, M. A. Wadud Shuvro and Umme Zakia, “Empirical Study on Personality Trait Classification by Food Related Preferences”, International Conference on Electrical, Computer and Communication Engineering (ECCE), Feb 2019.
- [6] Emine Yaman, Azra Musić-Kiliç, Zaid Zerdo, “Using Classification to Determine Whether Personality Profiles of Countries Affect Various National Indexes”, International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), May 2018
- [7] A.Victor Devadoss, J. Befija Minnie, “A Study of Personality Influence in Building Work Life Balance Using Fuzzy Relation Mapping(FRM)”, International Journal of Data Mining Techniques and Applications Volume: 02 Issue: 02, June 2013.
- [8] John, O. P., & Srivastava, S, “The big-five trait taxonomy: History, measurement and theoretical perspectives”, Oakland, CA: University of California. 1999.
- [9] Kerry L. Jang, W. John Livesley and Philip A. Vemon, “Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study”, Department of Psychiatry, University of British Columbia, Vancouver, Canada, 1996
- [10] Rajalaxmi Hegde, Sandeep Kumar Hegde, Sanjana, Sapna Kotian and Shreya C Shetty,” Personality classification using Data mining approach”, international Journal of Research and Analytical Reviews (IJRAR) volume 6, issue 1, March 2019
- [11] Song Lai; Bo Sun; Fati Wu; Rong Xiao, “Automatic Personality Identification Using Students’ Online Learning Behavior”, IEEE Transactions on Learning Technologies Volume: 13, Issue: 1, March 2020.

## APPENDIX - I

### SOURCE CODE

```
# dataset
data = read("file-path")

# cleaning data
missing values = sum(data.isnull())
data = data – missing values
data = data.reset_index()

# pre processing
Import MinMaxScaler
new_data = scaler(data)
data_sample = new_data[0:5000]
visualizer = elbowMethod(kmeans, k=(2, 15))
visualizer.fit(data_sample)
print(visualizer.graph)

# K-means
Kmeans = K-means(cluster=5)
K_fit = Kmeans.fit(new_data)
new_data["Cluster"] = K_fit.labels

# Question groups
new_data["extroversion"] = new_data.sum[0:10]/10
new_data["neurotic"] = new_data.sum[10:20]/10
new_data["agreeable"] = new_data.sum[20:30]/10
new_data["conscientious"] = new_data.sum[30:40]/10
new_data["open"] = new_data.sum[40:50]/10

# visualizing each cluster groups
cluster_group = new_data.group("clusters").mean()
for each_cluster in cluster_group:
    plot(bar_graph(each_cluster))
    plot(line_graph(each_cluster))
```

```

# visualize cluster
Import PCA
pca = PCA.fit(new_data)
plot(scatter_plot(pca))

# country based personality classification
data2= new_data[“country”, “cluster”]
data2 = filter (“country” if len(country) > 1000)
for country in data2:
    max_cluster = mode(country[“cluster”])
    print (country “---“max_cluster)

# classifying personality of new participant
data3 = read(“new-data-file-path”)
data3[“extroversion”] = data3.sum [0:10]/10
data3 [“neurotic”] = data3.sum [10:20]/10
data3 [“agreeable”] = data3.sum [20:30]/10
data3 [“conscientious”] = data3.sum [30:40]/10
data3 [“open”] = data3.sum [40:50]/10
cluster = k_fit.predict(data3)
data3[“cluster”] = cluster

# get top 5 performers in “extroversion”
new_data.sort(by= “extroversion”, ascending=False)
print(new_data[0:5])

```



## APPENDIX-II

### DATASHEETS

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6	EST7
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	1.0	4.0	4.0	2.0	2.0	2.0	2.0
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	2.0	3.0	4.0	1.0	3.0	1.0	2.0
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	4.0	4.0	4.0	2.0	2.0	2.0	2.0
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	3.0	3.0	3.0	2.0	3.0	2.0	2.0
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	1.0	5.0	5.0	3.0	1.0	1.0	1.0
EST8	EST9	EST10	AGR1	AGR2	AGR3	AGR4	AGR5	AGR6	AGR7	AGR8	AGR9	AGR10	CSN1	CSN2	CSN3	CSN4	CSN5
2.0	3.0	2.0	2.0	5.0	2.0	4.0	2.0	3.0	2.0	4.0	3.0	4.0	3.0	4.0	3.0	2.0	2.0
1.0	3.0	1.0	1.0	4.0	1.0	5.0	1.0	5.0	3.0	4.0	5.0	3.0	3.0	2.0	5.0	3.0	3.0
2.0	1.0	3.0	1.0	4.0	1.0	4.0	2.0	4.0	1.0	4.0	4.0	3.0	4.0	2.0	2.0	2.0	3.0
2.0	4.0	3.0	2.0	4.0	3.0	4.0	2.0	4.0	2.0	4.0	3.0	4.0	2.0	4.0	4.0	4.0	1.0
1.0	3.0	2.0	1.0	5.0	1.0	5.0	1.0	3.0	1.0	5.0	5.0	3.0	5.0	1.0	5.0	1.0	3.0
CSN6	CSN7	CSN8	CSN9	CSN10	OPN1	OPN2	OPN3	OPN4	OPN5	OPN6	OPN7	OPN8	OPN9	OPN10	EXT1_E		
4.0	4.0	2.0	4.0	4.0	5.0	1.0	4.0	1.0	4.0	1.0	5.0	3.0	4.0	5.0	9419.0		
1.0	3.0	3.0	5.0	3.0	1.0	2.0	4.0	2.0	3.0	1.0	4.0	2.0	5.0	3.0	7235.0		
3.0	4.0	2.0	4.0	2.0	5.0	1.0	2.0	1.0	4.0	2.0	5.0	3.0	4.0	4.0	4657.0		
2.0	2.0	3.0	1.0	4.0	4.0	2.0	5.0	2.0	3.0	1.0	4.0	4.0	3.0	3.0	3996.0		
1.0	5.0	1.0	5.0	5.0	5.0	1.0	5.0	1.0	5.0	1.0	5.0	3.0	5.0	5.0	6004.0		

EXT1_E	EXT2_E	EXT3_E	EXT4_E	EXT5_E	EXT6_E	EXT7_E	EXT8_E	EXT9_E	EXT10_E	EST1_E	EST2_E	EST3_E	EST4_E
9419.0	5491.0	3959.0	4821.0	5611.0	2756.0	2388.0	2113.0	5900.0	4110.0	6135.0	4150.0	5739.0	6364.0
7235.0	3598.0	3315.0	2564.0	2976.0	3050.0	4787.0	3228.0	3465.0	3309.0	9036.0	2406.0	3484.0	3359.0
4657.0	3549.0	2543.0	3335.0	5847.0	2540.0	4922.0	3142.0	14621.0	2191.0	5128.0	3675.0	3442.0	4546.0
3996.0	2896.0	5096.0	4240.0	5168.0	5456.0	4360.0	4496.0	5240.0	4000.0	3736.0	4616.0	3015.0	2711.0
6004.0	3965.0	2721.0	3706.0	2968.0	2426.0	7339.0	3302.0	16819.0	3731.0	4740.0	2856.0	7461.0	2179.0
EST5_E	EST6_E	EST7_E	EST8_E	EST9_E	EST10_E	AGR1_E	AGR2_E	AGR3_E	AGR4_E	AGR5_E	AGR6_E	AGR7_E	
3663.0	5070.0	5709.0	4285.0	2587.0	3997.0	4750.0	5475.0	11641.0	3115.0	3207.0	3260.0	10235.0	
3061.0	2539.0	4226.0	2962.0	1799.0	1607.0	2158.0	2090.0	2143.0	2807.0	3422.0	5324.0	4494.0	
8275.0	2185.0	2164.0	1175.0	3813.0	1593.0	1089.0	2203.0	3386.0	1464.0	2562.0	1493.0	3067.0	
3960.0	4064.0	4208.0	2936.0	7336.0	3896.0	6062.0	11952.0	1040.0	2264.0	3664.0	3049.0	4912.0	
3324.0	2255.0	4308.0	4506.0	3127.0	3115.0	6771.0	2819.0	3682.0	2511.0	16204.0	1736.0	28983.0	
AGR8_E	AGR9_E	AGR10_E	CSN1_E	CSN2_E	CSN3_E	CSN4_E	CSN5_E	CSN6_E	CSN7_E	CSN8_E	CSN9_E	CSN10_E	
5897.0	1758.0	3081.0	6602.0	5457.0	1569.0	2129.0	3762.0	4420.0	9382.0	5286.0	4983.0	6339.0	
3627.0	1850.0	1747.0	5163.0	5240.0	7208.0	2783.0	4103.0	3431.0	3347.0	2399.0	3360.0	5595.0	
13719.0	3892.0	4100.0	4286.0	4775.0	2713.0	2813.0	4237.0	6308.0	2690.0	1516.0	2379.0	2983.0	
7545.0	4632.0	6896.0	2824.0	520.0	2368.0	3225.0	2848.0	6264.0	3760.0	10472.0	3192.0	7704.0	
1612.0	2437.0	4532.0	3843.0	7019.0	3102.0	3153.0	2869.0	6550.0	1811.0	3682.0	21500.0	20587.0	
OPN1_E	OPN2_E	OPN3_E	OPN4_E	OPN5_E	OPN6_E	OPN7_E	OPN8_E	OPN9_E	OPN10_E	dateload	screenw	screenh	
3146.0	4067.0	2959.0	3411.0	2170.0	4920.0	4436.0	3116.0	2992.0	4354.0	2016-03-03 02:01:01	768.0	1024.0	
2624.0	4985.0	1684.0	3026.0	4742.0	3336.0	2718.0	3374.0	3096.0	3019.0	2016-03-03 02:01:20	1360.0	768.0	
1930.0	1470.0	1644.0	1683.0	2229.0	8114.0	2043.0	6295.0	1585.0	2529.0	2016-03-03 02:01:56	1366.0	768.0	
3456.0	6665.0	1977.0	3728.0	4128.0	3776.0	2984.0	4192.0	3480.0	3257.0	2016-03-03 02:02:02	1920.0	1200.0	
8458.0	3510.0	17042.0	7029.0	2327.0	5835.0	6846.0	5320.0	11401.0	8642.0	2016-03-03 02:02:57	1366.0	768.0	

introelapse	testelapse	endelapse	IPC	country	lat_appx_lots_of_err	long_appx_lots_of_err
9.0	234.0	6	1	GB	51.5448	0.1991
12.0	179.0	11	1	MY	3.1698	101.706
3.0	186.0	7	1	GB	54.9119	-1.3833
186.0	219.0	7	1	GB	51.75	-1.25
8.0	315.0	17	2	KE	1.0	38.0

## Machine Learning Based Personality Classification Using Clustering Algorithm

Sumukha A Manjula T R Ritwik Reddy

**Abstract:** The personality of an individual is a characteristic of person's attitude. Identifying a person's personality would be useful in assisting, choosing a right career or profession. Further the personality has greater impact on social, personal and economic dimensions of life. Assessing the personality of group of people in an organization would enable to understand the work feel factor and enhance the productivity. This paper presents machine learning method of classifying the personality pattern based on Big 5 model. The unsupervised K-Means algorithm is used to generate clusters to recognize the patterns in the dataset and group individuals according to their personality pattern. This would enable recognize the dominant personality of a country or assess personality patterns of an organization.

**Keywords:**

### 1. Introduction:

A personality of a person describes who one really is, a person's feelings and behavior make up most of the personality. It's commonly thought to be something that emerges from within an individual and is fairly constant across their lives. Humans are social creatures, knowing personality of others helps us to understand better and maintain better relationship with them. Personality awareness aids in the development of a positive outlook on life and the reduction of stress. It works like a GPS, helping one navigate through life.

There are infinite number of characteristics that combine to make a personality, it becomes quite challenging to find a way to classify them. Fortunately, enough studies have been done on this topic [1] and researchers have come up with five broad categories of personality traits, where each personality trait represents two extremes of a personality type. When we talk about high score in 'extraversion', it represents extrovert personality of a person, however, when we talk about low score in 'extraversion', it represents introvert personality of a person. Studies show the majority of people in the real world fall somewhere between the two polar ends of each dimension. [2].

There are various trait perspectives to personality, Regardless of psychologists differing opinions, the big five model is accepted all over the world, due to extensive research being conducted till date. Various questionnaires have been developed by psychologists to classify people into personality types. In this paper we have used questionnaire based on Big-five model[3].

Research is conducted on classification of personalities by using N-closest neighborhood with data mining. Researcher has built an application that gives the individual's personality and suitable career options after answering the set of questions. Provision is provided to view previous records after the login page [4]. There is also research made on personality of students using online learning behavior, the paper uses ENN (extended nearest neighbor) classification method[5].

[6] introduces a new method of recommendation system. It uses social media to predict personality of a person and use the personality to recommend the products in the absence of order history. In [7], the six-factor HEXACO personality model to rate the personality of the employees. It proposes a system where it aims to predict the personality of a candidate and

hence the job fit by analyzing candidate's responses to open ended interview questions. The method of using electroencephalography (EEG) to classify personality dimensions. [8] The classification was performed using support vector machine (SVM) to determine personality dimensions. The association between personality traits and facial attributes is proposed in [9]. It uses the dataset which consists of facial images and corresponding personality traits. The classifiers are trained on this labeled dataset and output is based on one-factor or two-factor personality classification. A study of personality based on a person's food preferences is proposed in [10]. It uses a dataset focusing on a questionnaire on people's dietary behavior and typical personality characteristics, based on a survey of 100 people. [11] focuses on individuals of a nation using dataset based on MBIT (Myers-Briggs Type Indicator) and determine whether personality affects various national indexes

## 2. Methodology

The personality classification using K means clustering algorithm is illustrated in the flow chart as shown in the Figure.1

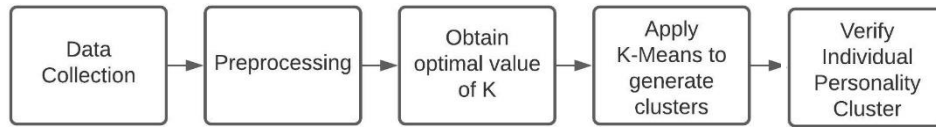


Fig.1. flow chart of personality classification using K Means clustering algorithm

### 2.1. Data set

The dataset was collected from Open Psychometrics during the year 2016-2018, at the start of the test, participants were informed that their responses would be registered and used for research, and at the end, they were asked to affirm their consent.

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	dateload	screenw	screenh	introelapse	testelapse	endelapse	IPC	country
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0	6	1	GB
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0	11	1	MY
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0	7	1	GB
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0	7	1	GB
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	2016-03-03 02:02:57	1366.0	768.0	8.0	315.0	17	2	KE

Fig.2. Sample of data set layout

There are 10 questions for each personality trait, therefore there are 50 questions. Some of the questions are reverse marked, i.e. in extroversion one of the question is "I don't talk a lot", here as we can observe, the question is addressing introversion, which is opposite of extroversion. As a result, when calculating the final score for this personality trait we must consider the reversed marks of the actual marks. To make the calculation simpler we have subtracted six from each of the reverse marked questions so that when we take the average of

a personality trait, we get the magnitude a person's personality in that trait. There were reversed questions in each of the personality traits, all of those were identified and modified as mentioned above to make the further steps easier. The dataset also contains a column ('IPC') which indicates the number of participants from one IP address, for accuracy purpose we have selected all the participants whose 'IPC' was no more than 2. The dataset also provides participants approximate location based on IP address and also their country. We have used these data to derive insights based on geographical location. Few of the columns from the dataset has been shown in Fig (2)

## 2.2. Big five personality model

The model used in the dataset for determining the personality is Big Five, many independent researchers contributed to the creation of the model. It has been studied in a variety of communities and cultures, it is the most widely accepted theory of personality today. The details of each personality trait has been given below.

(High and low scores exhibit the following, according to researchers [\[12\]](#))

6. Openness: It refers to a person's willingness to try new things, includes the ability to "think outside the box."
  - a. High Score: Curious, Creative, Imaginative and Unconventional
  - b. Low Score: Dislikes Change, Predictable, Follower, Prefers routine and Traditional
7. Conscientiousness: It describes a person's ability to manage their impulses in order to participate in goal-oriented activities.
  - a. High Score: Directed, purpose, and self-disciplined and Dutifulness.
  - b. Low Score: Disorganized, Inept and irresponsible, Procrastinates and Indiscipline.
8. Extraversion: It reflects the willingness of a person to interact with their social environment.
  - a. High Score: Sociable, thrives on excitement and loves being the focus of attention. and outgoing.
  - b. Low Score: Prefers solitude, tired by an excessive amount of social contact, Reflective and dislikes being center of attention.
9. Agreeableness: It refers to people's attitudes toward other people's relationships.
  - a. High Score: Forgiveness, forthrightness, a desire to assist, and sympathy and empathetic.
  - b. Low Score: Demanding, Stubborn, show-off, insults others and does not care about other's feelings.
10. Neuroticism: It describes the emotional stability of a person through how they view the world.
  - a. High Score: Anxious, Stress, Vulnerability, Mood swings, irritable and shy
  - b. Low Score: Calm, Confident, Resilient, Emotionally Stable and Balanced.

To find the factors that influence Big Five, Researchers have conducted a twin studies with 123 pairs of identical twins and 127 pairs of fraternal twins, and found that the amount of variance that can be attributed to genes is 40-60%. [\[13\]](#)

### Algorithm for personality classification:

#### Cleaning of data:

The real world datasets are full of noise and need to be preprocessed before feeding it to the algorithm. This dataset had total participants of 10,15,341. After removing missing values,

redundant participants (more than two participants from same IP Address), the total number participants were 800543.

#### Reverse Marking:

The answers to each questions can be given in numbers in the range one to five, where one represents disagree and five represents agree. The reversed marked questions were subtracted from six to represent it in a proper manner.

#### Pre-processing:

In order to decrease the variance in the dataset, the features were normalized using “Min-Max Scaler”. By doing so, all features will be transformed into the range (0-1). The mathematical formulation is given below. Here,  $x$  represents a single feature.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

All the rescaling is done feature-wise, so it is independent.

#### Need for Clustering:

In order to classify the participants into K number of groups, we need to determine patterns in the dataset, and classify participants to clusters depending on each of their personality patterns. By considering all the participants from a country and determining the mode of participant’s respective clusters we get the dominant personality of a country.

#### Finding optimum number of clusters:

In order to classify the participants, first we need to identify the optimal number of groups in this dataset. To identify this, we can use the distortion score elbow method [for more information refer [docs](#)]. This method fits the model with a range of values to find the right number of clusters. It uses sum of squared distance (SSE) between the data points and their respective assigned clusters centroid. The optimal point is the value where the point SSE starts to flatten out, forming an elbow. In other words, the distortion score starts to decrease exponentially, the point after which it starts decreasing linearly is the optimum point.

In the fig (3), the range for ‘k’ was set from 2 to 14, and the elbow was observed at k=5(Blue line).

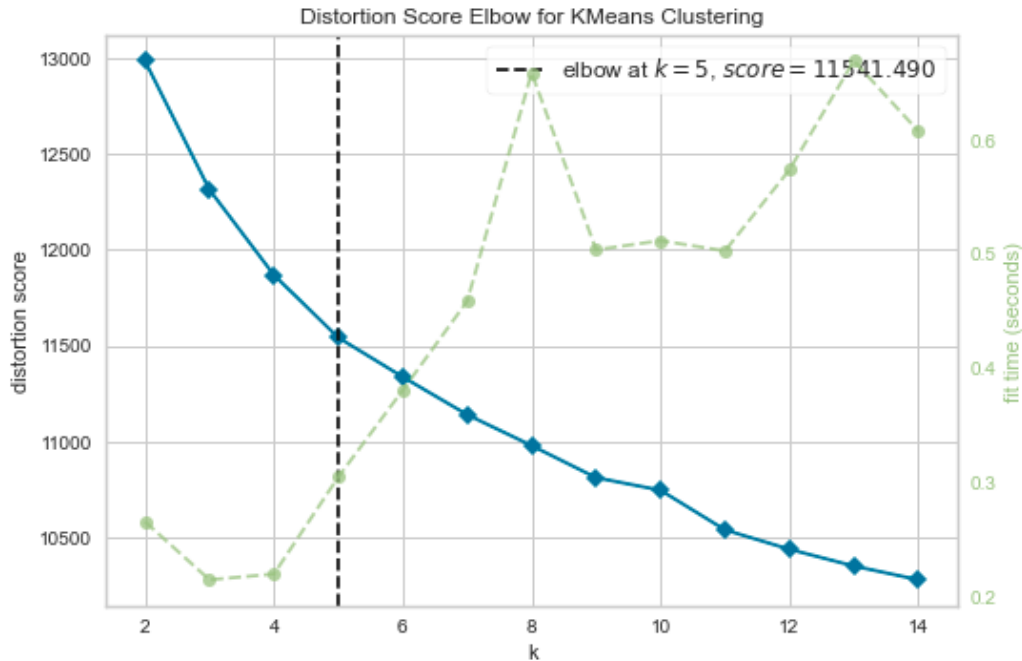


Fig.3. Elbow method of optimal k value

### K means Clustering:

The clustering algorithm chosen is K-Means. As it is fast (less memory intensive), simple. As we were looking for reducing the data to centroids, this algorithm was chosen.

The K Value is determined using the elbow method. For the first time the centroids are

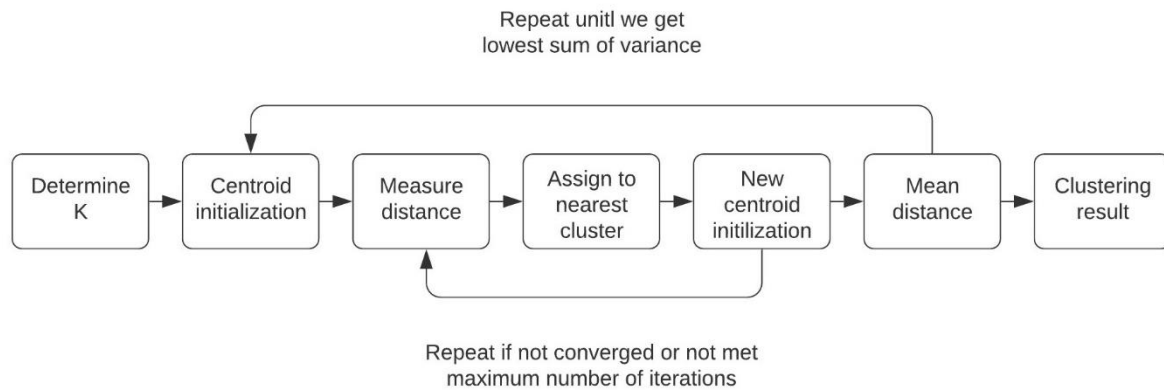


Fig.4. Working of clustering algorithm

randomly initialized. The Euclidian distance is calculated using the below expression.

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

The clusters are assigned to the nearest data points to the centroids. Centroids are now updated with respect to the mean of assigned values to them. Again the distance is computed, the position of centroids are updated and the sum of variance is recorded. The sum of variance should decrease after every iteration. If it remains the same then that will be end of iteration and the K means is said to be converged. This will be the result of K means algorithm.

After obtaining the Clustering result, we save the clusters in the clusters column in the dataset. For better understandability of each personality traits the average of each personality trait's questions was taken and saved in a new column with column name same as that of personality trait respectively.

For visualizing each clusters, the mean of all data points belonging to a cluster was considered. Using 'matplotlib' library the visualization of each cluster was achieved. Each of the clusters represented the pattern of personality trait of all the participants belonging to that cluster.



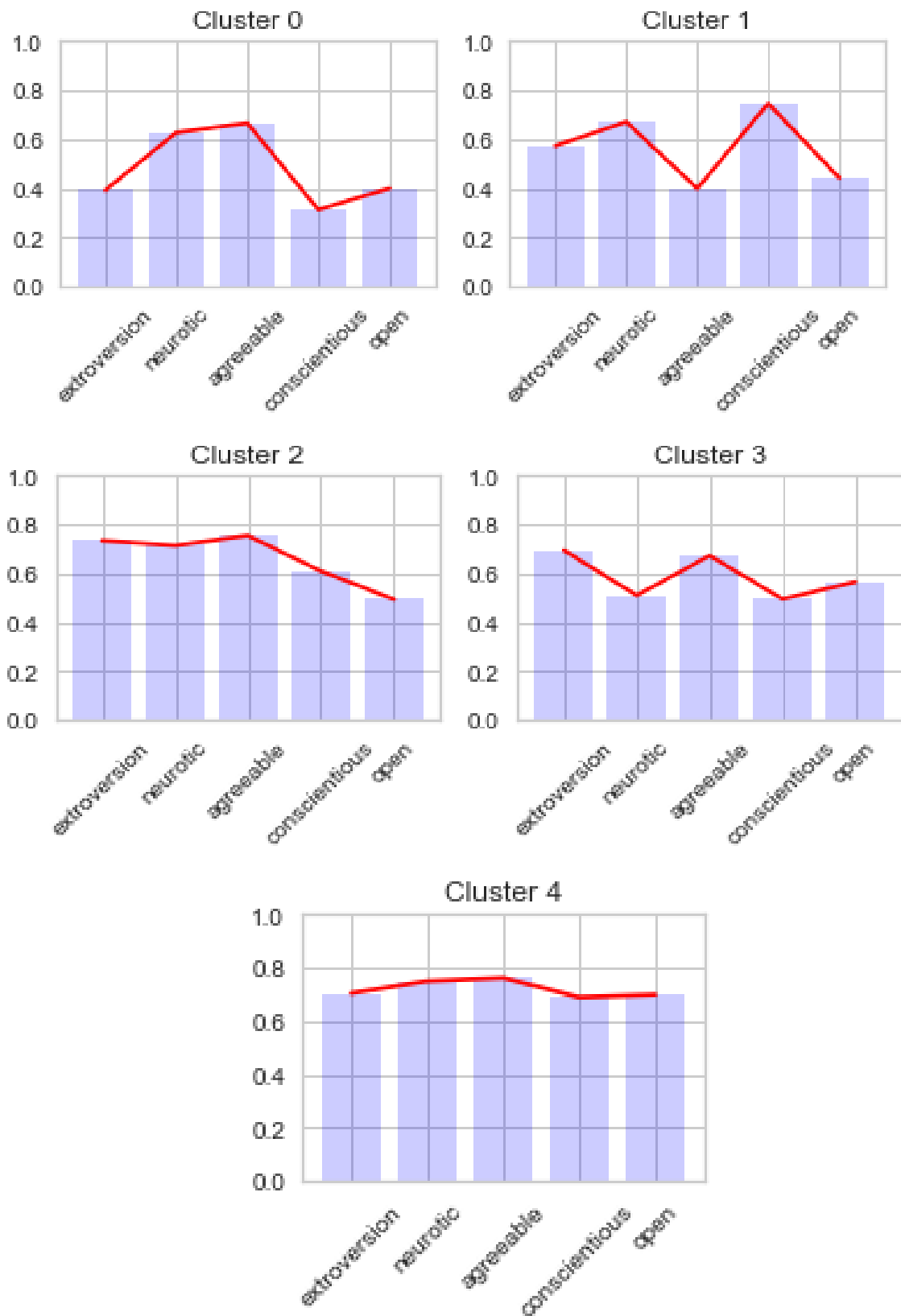


Fig.5. Average personality pattern of each clusters

For visualizing the data points as clusters, PCA (Principal Component Analysis) was used. PCA is used to create visualization of data that maximizes the variance of the projection coordinates while minimizing residual variance in the least squares sense. (for more information refer [docs](#))

PCA was used from sklearn and seaborn packages of python, the clusters were visualized.

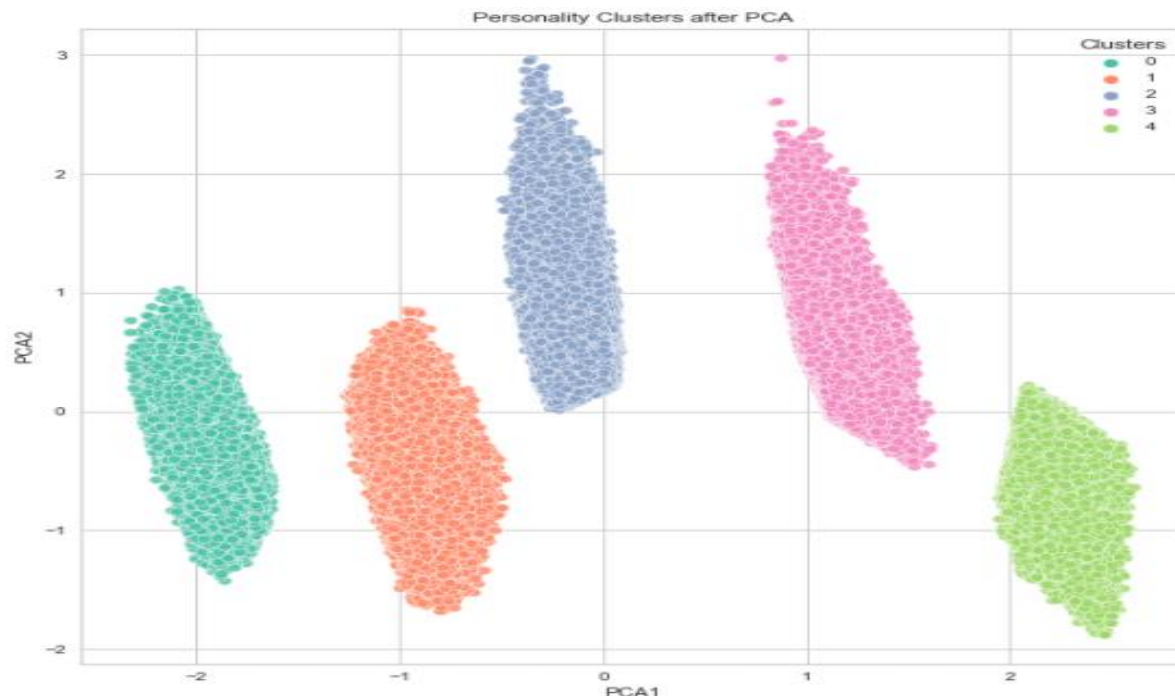


Fig.6. Cluster visualization using PCA

#### Country based dominant personalities:

As we have discussed in the dataset section that we have country column in our dataset which indicates the participant's country, we can use that to find the dominant personality of that country. First, we take all the participants from a country and consider their corresponding cluster, then we apply mode operation to get the dominant personality of the country.

From clustering the participants into groups, we are able to understand that the personality traits of a cluster, is the mean of personality traits of all participants belonging to that cluster. With the help of this clusters we can find the dominant personality of a county by simply considering the mode of cluster all the participants from that country.

```
In [15]: # Creating K-means Cluster Model
from sklearn.cluster import KMeans

# I define 5 clusters and fit my model
kmeans = KMeans(n_clusters=5, verbose=0)
k_fit = kmeans.fit(df)
kmeans.inertia_
```

Out[15]: 1879676.5329615066

Fig.7. Obtained inertia value

Due to the high dimensionality of the dataset with 50 columns i.e. 10 columns for each of five personality traits, the inertia obtained was quite high also called as ‘curse of dimensionality’. Although steps were taken to reduce dimensionality the model which in turn reduced the inertia value, it was observed that prediction was highly skewed. In order to avoid that, the dimensionality reduction was discarded.

#### Clustering of new participants:

If a new participant answers the questions, the new responses can also be put in appropriate cluster. By doing this, the accuracy of country based dominant personalities can be increased. We can take the average of each personality groups and find the individual’s dominant personality trait. New participant’s data can be used to classify to a cluster.

```
In [34]: my_data = pd.read_excel(r"H:\MY projects\project\Project\my_personality_test.xls")
my_data
```

```
Out[34]:
```

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	EST1	EST2	EST3	EST4	EST5	EST6
0	2	5	3	3	4	4	3	2	4	3	2	2	3	4	3	4

```
In [35]: my_personality = k_fit.predict(my_data)
print("My Personality Cluster: ", my_personality)
```

My Personality Cluster: [4]

Fig.8A. New participant’s cluster given by K Means model

Sum of my question groups

```
Out[44]:
```

	extroversion	neurotic	agreeable	conscientious	open	cluster
0	3.3	2.6	3.2	3.1	3.2	4

Fig .8B. Average scores of each personality traits

The average of each of personality trait’s responses are computed and stored in respective trait’s names. The dominant personality trait of the new participant is extroversion (as the average is highest compared to other traits).



**Conclusion:** In this project, we understood how clustering can be used to derive patterns in the dataset and classify the data points to corresponding clusters. We also understood how we can use these clusters to determine the approximate dominant personality of a country, also we saw how we can add new participants to our dataset to get more generalized model.

Future Scope: It can be used by the recommendation system to recommend products without the consumer’s order history. It can be used by hiring managers to determine the behavior of potential candidate and analyze if the candidate is right fit or not. It can also be used by the companies to get an insight whether a product succeeds in a country by analyzing the dominant personality. It can be used by teachers to find the optimal methods of teaching by analyzing the pupil’s personalities.

## References:

- Dataset: : <https://www.kaggle.com/tunguz/big-five-personality-test>
- Research Papers:
  1. Personality Research, Methods, and Theory. A Festschrift Honoring Donald W. Fiske. University of Massachusetts at Amherst
  2. A Study of Personality Influence in Building Work Life Balance Using Fuzzy Relation Mapping (FRM) by A.Victor Devadoss and J. Befija Minnie (2013)
  3. John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives
  4. Personality classification using Data mining approach by Rajalaxmi Hegde , Sandeep Kumar Hegde , Sanjana , Sapna Kotian and Shreya C Shetty. (2019)
  5. Automatic Personality Identification Using Students' Online Learning Behavior (2019)
  6. Personality-Aware Product Recommendation System Based on User Interests Mining and Metapath Discovery by Sahraoui Dhelim, Huansheng Ning, Nyothiri Aung, Runhe Huang and Jianhua Ma (2021)
  7. Predicting Personality Using Answers to Open-Ended Interview Questions by Madhura Jayaratne and Buddhi Jayatilleke (2020)
  8. Personality Dimensions Classification with EEG Analysis using Support Vector Machine by Fadhilah Qalibi Annisa, Eko Supriyanto, Sahar Taheri (2020)
  9. A Computational Personality Traits Analysis Based on Facial Geometric Features by Mingliang Xue; Xiaodong Duan; Yuangang Wang; Yuting Liu (2019)
  10. Empirical Study on Personality Trait Classification by Food Related Preferences by Tasfia Hoque, Raqeebir Rab, Khushnoor Rafsan Jani Alam, Saif Hasan Khan, M. A. Wadud Shuvro and Umme Zakia (2019)
  11. Using Classification to Determine Whether Personality Profiles of Countries Affect Various National Indexes by Emine Yaman, Azra Musić-Kiliç, Zaid Zerdo (2018)
  12. The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives by Oliver P. John and Sanjay Srivastava. University of California at Berkeley (1999)
  13. Heritability of the Big Five Personality Dimensions and Their Facets: A Twin Study by Kerry L. Jang, W. John Livesley and Philip A. Vernon (1996).

### INFORMATION REGARDING STUDENTS

STUDENT NAME	EMAIL ID	PERMANENT ADDRESS	PHONE NUMBER	LANDLINE NUMBER	PLACEMENT DETAILS	PHOTOGRAPH
SUMUKHA A	<a href="mailto:SUMUKHAA204@GMAIL.COM">SUMUKHAA204@GMAIL.COM</a>	G002 MANTRI WOODLANDS BANNERGHATTA ROAD AREKERE GATE BANGALORE	9606626914	42049224	NOT PLACED	
SAI RITWIK REDDY K	<a href="mailto:RRITWIK586@GMAIL.COM">RRITWIK586@GMAIL.COM</a>	PLOT NO 152, SAI RAM NAGAR COLONY, CHAMPAPE, HYDRABAD	8884880644	24075657	OPTED HIGHER STUDIES	

**BATCH PHOTOGRAPH ALONG WITH GUIDE**



Dr. Manjula TR  
Assistant Professor



Sumukha A  
17BTREC108



Sai Ritwik Reddy K  
17BTREC006