ABSTRACT:

This study employs clustering analysis, an unsupervised learning technique, to uncover hidden patterns in a large-scale dataset consisting of personality traits.

The aim of the study is to predict the intensity of the Big Five Personality Traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). The dataset includes responses from over 10 lakh individuals, each responding to a set of questions about their personality traits, with each question having a 5-point agreement scale. The study aims to identify groups of individuals with similar personality traits using the clustering techniques, k-means clustering, hierarchical clustering. The results of the clustering analysis will be evaluated using various metrics, such as silhouette coefficient and elbow method. The study will also explore and predict the relationship between the identified clusters and the Big Five Personality Traits, which can provide valuable insights into the nature of personality traits and their manifestation in individuals. The findings of this research can be useful in various fields, such as psychology, sociology, and marketing, where the understanding of personality traits and their relationships can be beneficial.

We later deploy this model on the web using the streamlit library. The website gives us the information about the test and allows us to give the test and get results immediately.

BACKGROUND:

Personality refers to the unique set of characteristics, traits, behaviors, and patterns of thought that define an individual and distinguish them from others. It is the way in which we perceive, interact with, and respond to the world around us. Personality is believed to be a combination of genetic, environmental, and social factors, and it can influence every aspect of our lives, from our relationships and career choices to our emotional well-being and overall happiness.

Everyone should be aware of their own personality traits because it can help them understand themselves better, improve their communication and relationships with others, and make informed decisions about their personal and professional lives. By understanding your own personality traits, you can identify your strengths and weaknesses, and work on areas that need improvement. It can also help you understand how you react to different situations and why you may be more drawn to certain activities or types of people.

In addition, being aware of your own personality can help you better understand and interact with others. By recognizing different personality traits in others, you can adjust your communication style and behavior to better fit their needs and preferences, leading to more harmonious relationships and effective collaboration.
Overall, understanding your own personality can lead to increased self-awareness, personal growth, and more fulfilling relationships and experiences in life.

The Big Five personality traits model is a widely recognized framework for describing and understanding human personality. It is also known as the Five Factor Model (FFM) and consists of five broad dimensions of personality. These dimensions are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. Let's explore each of these dimensions in detail:

1. Openness to experience: This dimension refers to a person's openness and receptivity to new experiences, ideas, and perspectives. People who score high in openness tend to be curious, imaginative, creative, and open-minded. They enjoy exploring new ideas and concepts and are often interested in art, music, and culture. People who score low in openness tend to be more traditional, practical, and conventional.

2. Conscientiousness: This dimension refers to a person's level of organization, responsibility, and self-discipline. People who score high in conscientiousness tend to be reliable, organized, and hard-working. They set high standards for themselves and take their obligations seriously. People who score low in conscientiousness tend to be more relaxed, spontaneous, and carefree.

3. Extraversion: This dimension refers to a person's level of sociability, assertiveness, and energy. People who score high in extraversion tend to be outgoing, talkative, and energetic. They enjoy being around others and often seek out social interactions. People who score low in extraversion tend to be more reserved, quiet, and introspective.

4. Agreeableness: This dimension refers to a person's level of empathy, cooperation, and kindness. People who score high in agreeableness tend to be compassionate, cooperative, and caring. They value harmony and tend to

be supportive of others. People who score low in agreeableness tend to be more competitive, assertive, and sometimes critical.

5. Neuroticism: This dimension refers to a person's level of emotional stability, anxiety, and sensitivity. People who score high in neuroticism tend to be more anxious, worried, and emotionally reactive. They may be more sensitive to stress and may experience a wider range of emotions. People who score low in neuroticism tend to be more emotionally stable and resilient.

The Big Five personality traits model is widely used in psychology research and has been found to be reliable and valid across different cultures and populations. It can be useful in many areas, including personal growth, career development, and relationships. By understanding our own personality traits and those of others, we can develop greater self-awareness and improve our interactions with others.

The Big Five personality traits model works by measuring an individual's level of each of the five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. There are various methods used to measure these dimensions, including self-report questionnaires, observer ratings, and interviews.

METHODOLOGY:

The Big Five personality traits model works by measuring an individual's level of each of the five dimensions of personality: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. There are various methods used to measure these dimensions, including self-report questionnaires, observer ratings, and interviews.

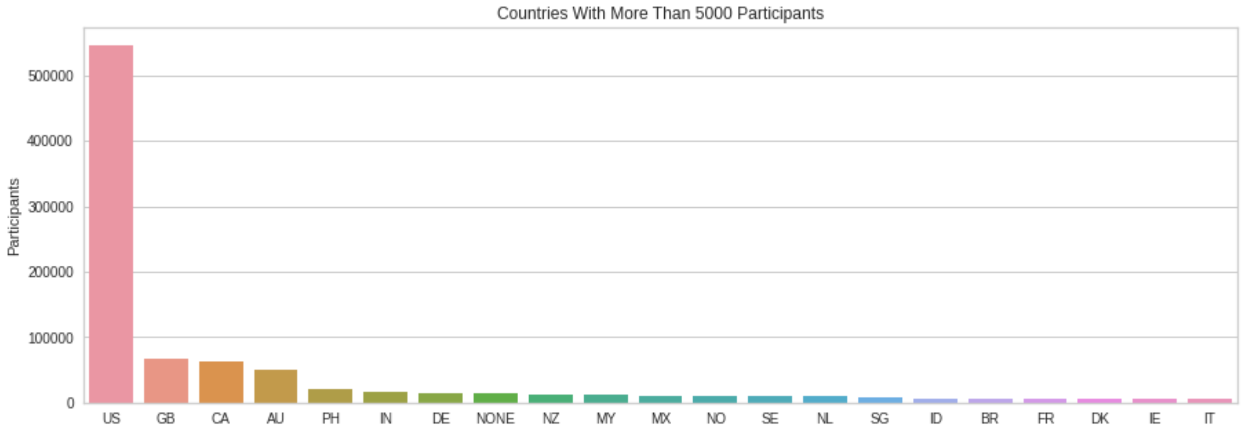
In self-report questionnaires, individuals answer a series of questions that are designed to assess their personality traits. For example, a questionnaire might ask how much a person agrees or disagrees with statements such as "I am imaginative" (openness), "I am careful and diligent" (conscientiousness), "I am outgoing and sociable" (extraversion), "I am considerate and kind to others" (agreeableness), and "I worry a lot" (neuroticism). The responses are then scored and analyzed to provide a profile of the individual's personality traits.

Exploring our Dataset

Now, the dataset that we are working with is a self-report questionnaire having 50 questions overall, 10 questions for each of the five traits. Along with those questions, the time taken by an individual for answering each of the questions is also provided. Although we will not be analysing the time aspect of the questions so we will be dropping those columns.

The dataset had around 89 thousand missing values in total. Since the size of the dataset is around 10 Lakh, we can afford to drop those datapoints without compromising on our model building step.

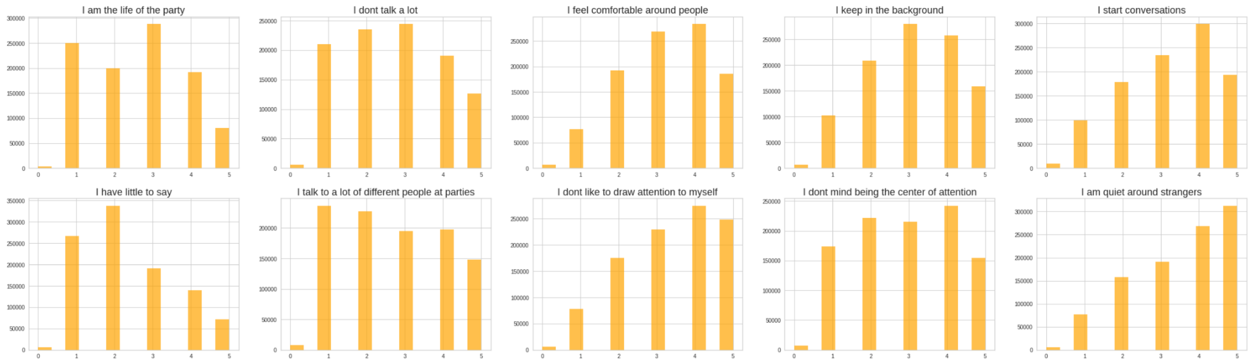
The country of the individuals participating in the study was also included and the following bar plot shows us the countries with more than 5000 participants.



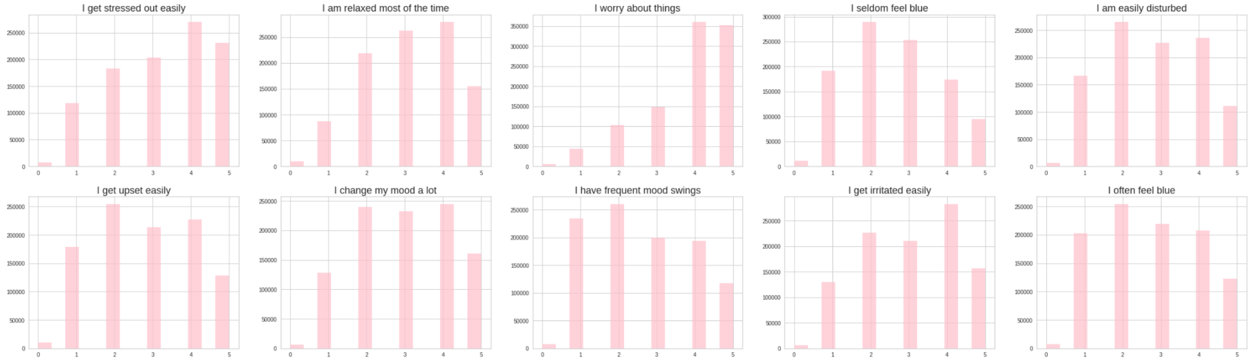
As we can observe, out of our 10 lakh individuals, more than half of them are from the United States of America followed by Great Britain and Canada.

The questions asked for each of the traits are of the 5 point agreement scale and we can visualize them using a frequency plot or bar plot. Each of the sub-graphs explore the distribution of answers of each of the questions which is the title of each. The X-axis is the 5-point scale in addition to 0 (meaning not answered) and the Y-axis is the number of individuals.

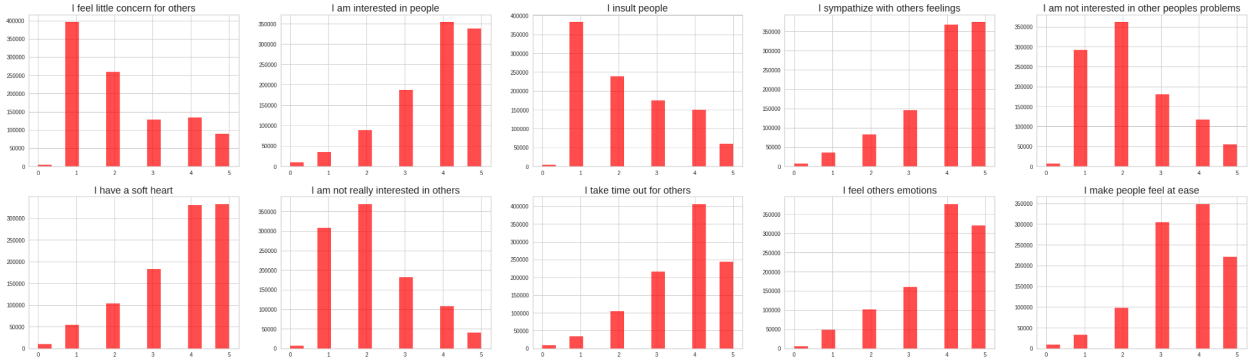
The questions related to the trait Extraversion are:



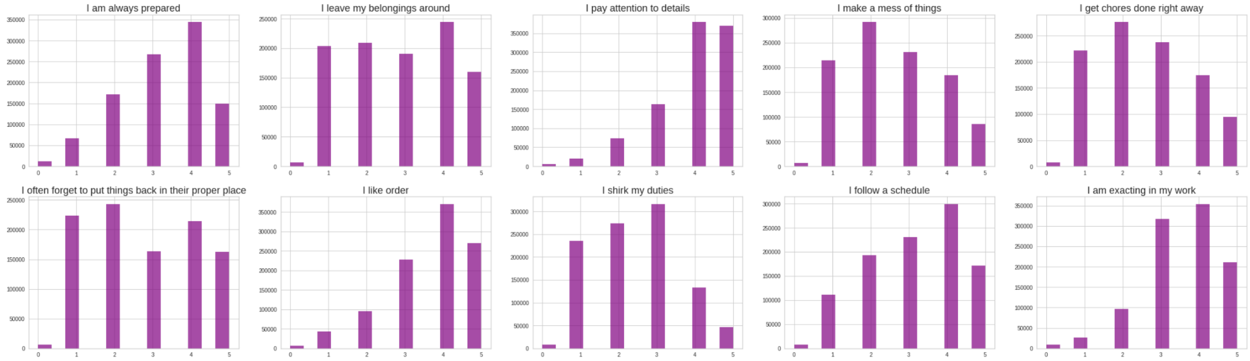
The questions related to the trait Neuroticism are:



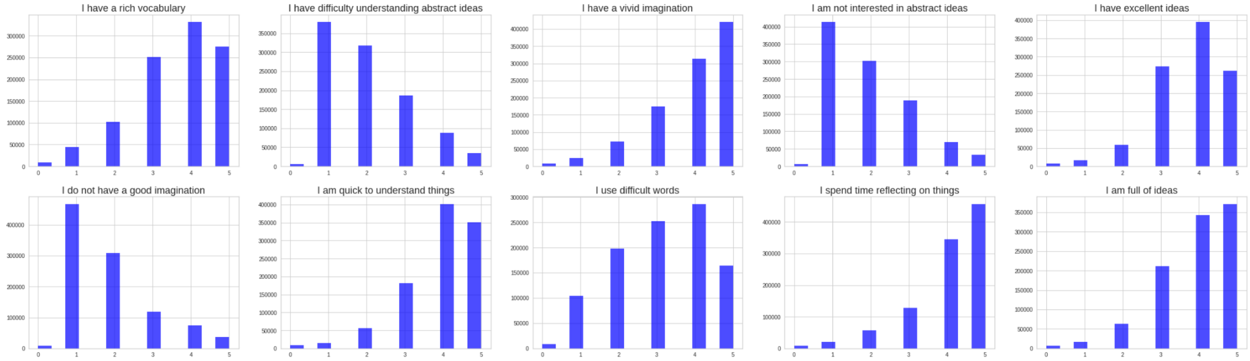
The questions related to the trait Agreeableness are:



The questions related to the trait Conscientiousness are:



The questions related to the trait Openness to Experience are:



Clustering

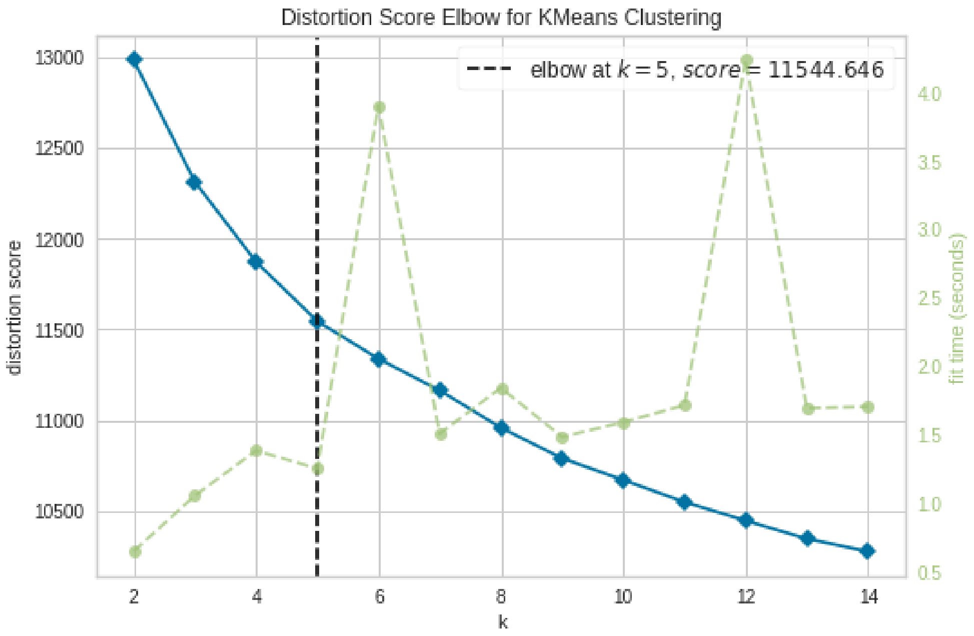
Clustering is a unsupervised machine learning technique is used to group similar data points together in a dataset. We will be using the popular algorithm of Kmeans to cluster our data points together.

K-means clustering is an iterative algorithm that works by dividing a set of observations into k clusters based on their similarities. It starts by randomly selecting k initial cluster centroids, where k is a predefined number of clusters. Each observation is then assigned to the cluster whose centroid is closest to it in terms of Euclidean distance.The algorithm then computes the mean of all the observations in each cluster, and moves the centroid to that mean. This process is repeated until the centroids no longer move or until a maximum number of iterations is reached.

One of the parameters that we need to pass to the Kmeans algorithm is the number of clusters. Now, it is important to have the optimal number of clusters in KMeans clustering because it directly affects the quality of the clustering results. If the number of clusters is too low, then the clusters may be too broad, resulting in loss of important information and patterns in the data. On the other hand, if the number of clusters is too high, then the clusters may be too specific, resulting in overfitting and unnecessary complexity.

Having the optimal number of clusters helps to find the right balance between these two extremes, resulting in more accurate and useful clustering results. This can be particularly important in applications such as customer segmentation or anomaly detection, where the clustering results can have significant impacts on business decisions.

For getting the optimal number of clusters, we use the KElbowVisualizer function from yellowbricks library. It works by plotting the sum of squared distances of samples to their closest cluster center (inertia) against the number of clusters. The "elbow" point on the plot indicates the optimal number of clusters where adding more clusters does not significantly reduce the inertia anymore. The visual representation of the method is as follows:



As we can observe, at value of k being 5, the fitting time of model appears to drop drastically and we observe the distortion score of 11544.646. Taking into account the large number of data points and our computational capacity, we can finalize the number of clusters as 5 and move forward with fitting the Kmeans model to the data.

Before we start with model fitting we first use the MinMaxScaler function to scale the data points in the range 0 to 1. We opt for the MinMaxScaler as we know the precised range in which the readings lie. On the other hand, if we wouldn’t have known the range, we would’ve chosen the StandardScaler function.

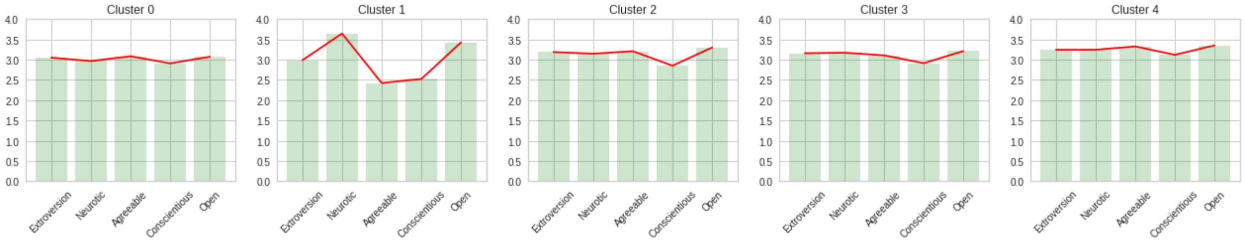
Now that we have our data scaled, we fit the Kmeans model with number of clusters as 5 and we get the silhouette score of 0.0711, which is pretty satisfactory for the scale of dataset we are dealing with.

Factor Analysis

Factor analysis is a statistical technique used to uncover the underlying structure of a set of variables. Specifically, it seeks to identify the common factors that contribute to the observed variation in a set of variables. Factor analysis can be used for a variety of purposes, including data reduction, construct validation and hypothesis generation.

The reason we are performing factor analysis is for construct validation. Since we know the way questionnaire was built, we will use that information to our aid. In the questionnaire, each question belongs to one of the five personality traits in our model and each trait has exactly 10 questions. EXT_1 to EXT_10 columns store the answers for the Extraversion trait, OPN_1 to OPN_10 store the answers for Openness to Experience trait, and so on.

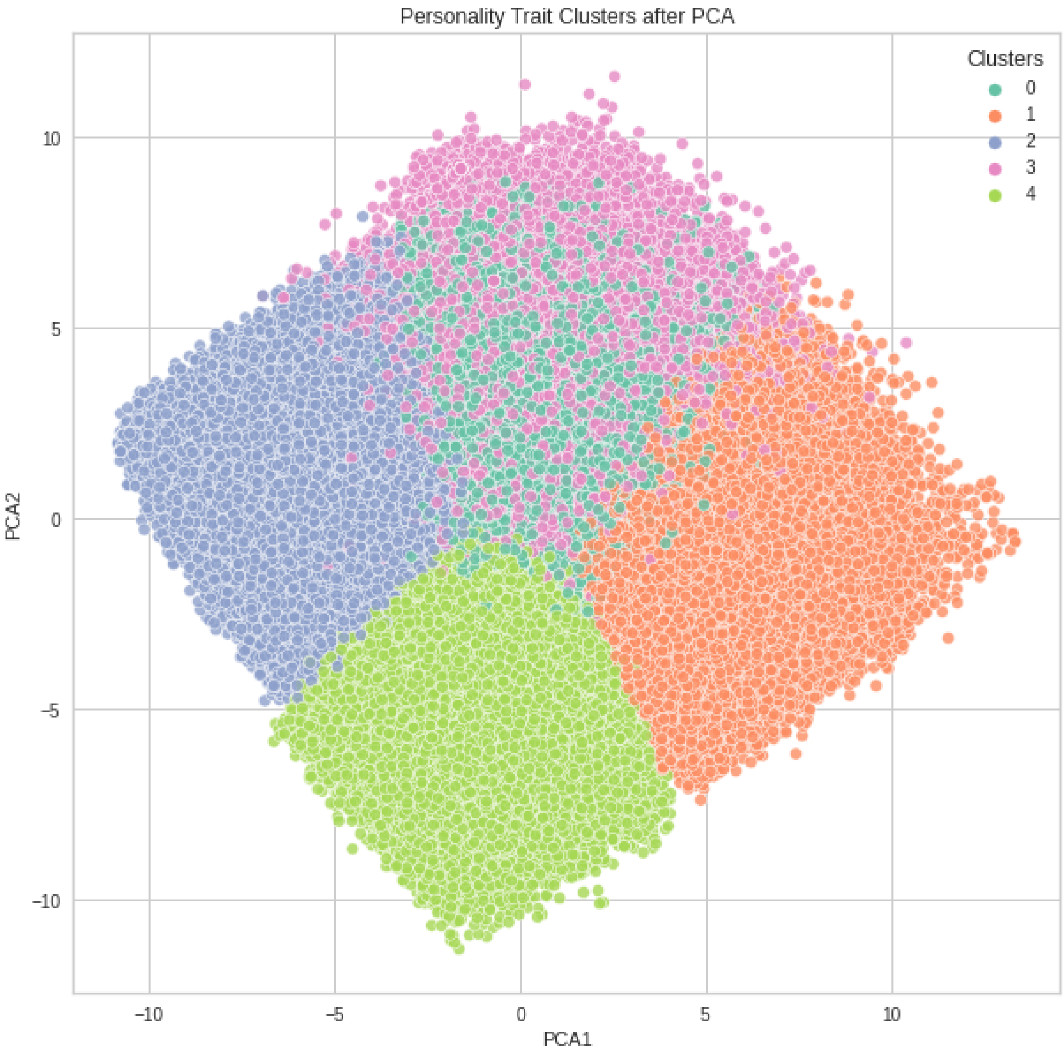
Hence, we combine the columns with respect to the traits they belong to and take the mean of the datapoints to aggregate. We plot a bar graph for each of the clusters showing the means of each of the traits below:



Visualizing the Clusters

After performing Factor Analysis, we have our five factors for each of the five traits in our model. But, we cannot visualize the data points since this is in 5 dimensions. Hence, for ease in visualization, we further perform Principal Component Analysis in order to reduce the dimensionality to the 2-D space.

The visual representation of our data points in the 2D space along with clusters is:



Difficulties faced

Due to the sheer scale of data that I chose, fitting and visualising was hard to perform on the standard notebook we get in Google Colaboratory. Even while running the KElbowVisualizer function, I had to run it on the initial 5000 data points only.

The scale not only restricted in visualising using TSNE but also in the number of models I could experiment with. For fitting the DBSCAN algorithm, I had to reduce the dataset size to the initial 1 Lakh data points instead of all the 10 Lakh data points. The DBSCAN model was fitted with the parameters of epsilon equal to the silhouette score we had got in the Kmeans model i.e. 0.0711 and min_samples as 50. Even with these parameters, the model classified 99% of the data points as noise and the other 1% were grouped in 3 clusters.

We observe some similar results when we applied the OPTICS algorithm for clustering with similar parameters and min_samples as 5 this time. Even with these parameters, the model classified 99% of the data points as noise and the other 1% were grouped in 7 clusters.

The results of both the DBSCAN and OPTICS model seemed unsatisfactory to me and hence I need to better understand how these algorithms work inorder to optimize the fitting.

Deployment using Streamlit:

Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc. With Streamlit, no callbacks are needed since widgets are treated as variables. Data caching simplifies and speeds up computation pipelines. Streamlit watches for changes on updates of the linked Git repository and the application will be deployed automatically in the shared link.

We have used our k-means model as the prediction model. The website involves the whole test for you to give it a try. We can see the results and interpret it right on the website itself.

CONCLUSION:

In this study, we were able to perform cluster analysis on a huge data set using the Kmeans clustering algorithm. We got the optimal silhouette score of 0.0711 on this model. We performed Factor analysis for constructing validation for the Big Five Personality Trait model and then predict the scores for each trait for any individual.

Further improvements are required in this study with better resources and better understanding of the underlying mathematics of each model.

REFERENCES:

1. The Big Five Personality Traits
2. Jupyter Notebook
3. Drive Folder
4. Dataset
5. Meta Data
6. Model file