Iteration 03: Project Methodology

Julian Benitez Mages, Anaelle Surprise

March 11, 2025

1 Introduction

This paper explores machine learning techniques for analyzing personality data collected through an online interactive test based on the Big Five Factor Model. The dataset contains over one million responses from 2016 to 2018, measuring personality traits using the International Personality Item Pool (IPIP).

The first of our project's objectives is to explore clustering for personality survey scores. From the survey results, the IPIP projects instructions to measure overall scores for each of the five traits, and we are interested in clustering this data into distinct, separate "personalities".

The second objective of our project is to use supervised predictive models to better understand the relationship between factors in our dataset. We are interested in analyzing the importance of geographical factors such as the user's country (provided in the dataset) in predicting overall personality scores. Additionally, we want to understand the relationship between time spent answering the survey questions and personality.

Ultimately, upon finalization of personality clusters, we will use Large Language Models to generate descriptions of each cluster using natural language to describe the different "personalities".

Our project will culminate in a front-end environment where the user can take the IPIP personality test for themselves, and our models will be applied to their individual result. The user will thus be able to see which personality cluster they belong to, and a prediction of their country and geographic coordinates.

1.1 Background

The Big Five personality traits, also known as the Five Factor Model, are among the most widely used frameworks for psychological assessment. These traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—were derived from linguistic analyses of descriptors used to characterize human behavior. The dataset used in this study was obtained through an interactive online personality test constructed using the "Big-Five Factor Markers" from the IPIP [1].

1.2 Defining the Problem

2 Dataset Collection & Preparation

2.1 Data Sources

Our data set contains 1,015,342 instances of Big Five survey results from the test administered by Open Psychometrics. We obtained access to the dataset via a Kaggle posting, courtesy of Bojan Tunguz.

2.2 Data Description

The dataset contains 1,015,342 rows and 110 columns, totaling 416.27 MB of data. The columns contain the following information:

- Survey Question Responses
- Time spent per survey question
- Time spent on landing/finalization pages
- Survey Time Stamp
- User Screen Size
- Number of Records per IP address
- User's Country and Approximate Coordinates

The survey contains 10 statements corresponding to each personality trait, with the user asked to rate their relatability to the statement on a scale from 1 to 5. Statements corresponding to Openness are coded as 'OPN'; Statements corresponding to Conscientiousness are coded as 'CSN'; Statements corresponding to Extraversion are coded as 'EXT'; Statements corresponding to Agreeableness are coded as 'AGR', and statements corresponding to Neuroticism are coded as 'EST'. A protion of the questions are reversed, meaning that the statement embodies a lack of the corresponding trait. For these questions, the 1-to-5 scale is reversed, with 5 indicating minimal relatability, and 1 indicating maximum relatability. The full list of statements is provided in the appendix.

The time spent columns measure the number of milliseconds the user spent on each question. This information was measured by taking the difference in time spent between clicking the answer for subsequent survey questions. There is also information on time spent on the opening page, the survey questions page, and the finalization page, totaling 53 such columns.

Among the remaining columns, the dataset contains dimensions of the user's screen in pixels, the number of records from the user's IP address, the user's country, and the user's approximate longitude and latitude.

2.3 Data Pre-processing

To prepare the data for clustering and modeling, there were a number of pre-processing steps we had to take. Upon obtaining the dataset from the 'kagglehub' package, we had to programatically convert the data from raw CSV form, and convert numeric columns into numeric data types.

We filtered the dataset to only include rows where the column corresponding to records for the user's IP address is 1, as we wanted each row to correspond to just one survey response. High values of column are likely attributed to university or workplace contexts, where several users took the survey under the same IP address, or one user took the survey multiple times. The categorical variable country was encoded using LabelEncoder to transform it into numerical values suitable for modeling.

Our most important pre-processing step involved calculating the total scores for each trait, for each survey response. We followed the scoring instructions provided by the IPIP, which directed us to sum the scores for the questions corresponding to each trait, resulting in five new columns, one for each trait score.

Lastly, we removed rows with N/A values under the first 50 columns (survey response columns) as these incomplete survey results were unable to generate accurate trait scores.

Upon completion of our pre-processing steps, our dataset now contained 696,845 rows and 115 columns, totaling 451.16 MB of data.

3 Model Selection & Development

3.1 Clustering

For clustering the data, our initiative was to apply a range of clustering algorithms onto the trait score columns which we had calculated during data pre-processing. We deployed the following clustering algorithms onto our data:

- K-Means Clustering
- Gaussian Mixture Modeling
- Hierarchical Clustering
- DBScan

Prior to testing the aforementioned algorithms, we separated the data into three sections:

- Survey Results Data
- Time Spent Data
- Trait Score Data

We split the data into three tables, and used scikit-learn's StandardScaler to normalize the data for a mean of 0 and a standard deviation of 1.

3.2 Predictive Modeling

Traditional machine learning models were considered over deep learning models due to the dataset being structured tabular data. Between a Random Forest Classier, XGBoost and Logistic Regression and Support Vector Machines, we chose a Random Forest Classifier. At first the Random Forest Classifier was chosen for its ability to handle both categorical and numerical features, and the efficacy of the class$_{w}eightparameter.Buttheruntimewasstillquiteexcessive.SupportVectorMachineswhileeffect$

For this model, we used the respective OCEAN scores, and timelapsed variables to predict the country. The model handled the immense class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). We applied SMOTE with a custom sampling strategy to ensure that no class had more than 5000 samples to aid in the model's bias toward majority classes. Then we trained the model with XGBoost.

The parameters chosen were 300 number of trees valuing accuracy over speed due to the low predictive power of our features. The tree depth is 7 to prevent overfitting. The stepsize chosen was .1, again to improve accuracy.

3.3 Feature Modeling

4 Model Evaluation & Comparison

4.1 Clustering

For evaluation and comparison of clustering models, we compared different statistics to arrive at the optimal model and number of clusters.

To evaluate our model selection and n-clusters, we relied on the Davies-Bouldin score and the Calinski-Harabasz scores. We chose the KMeans and Gaussian Mixture Model algorithms for their optimal scores, and plotted these measures as a function of n-clusters for each model, obtaining the following results:

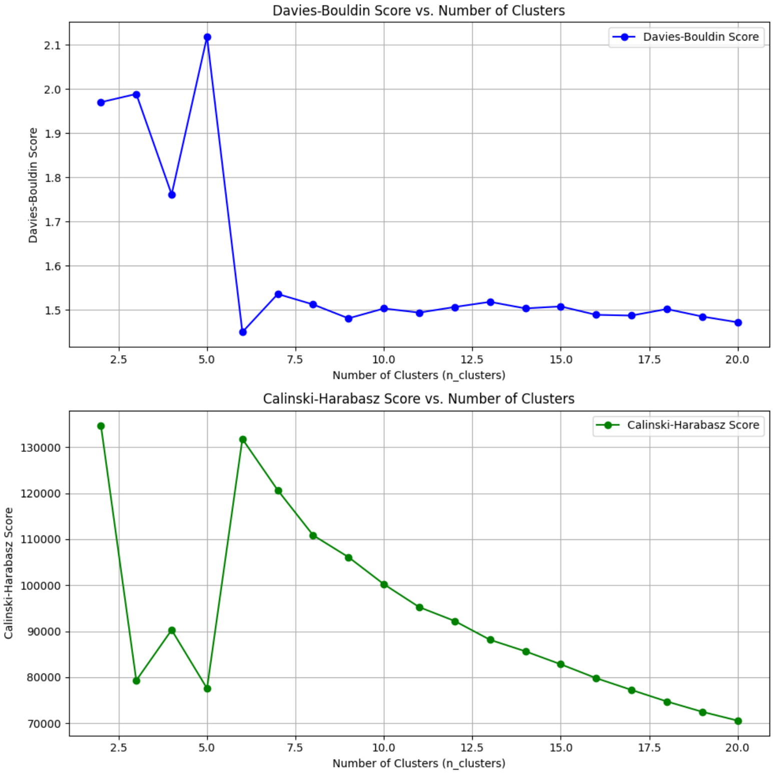

Figure 1: Evaluation of KMeans Clusters

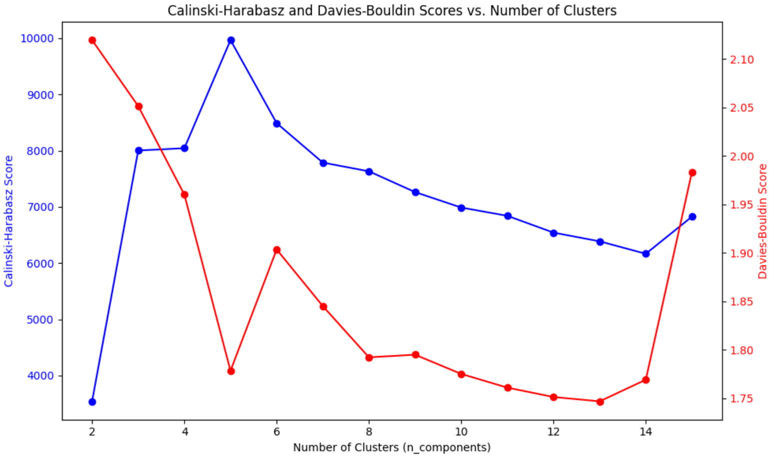

Figure 2: Evaluation of GMM Clusters

The findings from these results pointed to an optimal number of clusters of 6 for KMeans clustering, and 5 for Gaussian Mixture Modeling.

4.2 Feature Modeling

A Survey Statements

Below are the survey statements used in the dataset, grouped by personality trait:

A.1 Openness (OPN)

- OPN1: I have a vivid imagination.
- OPN2: I have excellent ideas.
- OPN3: I am full of ideas.
- OPN4: I am quick to understand things.
- OPN5: I use difficult words.
- OPN6: I spend time reflecting on things.
- OPN7: I am interested in abstract ideas.
- OPN8: I do not have a good imagination. (Reversed)
- OPN9: I have difficulty understanding abstract ideas. (Reversed)
- OPN10: I do not enjoy thinking about theoretical ideas. (Reversed)

A.2 Conscientiousness (CSN)

- CSN1: I am always prepared.
- CSN2: I follow a schedule.
- CSN3: I get chores done right away.
- CSN4: I pay attention to details.
- CSN5: I like order.
- CSN6: I make plans and stick to them.
- CSN7: I do things according to a plan.
- CSN8: I waste my time. (Reversed)
- CSN9: I do not finish what I start. (Reversed)
- CSN10: I find it difficult to get down to work. (Reversed)

A.3 Extraversion (EXT)

- EXT1: I am the life of the party.
- EXT2: I talk a lot.
- EXT3: I keep in the background. (Reversed)
- EXT4: I do not talk a lot. (Reversed)
- EXT5: I feel comfortable around people.
- EXT6: I start conversations.
- EXT7: I have little to say. (Reversed)
- EXT8: I don't mind being the center of attention.
- EXT9: I am quiet around strangers. (Reversed)
- EXT10: I am reserved. (Reversed)

A.4 Agreeableness (AGR)

- AGR1: I am interested in people.
- AGR2: I sympathize with others' feelings.
- AGR3: I have a soft heart.
- AGR4: I take time out for others.
- AGR5: I make people feel at ease.
- AGR6: I feel others' emotions.
- AGR7: I have difficulty understanding others. (Reversed)
- AGR8: I am not really interested in others. (Reversed)
- AGR9: I insult people. (Reversed)
- AGR10: I feel little concern for others. (Reversed)

A.5 Neuroticism (EST)

- EST1: I get stressed out easily.
- EST2: I worry a lot.
- EST3: I am easily disturbed.
- EST4: I get upset easily.

- EST5: I have frequent mood swings.
- EST6: I get irritated easily.
- EST7: I often feel blue.
- EST8: I am relaxed most of the time. (Reversed)
- EST9: I seldom feel blue. (Reversed)
- EST10: I do not get stressed out easily. (Reversed)

References

[1] International Personality Item Pool, https://ipip.ori.org/.