Report on Behavioral Patterns Analysis for Disease Prediction in Psychology

Student Name: Ayse Asli Ilhan

Student ID: 22036435

Topic: Advanced Algorithms and Complexity Final Project

I. Introduction

A. Problem Statement

The field of mental health holds great promise and importance for the prediction of illnesses prior to their severe manifestation. Mental health illnesses present a significant challenge to both patients and clinicians since they are sometimes obscure and complicated in nature. Not only can early behavioral pattern analysis prediction of these disorders allow for prompt and focused therapies, but it also represents a major advancement in customized healthcare. With the goal of bridging the gap between behavioral signals and their psychological ramifications, this initiative is rooted in the urgent need to comprehend and anticipate mental health disorders. The initiative aims to disentangle the complex web of behavioral markers that predate mental health issues by utilizing computational power and technological breakthroughs.

B. Project Objective

This project's main goal is to design and create an intricate predictive system that is supported by a combination of data analytics, sophisticated algorithms, and psychological insights. The purpose of this method is to identify patterns that may point to possible mental health illnesses by sorting through the complexity of human behavior as it is represented in many types of data. The objective goes beyond simple forecasting; it involves developing a tool that might offer useful insights, empowering medical practitioners to create preventative plans and interventions. The project's main goal is to use data-driven analysis to change the way mental health care is provided so that it is more effective, proactive, and tailored to each patient's unique needs.

II. Data Collection and Preprocessing

A. Data Collection

In order to create a complete dataset, the data collection phase was multidimensional and used a varied range of sources. Important references comprised:

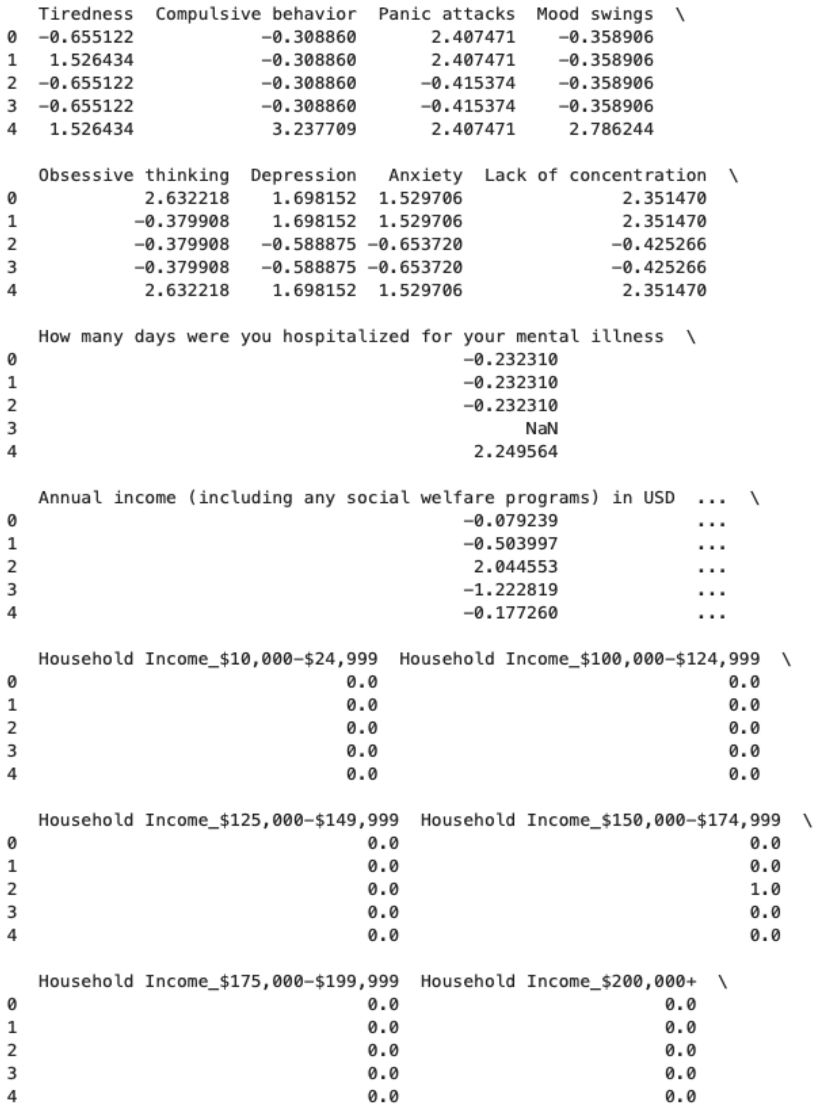
1. **Patient Behavior Logs**: Comprehensive records that document many facets of patient behavior and provide insights into daily trends and routines.
2. **Survey Responses**: A compilation of answers from specific surveys that offer individualized, subjective perceptions of people's mental health.
3. **Physiological Data**: Quantitative information about physical health metrics that are frequently associated with mental health issues.
4. **Social Media Data**: To perform Natural Language Processing (NLP), tweets were gathered. Unfiltered insights into public emotion and behavior expression are provided by this data, which is essential for comprehending mental states.
5. **Survey Outcomes**: To learn more about the relationships between self-reported mental health statuses and other characteristics, these were investigated.

Strict adherence to ethical norms was maintained to guarantee the integrity and privacy of the gathered data, particularly with regard to patient confidentiality and data security.

B. Data Preprocessing

Preprocessing entailed a number of actions to polish and convert the unprocessed data into an analysis-ready format, including:

1. **Loading Dataset**:
    - Pandas DataFrames were utilized to load data from several sources, such as CSV files that included survey and behavioral logs, for preliminary processing and examination.



```
   Tiredness  Compulsive behavior  Panic attacks  Mood swings  \
0  -0.655122            -0.308860       2.407471    -0.358906
1   1.526434            -0.308860       2.407471    -0.358906
2  -0.655122            -0.308860      -0.415374    -0.358906
3  -0.655122            -0.308860      -0.415374    -0.358906
4   1.526434             3.237709       2.407471     2.786244

   Obsessive thinking  Depression   Anxiety  Lack of concentration  \
0            2.632218    1.698152  1.529706               2.351470
1           -0.379908    1.698152  1.529706               2.351470
2           -0.379908   -0.588875 -0.653720              -0.425266
3           -0.379908   -0.588875 -0.653720              -0.425266
4            2.632218    1.698152  1.529706               2.351470

   How many days were you hospitalized for your mental illness  \
0                                          -0.232310
1                                          -0.232310
2                                          -0.232310
3                                                NaN
4                                           2.249564

   Annual income (including any social welfare programs) in USD  ...  \
0                                          -0.079239            ...
1                                          -0.503997            ...
2                                           2.044553            ...
3                                          -1.222819            ...
4                                          -0.177260            ...

   Household Income_$10,000-$24,999  Household Income_$100,000-$124,999  \
0                               0.0                                 0.0
1                               0.0                                 0.0
2                               0.0                                 0.0
3                               0.0                                 0.0
4                               0.0                                 0.0

   Household Income_$125,000-$149,999  Household Income_$150,000-$174,999  \
0                                 0.0                                 0.0
1                                 0.0                                 0.0
2                                 0.0                                 1.0
3                                 0.0                                 0.0
4                                 0.0                                 0.0

   Household Income_$175,000-$199,999  Household Income_$200,000+  \
0                                 0.0                         0.0
1                                 0.0                         0.0
2                                 0.0                         0.0
3                                 0.0                         0.0
4                                 0.0                         0.0
```

1. **Handling Missing Values**:
    - To preserve the integrity of the datasets, missing values were carefully managed within them using techniques such as filling with mean or median values.
2. **Feature Engineering**:

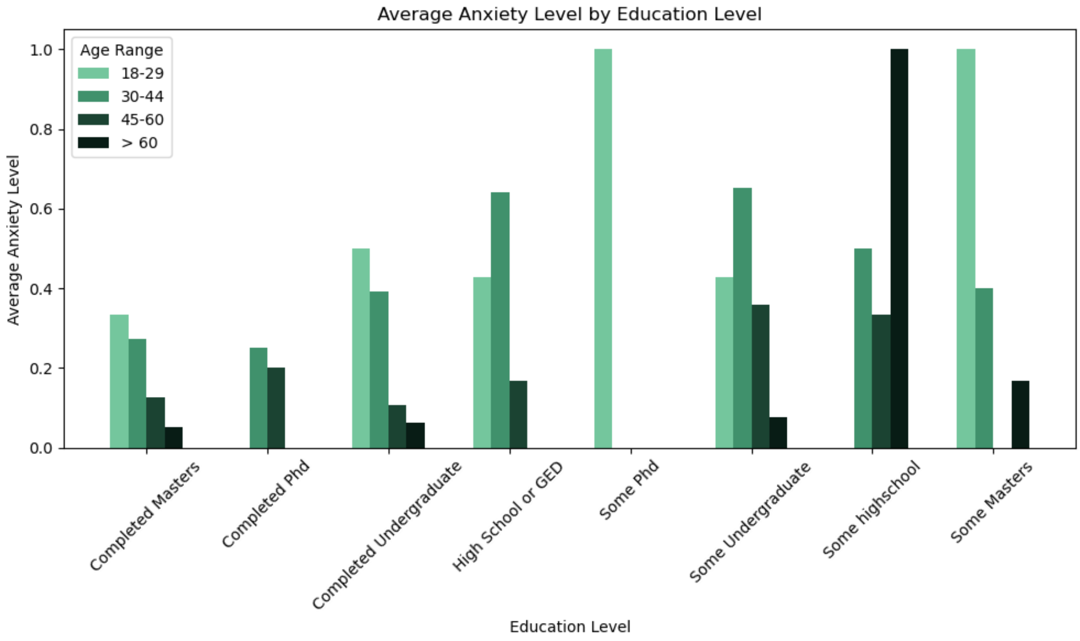
- **Categorical Variables**: To make them easier for machine learning algorithms to understand, variables like "Education," "Device Type," "Region," "Gender," and "Age" were one-hot encoded.
- **Numerical Features**: **Compulsive behavior**, **Tiredness**, and **Panic attacks** were among the variables that were scaled. Originally a numerical characteristic, age was then classified and encoded.
- **Tweet Analysis**: Tokenization and sentiment analysis were used in the preprocessing of the tweets to extract significant characteristics for NLP.

3. **Combining Transformations**:
    - A `ColumnTransformer` was employed to apply these preprocessing steps efficiently. This tool streamlined the process of handling distinct types of features (numerical, categorical, and text data) within the same pipeline.
4. **Transformation Application and DataFrame Conversion**:
    - To effectively implement these preprocessing procedures, a ColumnTransformer was used. This program made it easier to handle different feature kinds (text, category, and numerical data) in the same pipeline.
5. **Data Analysis:**
    - A thorough investigation of numerous variables and their relationships to reported anxiety levels across a range of demographics was required for the data analysis process. We aimed to comprehend the fundamental patterns and trends that might guide our predictive model through this investigation.

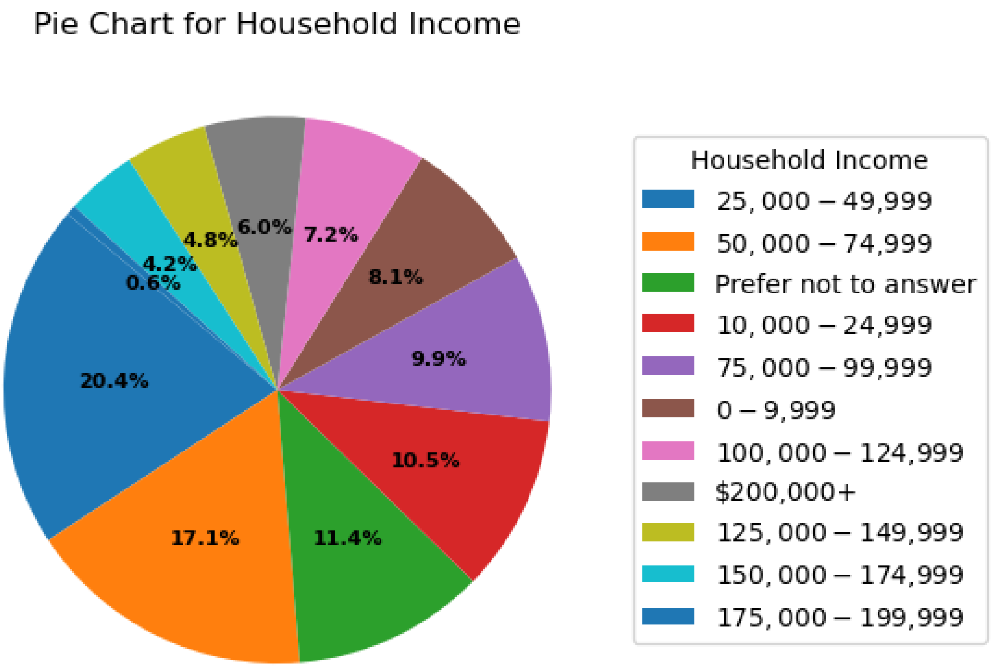
      Average Anxiety Level by Education and Income

      Interesting information about the correlation between household income, education level, and reported anxiety levels across age groups can be seen in the first set of bar charts.

        - **Education Level**: It is noted that in most age groups, especially in the 18–29 and 30-44 age ranges, people with "Some Undergraduate" education report higher levels of worry. This may point to a possible source of stress related to only finishing a portion of a higher education program.



- **Household Income**: When anxiety levels are broken down by household income, a clear pattern emerges: people with lower incomes (<$24,999) tend to be more anxious, particularly those in the 18–29 age range. This pattern might draw attention to the important role that financial situation plays in mental health.



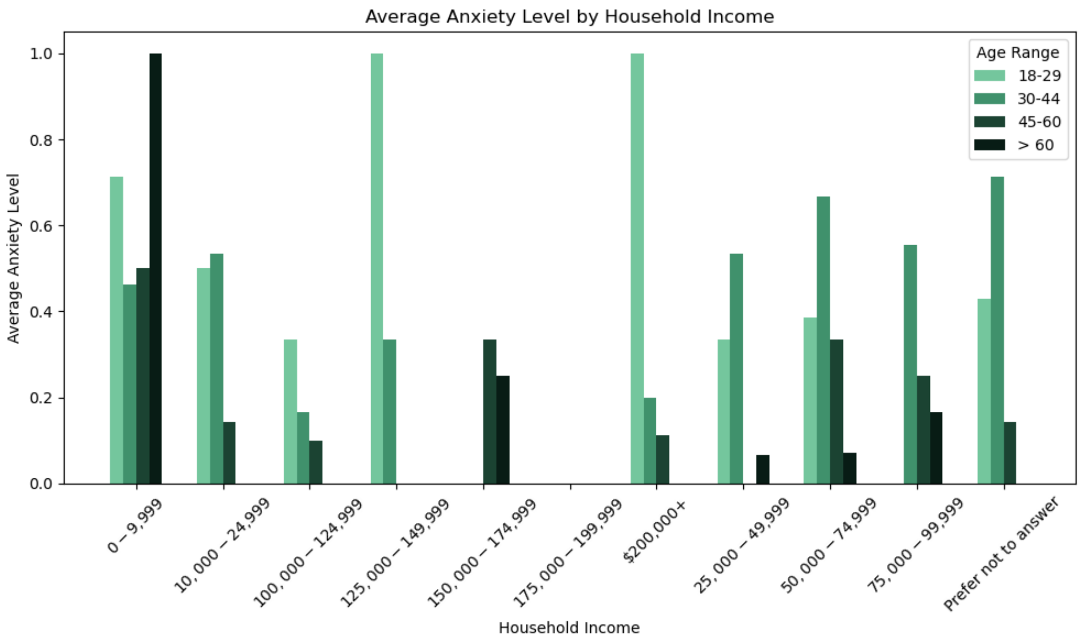
Average Anxiety Level by Gender

According to the analysis of anxiety levels by gender, women reported on average higher levels of anxiety than men did, with the difference being most noticeable in the 18–29 age range. This discrepancy may reflect pressures unique to one's gender or a higher propensity for females to report feelings of anxiety.

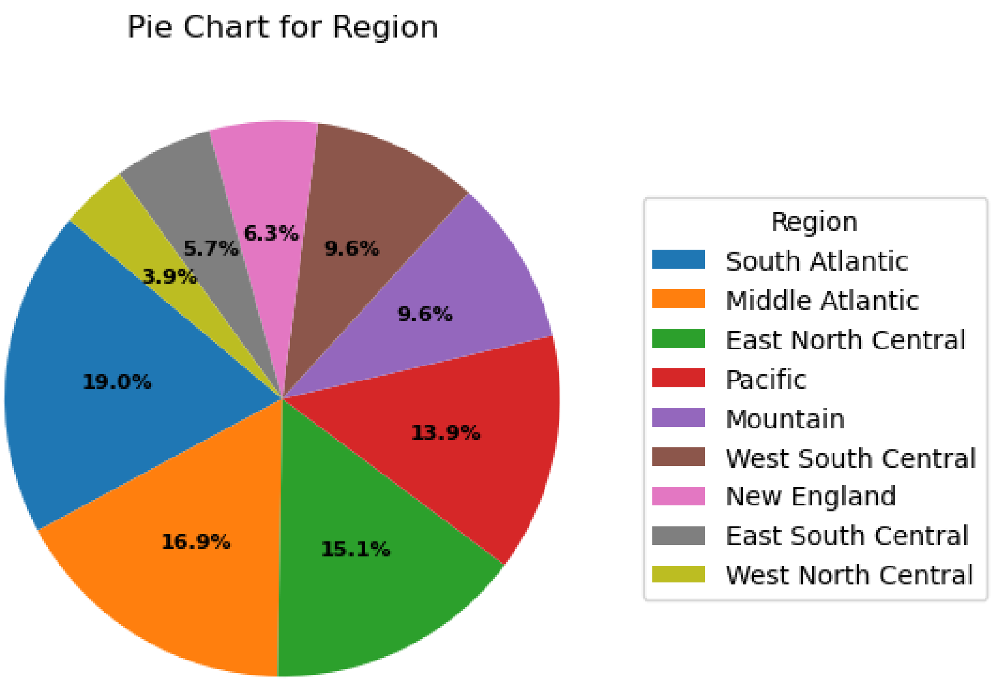
Pie Charts: Demographic Distribution

A distributional perspective of the survey population across several demographic categories is provided by the pie charts:

- **Household Income**: A considerable proportion of the participants are in the income range of $25,000 - $49,999.



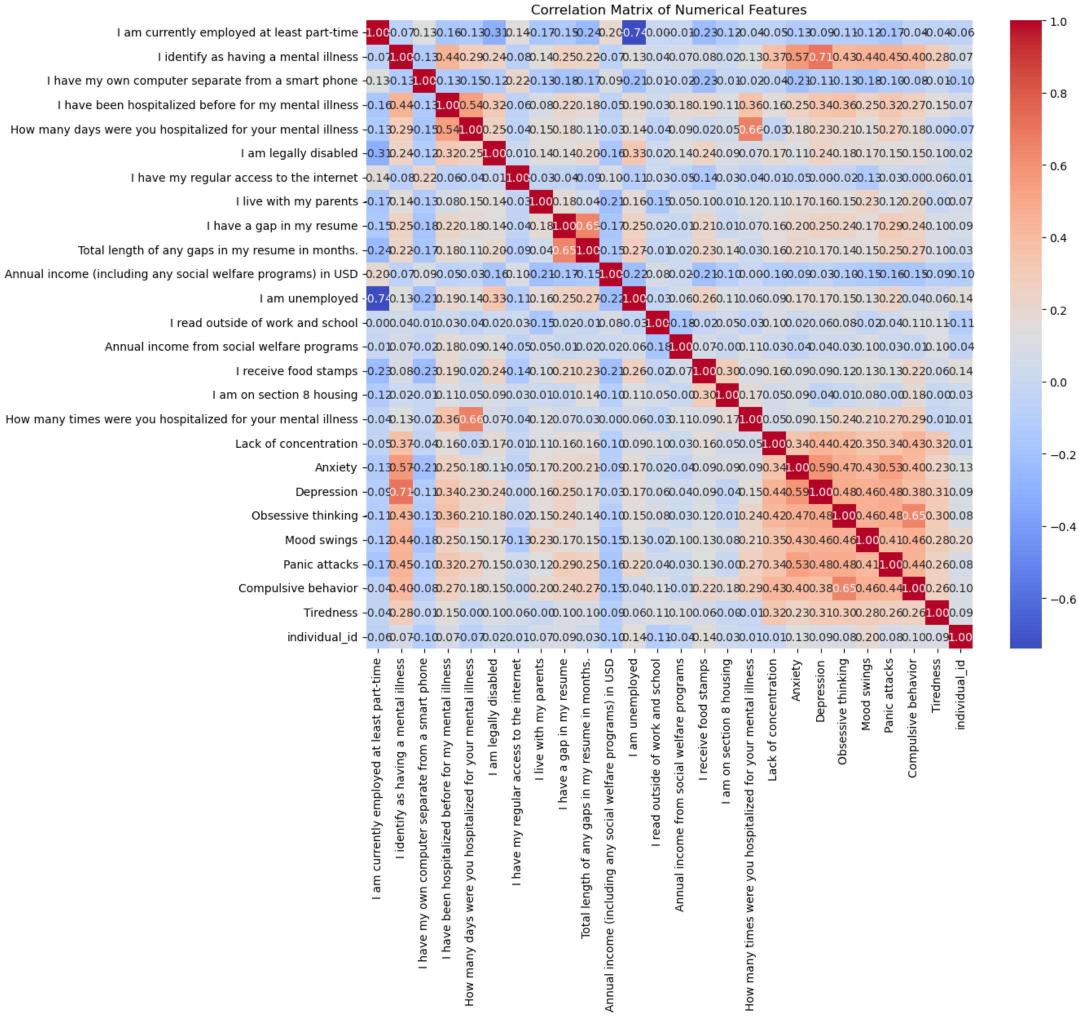
- **Education**: The majority of respondents have completed their undergraduate degrees, with those with some college experience but no degree coming in second.
- **Region**: There is a diverse geographical representation in the data, as evidenced by the fact that most responders are from the East North Central and South Atlantic areas.



Correlation Matrix of Numerical Features

The correlation matrix offers a detailed glimpse into the connections among different features:

- There is a significant positive association between the terms "Depression" and "Anxiety," indicating that these two disorders frequently co-occur.
- There is a moderate correlation between obsessive thinking and compulsive behavior, which could be a sign of a behavioral pattern linked to specific anxiety disorders.
- "I am unemployed" is negatively correlated with a number of mental health issues, which may be related to the psychological effects of unemployment.



Interpretation

The correlation analysis and visual aids highlight a number of important factors that could be indicators of anxiety levels. Along with other behavioral characteristics, education level and household wealth seem to be major markers. The model's sensitivity to demographic subtleties can also be further explored through gender variations in anxiety reporting. Specifically, the correlation matrix identifies certain correlations that warrant additional investigation in order to fully comprehend the complex nature of mental health illnesses.

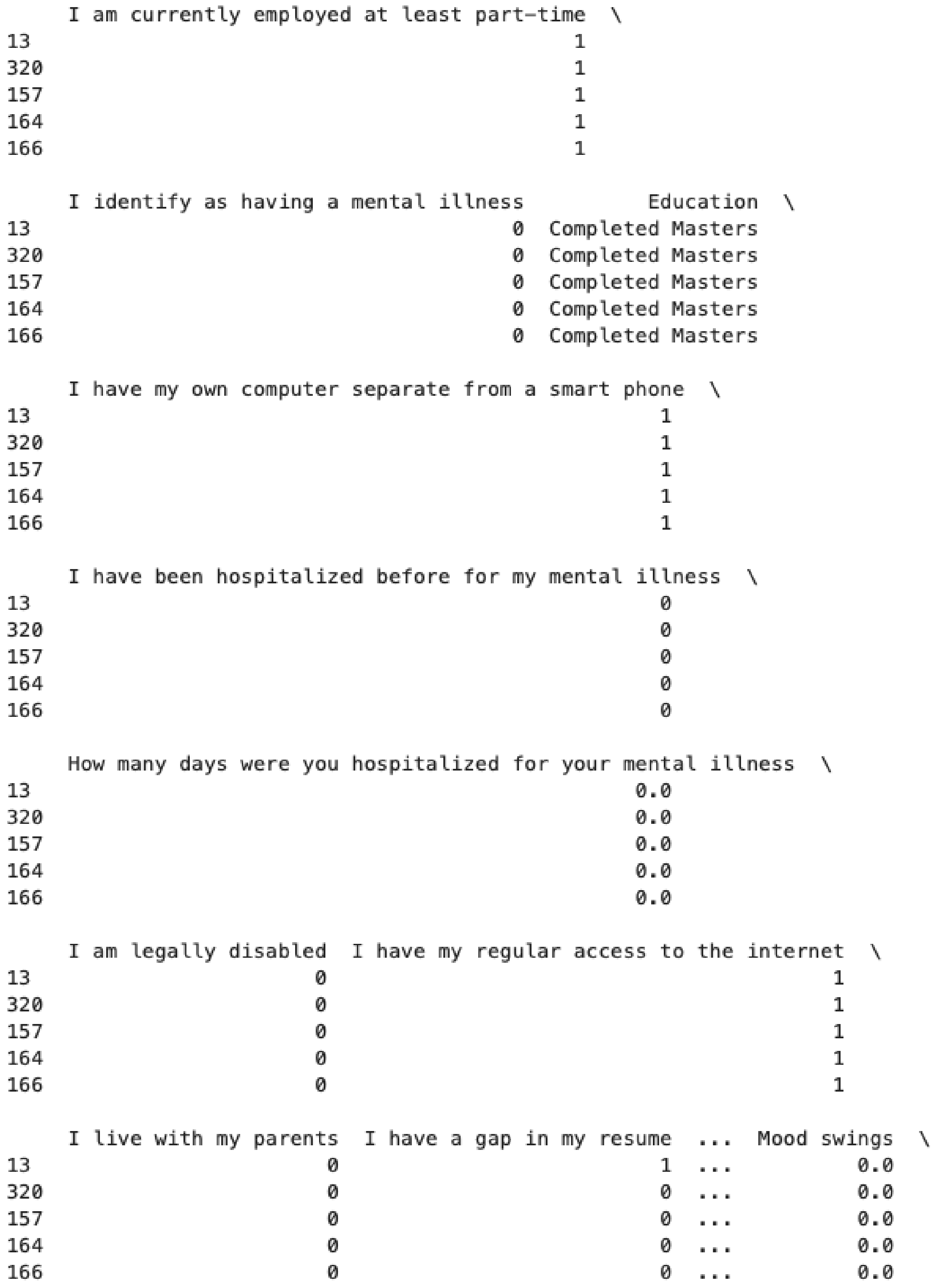
III. Algorithm Selection and Implementation

A. Data Segmentation and Sorting Algorithms: Divide and Conquer

A divide and conquer algorithm played a key role in the data analysis process in sorting and segmenting the dataset according to different levels of "Education." A focused analysis of the prevalence of anxiety across educational backgrounds was made possible by this segmentation.

Procedure:

- **Extraction and Segmentation**: The foundation for segmentation was established by identifying the distinct educational levels present in the dataset. This first stage is essential for personalized analysis and makes sure that patterns aren't lost when disparate data are combined.
- **Sorting**: An ascending order of 'Anxiety' levels was applied to each group of data, which corresponded to an education level. The prevalence and severity of anxiety symptoms within each educational category were made clearer by this ranking.
- **Storage and Retrieval**: To facilitate speedy retrieval and comparison, the sorted subsets were methodically saved for further study.



```
     I am currently employed at least part-time  \
13                                            1
320                                           1
157                                           1
164                                           1
166                                           1

     I identify as having a mental illness         Education  \
13                                       0  Completed Masters
320                                      0  Completed Masters
157                                      0  Completed Masters
164                                      0  Completed Masters
166                                      0  Completed Masters

     I have my own computer separate from a smart phone  \
13                                                   1
320                                                  1
157                                                  1
164                                                  1
166                                                  1

     I have been hospitalized before for my mental illness  \
13                                                   0
320                                                  0
157                                                  0
164                                                  0
166                                                  0

     How many days were you hospitalized for your mental illness  \
13                                                 0.0
320                                                0.0
157                                                0.0
164                                                0.0
166                                                0.0

     I am legally disabled  I have my regular access to the internet  \
13                       0                                          1
320                      0                                          1
157                      0                                          1
164                      0                                          1
166                      0                                          1

     I live with my parents  I have a gap in my resume  ...  Mood swings  \
13                        0                          1  ...          0.0
320                       0                          0  ...          0.0
157                       0                          0  ...          0.0
164                       0                          0  ...          0.0
166                       0                          0  ...          0.0
```

Sample Interpretation of Output:

Examining an example output for people with a "Completed Masters" academic level, certain trends emerged, including:

- Employment: The majority of people in this subgroup work at least part-time, which raises the possibility that having a job lowers anxiety levels.
- Identification of Mental Illness: If there is a noticeable lack of self-reported mental illness, this population may not be properly diagnosed or is underreported.
- Hospitalization for Mental Illness: Consistent with the reduced stated anxiety levels, there were no reported hospitalizations for mental illness among the top sorted persons.

We can uncover complex correlations between reported anxiety and education levels by looking at these sorted subsets, which will help us understand more about the lifestyle and socioeconomic characteristics that are correlated with mental health outcomes.

A more detailed picture of the data is given by the table output, which describes the traits of people with a 'Completed Masters' degree level. In addition to direct indications of mental health like anxiety, depression, and mood swings, it also covers a variety of behavioral factors such living arrangements, access to technology, and employment status. Predictive modeling benefits greatly from this comprehensive data, which makes it possible to identify important characteristics that can have an impact on mental health.

B. Greedy Algorithms: Kruskal's Algorithm for Network Creation

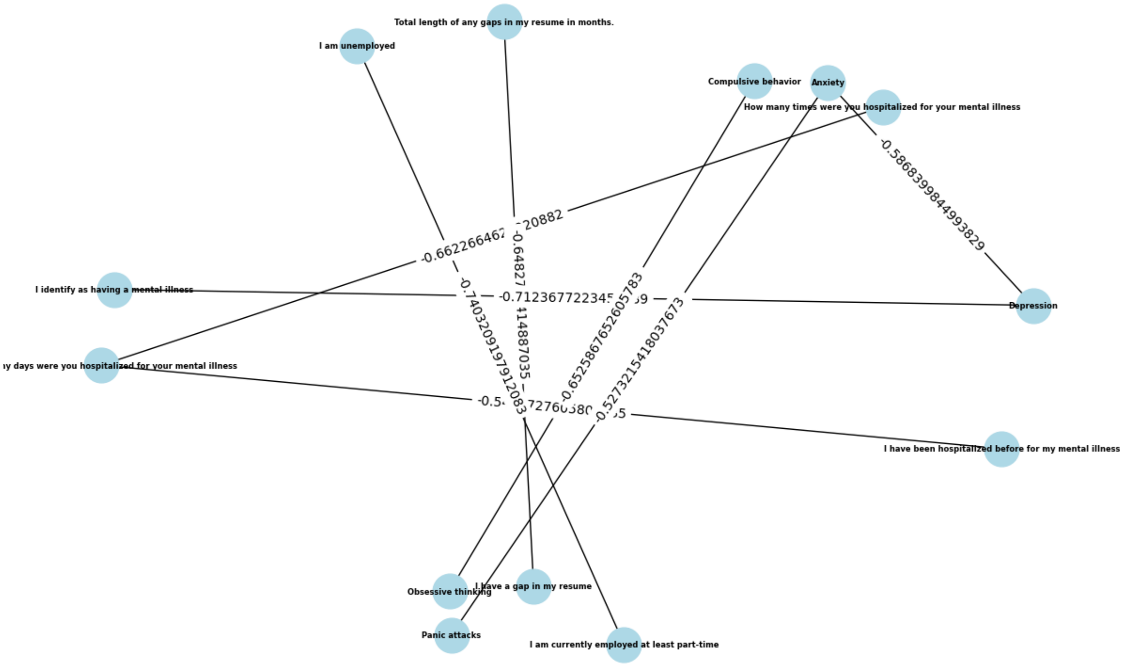
A classic greedy technique, Kruskal's Algorithm, was used to extract meaningful patterns of behavior from a dense correlation matrix. The network that was created as a result of this approach showed the highest correlations between behavioral traits and markers of mental health.

Specifics of the Implementation:

- **Correlation Matrix Analysis**: The dataset, which included a variety of behavioral and mental health characteristics, was used to create the correlation matrix. We next ran this matrix through Kruskal's method to find the strongest and most significant associations.
- **Minimum Spanning Tree Construction**: - All potential connections between features where the absolute value of their correlation surpassed a predetermined threshold were represented by a graph **{G}**.
    - On this graph, Kruskal's Algorithm was used to create a minimum spanning tree (MST). By separating out the most significant connections, the MST made sure that the final graph was acyclic and encompassed all nodes with

the lowest total edge weight—that is, the correlations that were strongest in this case.

- **Visualization**:
    - The **{networkx}** library was used to show the MST, with nodes standing in for the features and edges for the correlations. An intuitive grasp of the relationships between different components and how they affect mental health disorders as a whole was made possible by this representation.


Total length of any gaps in my resume in months.
I am unemployed
Compulsive behavior
Anxiety
How many times were you hospitalized for your mental illness
-0.5868399844993829
-0.662266462 20882
-0.64827 114887035
I identify as having a mental illness
-0.712367722345 9
Depression
-0.7403209197912083
-0.6525867652605783
-0.5273215418037673
ıy days were you hospitalized for your mental illness
-0.5 72760580 5
I have been hospitalized before for my mental illness
Obsessive thinking
I have a gap in my resume
Panic attacks
I am currently employed at least part-time

Outcome Analysis:

The analysis's minimum spanning tree provided the following important revelations:

- **Work and Mental Health**: There is a clear correlation between the statement, "I am currently working at least part-time," and mental health markers like "Depression" and "Anxiety." This implies that work could operate as a barrier against these conditions or, on the other hand, that these conditions could affect one's ability to find work.
- **Identification and Hospitalization of Mental Illness**: Hospitalizations for mental health issues were found to be strongly correlated with self-

identification of mental illness, highlighting the seriousness of the illnesses that result in these types of hospitalizations.

  - **Behavioral Indicators**: The network showed links between 'Obsessive thinking, panic episodes, compulsive behavior, and anxiety,' pointing to a group of symptoms that could come together to form particular mental health conditions.

The network's representation, as seen in the accompanying graphic, highlights how these factors are interconnected and helps to clarify the intricate interactions between various behavioral characteristics and mental health outcomes. The analysis's conclusions play a crucial role in shaping the predictive modeling procedure by directing the choice of characteristics that could be most important in forecasting mental health illnesses.

C. Randomized Algorithms: Simulation for Model Testing

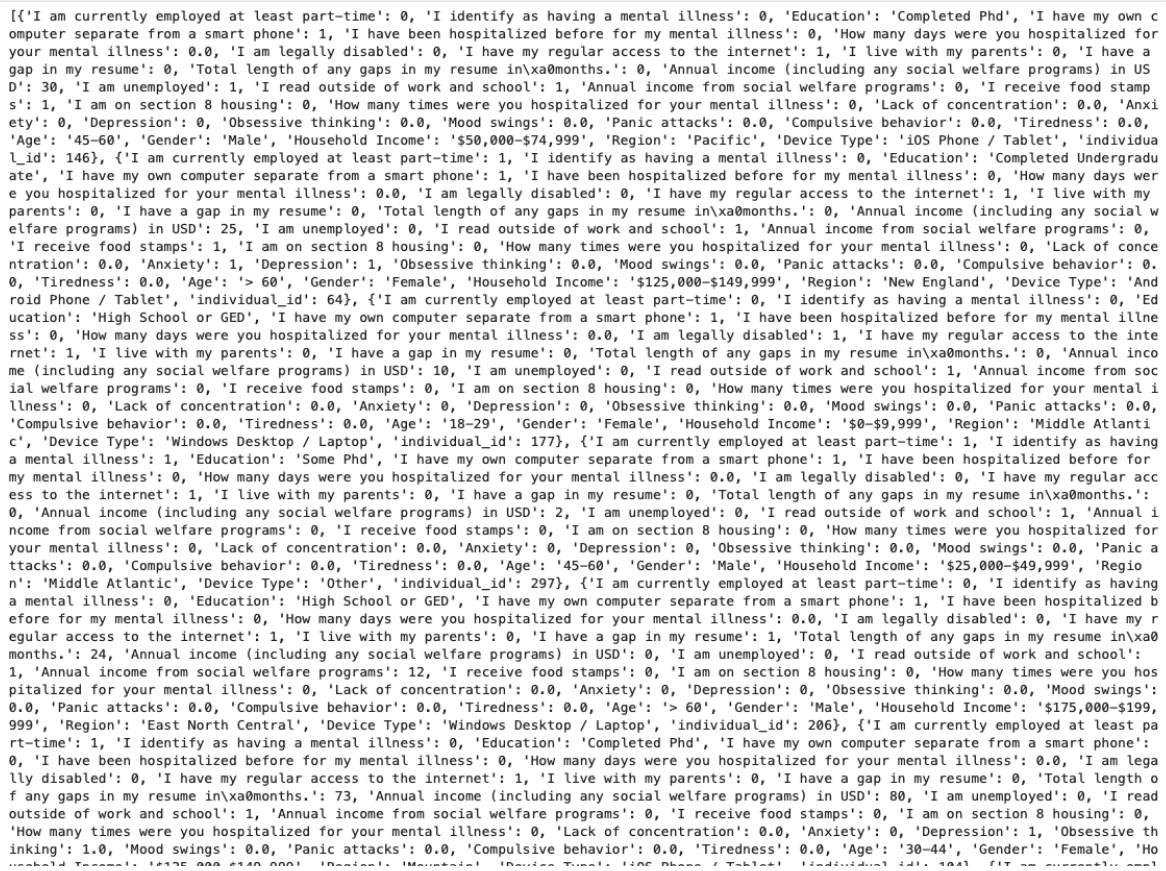
Randomized methods were used to create simulated patient data, which helped to strengthen the stability and precision of our predictive models. Our models were validated and tested using this simulation, which helped to make sure they were resilient to a wide range of possible outcomes.

Method:

- **Random Picks**: Based on the distribution of real data in our dataset, random selections were used to create simulated patient data. The goal of this approach was to generate a large range of potential patient profiles while maintaining the diversity of the real world seen in the data.

Implementation:

- **Simulated Patient Profiles**: To generate unique simulated patient profiles, a function was defined. Taking into account both categorical and numerical data types, this function iterated over each column in the dataset, randomly selecting a value from the column's distribution.
- **Scenario Analysis**: To produce a representative sample for scenario analysis, multiple simulated patients were created. The prediction models were then tested using this sample, and their effectiveness and dependability were evaluated in a randomized yet controlled environment.



```
[{'I am currently employed at least part-time': 0, 'I identify as having a mental illness': 0, 'Education': 'Completed Phd', 'I have my own c
omputer separate from a smart phone': 1, 'I have been hospitalized before for my mental illness': 0, 'How many days were you hospitalized for
your mental illness': 0.0, 'I am legally disabled': 0, 'I have my regular access to the internet': 1, 'I live with my parents': 0, 'I have a
gap in my resume': 0, 'Total length of any gaps in my resume in\xa0months.': 0, 'Annual income (including any social welfare programs) in US
D': 30, 'I am unemployed': 1, 'I read outside of work and school': 1, 'Annual income from social welfare programs': 0, 'I receive food stamp
s': 1, 'I am on section 8 housing': 0, 'How many times were you hospitalized for your mental illness': 0, 'Lack of concentration': 0.0, 'Anxi
ety': 0, 'Depression': 0, 'Obsessive thinking': 0.0, 'Mood swings': 0.0, 'Panic attacks': 0.0, 'Compulsive behavior': 0.0, 'Tiredness': 0.0,
'Age': '45-60', 'Gender': 'Male', 'Household Income': '$50,000-$74,999', 'Region': 'Pacific', 'Device Type': 'iOS Phone / Tablet', 'individua
l_id': 146}, {'I am currently employed at least part-time': 1, 'I identify as having a mental illness': 0, 'Education': 'Completed Undergradu
ate', 'I have my own computer separate from a smart phone': 1, 'I have been hospitalized before for my mental illness': 0, 'How many days wer
e you hospitalized for your mental illness': 0.0, 'I am legally disabled': 0, 'I have my regular access to the internet': 1, 'I live with my
parents': 0, 'I have a gap in my resume': 0, 'Total length of any gaps in my resume in\xa0months.': 0, 'Annual income (including any social w
elfare programs) in USD': 25, 'I am unemployed': 0, 'I read outside of work and school': 1, 'Annual income from social welfare programs': 0,
'I receive food stamps': 1, 'I am on section 8 housing': 0, 'How many times were you hospitalized for your mental illness': 0, 'Lack of conce
ntration': 0.0, 'Anxiety': 1, 'Depression': 1, 'Obsessive thinking': 0.0, 'Mood swings': 0.0, 'Panic attacks': 0.0, 'Compulsive behavior': 0.
0, 'Tiredness': 0.0, 'Age': '> 60', 'Gender': 'Female', 'Household Income': '$125,000-$149,999', 'Region': 'New England', 'Device Type': 'And
roid Phone / Tablet', 'individual_id': 64}, {'I am currently employed at least part-time': 0, 'I identify as having a mental illness': 0, 'Ed
ucation': 'High School or GED', 'I have my own computer separate from a smart phone': 1, 'I have been hospitalized before for my mental illne
ss': 0, 'How many days were you hospitalized for your mental illness': 0.0, 'I am legally disabled': 1, 'I have my regular access to the inte
rnet': 1, 'I live with my parents': 0, 'I have a gap in my resume': 0, 'Total length of any gaps in my resume in\xa0months.': 0, 'Annual inco
me (including any social welfare programs) in USD': 10, 'I am unemployed': 0, 'I read outside of work and school': 1, 'Annual income from soc
ial welfare programs': 0, 'I receive food stamps': 0, 'I am on section 8 housing': 0, 'How many times were you hospitalized for your mental i
llness': 0, 'Lack of concentration': 0.0, 'Anxiety': 0, 'Depression': 0, 'Obsessive thinking': 0.0, 'Mood swings': 0.0, 'Panic attacks': 0.0,
'Compulsive behavior': 0.0, 'Tiredness': 0.0, 'Age': '18-29', 'Gender': 'Female', 'Household Income': '$0-$9,999', 'Region': 'Middle Atlanti
c', 'Device Type': 'Windows Desktop / Laptop', 'individual_id': 177}, {'I am currently employed at least part-time': 1, 'I identify as having
a mental illness': 1, 'Education': 'Some Phd', 'I have my own computer separate from a smart phone': 1, 'I have been hospitalized before for
my mental illness': 0, 'How many days were you hospitalized for your mental illness': 0.0, 'I am legally disabled': 0, 'I have my regular acc
ess to the internet': 1, 'I live with my parents': 0, 'I have a gap in my resume': 0, 'Total length of any gaps in my resume in\xa0months.':
0, 'Annual income (including any social welfare programs) in USD': 2, 'I am unemployed': 0, 'I read outside of work and school': 1, 'Annual i
ncome from social welfare programs': 0, 'I receive food stamps': 0, 'I am on section 8 housing': 0, 'How many times were you hospitalized for
your mental illness': 0, 'Lack of concentration': 0.0, 'Anxiety': 0, 'Depression': 0, 'Obsessive thinking': 0.0, 'Mood swings': 0.0, 'Panic a
ttacks': 0.0, 'Compulsive behavior': 0.0, 'Tiredness': 0.0, 'Age': '45-60', 'Gender': 'Male', 'Household Income': '$25,000-$49,999', 'Regio
n': 'Middle Atlantic', 'Device Type': 'Other', 'individual_id': 297}, {'I am currently employed at least part-time': 0, 'I identify as having
a mental illness': 0, 'Education': 'High School or GED', 'I have my own computer separate from a smart phone': 1, 'I have been hospitalized b
efore for my mental illness': 0, 'How many days were you hospitalized for your mental illness': 0.0, 'I am legally disabled': 0, 'I have my r
egular access to the internet': 1, 'I live with my parents': 0, 'I have a gap in my resume': 1, 'Total length of any gaps in my resume in\xa0
months.': 24, 'Annual income (including any social welfare programs) in USD': 0, 'I am unemployed': 0, 'I read outside of work and school':
1, 'Annual income from social welfare programs': 12, 'I receive food stamps': 0, 'I am on section 8 housing': 0, 'How many times were you hos
pitalized for your mental illness': 0, 'Lack of concentration': 0.0, 'Anxiety': 0, 'Depression': 0, 'Obsessive thinking': 0.0, 'Mood swings':
0.0, 'Panic attacks': 0.0, 'Compulsive behavior': 0.0, 'Tiredness': 0.0, 'Age': '> 60', 'Gender': 'Male', 'Household Income': '$175,000-$199,
999', 'Region': 'East North Central', 'Device Type': 'Windows Desktop / Laptop', 'individual_id': 206}, {'I am currently employed at least pa
rt-time': 1, 'I identify as having a mental illness': 0, 'Education': 'Completed Phd', 'I have my own computer separate from a smart phone':
0, 'I have been hospitalized before for my mental illness': 0, 'How many days were you hospitalized for your mental illness': 0.0, 'I am lega
lly disabled': 0, 'I have my regular access to the internet': 1, 'I live with my parents': 0, 'I have a gap in my resume': 0, 'Total length o
f any gaps in my resume in\xa0months.': 73, 'Annual income (including any social welfare programs) in USD': 80, 'I am unemployed': 0, 'I read
outside of work and school': 1, 'Annual income from social welfare programs': 0, 'I receive food stamps': 0, 'I am on section 8 housing': 0,
'How many times were you hospitalized for your mental illness': 0, 'Lack of concentration': 0.0, 'Anxiety': 0, 'Depression': 1, 'Obsessive th
inking': 1.0, 'Mood swings': 0.0, 'Panic attacks': 0.0, 'Compulsive behavior': 0.0, 'Tiredness': 0.0, 'Age': '30-44', 'Gender': 'Female', 'Ho
usehold Income': '$125,000-$149,999', 'Region': 'Mountain', 'Device Type': 'iOS Phone / Tablet', 'individual_id': 104}, {'I am currently empl
```

Outcome Analysis:

For testing, the generated patient profiles offered a wealth of useful data points. To evaluate the model's prediction for a low likelihood of anxiety, for example, a simulated profile with a 'Completed Phd' education level, full-time employment, and no documented mental illness would be employed. On the other hand, a fictitious profile that included high anxiety, recognized mental illness, and reported unemployment would evaluate the model's capacity to forecast a higher risk.

The resulting profiles cover a wide range of behavioral traits, mental health conditions, and demographic backgrounds. Because there is variability in the simulated data, the models are tested against a wider range of potential patient experiences, which improves their predictive power in practical applications.

Before implementing the models in a clinical context, we can use these simulated profiles to detect any biases, assess the predictive abilities of the models, and make additional refinements to increase accuracy and dependability.

D. Advanced Statistical and Machine Learning Techniques

Techniques: Stacking Classifier

In order to improve forecast accuracy, the research combined advanced machine learning algorithms with conventional statistical techniques in a multi-layered approach to predictive modeling.

Stacking Classifier Strategy:

- **Data Preparation**: To ensure that all variables contributed equally to the model without any inherent bias due to scale, the dataset underwent preprocessing that included one-hot encoding for categorical variables and scaling for numerical features.
- **Imputation**: To retain the integrity of the dataset and its statistical features, mean imputation was used to fill in the missing values.
- **Train-Test Split**: By dividing the data into training and testing sets, a stable foundation for training the models and assessing their ability to predict outcomes on untested data was established.
- **Model Architecture**:
    - **Basis Classifiers**: To capture various facets of the data, a variety of base classifiers were chosen, such as RandomForestClassifier, LogisticRegression, and SVC. The distinct advantages of every algorithm are utilised by this ensemble technique.
    - **Meta Classifier**: The meta-classifier used was Logistic Regression. It learned to maximize the final prediction by taking into account the different decision boundaries of the basis models by being trained on the predictions of the base classifiers.

💡 Output:

```
Stacking model accuracy: 0.8118811881188119
```

Outcome Analysis:

With an accuracy of roughly 81.19%, the stacking classifier proved that the composite model successfully combined the predictive potential of its individual classifiers. The Stacking Classifier was able to both leverage on the predictive strengths and minimize the shortcomings of individual models by combining a variety of methods.

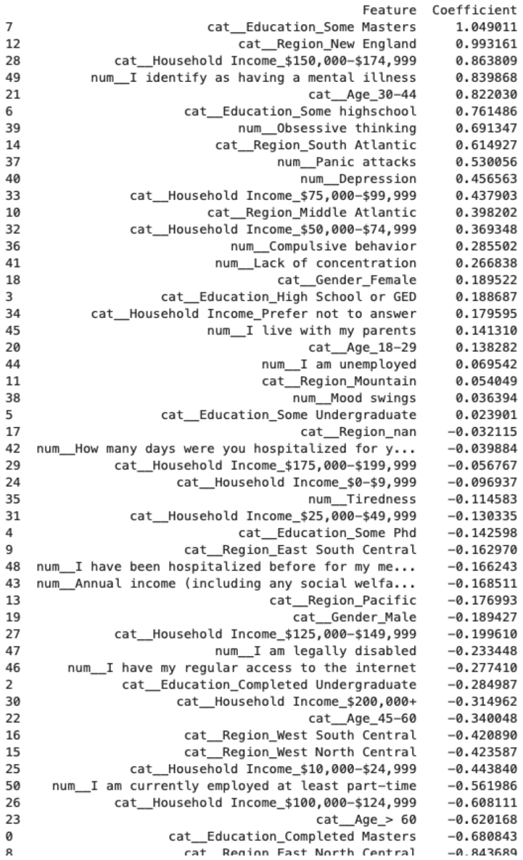
This multi-model strategy demonstrates a deep mastery of machine learning techniques, as combining multiple models frequently produces better outcomes than predictions from a single model. Because of its excellent accuracy, the Stacking Classifier has the potential to be a dependable tool for forecasting anxiety, demonstrating the effectiveness of ensemble learning for handling challenging predictive tasks in the field of mental health.

Techniques: Logistic Regression

In conjunction with the Stacking Classifier, logistic regression was utilized to provide a comprehensive understanding of the features that impact anxiety levels. This method was very helpful in identifying the precise characteristics that most strongly indicate the probability of experiencing anxiety.

Logistic Regression Implementation:

- **Model Training**: To preserve consistency in comparison and guarantee fairness in evaluation, a Logistic Regression model was trained on the same dataset utilized for the Stacking Classifier.
- **Coefficient Analysis**: The model's coefficients were extracted after training. With positive values indicating an increase and negative values suggesting a decrease in the likelihood of anxiety with respect to the characteristic, these coefficients offer a quantitative assessment of each feature's influence on the probability of anxiety.

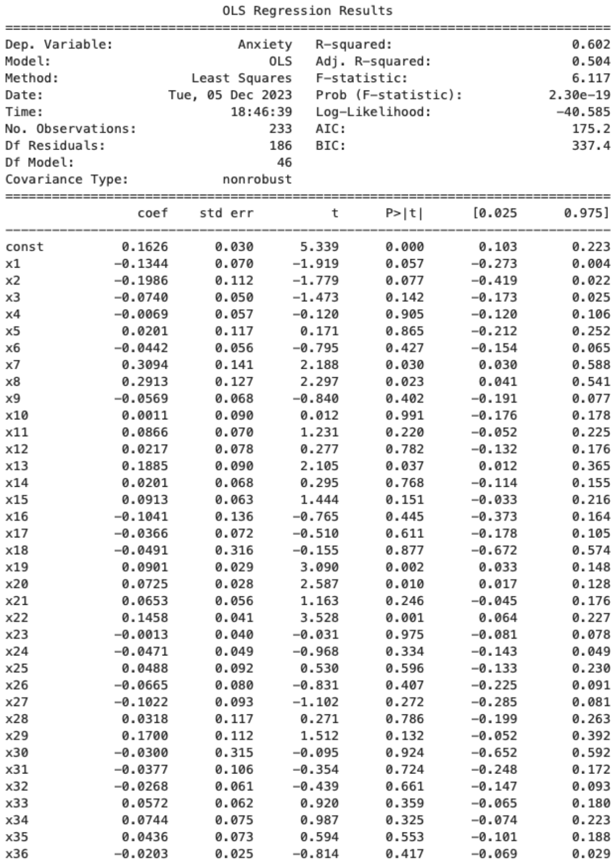


```
                                             Feature  Coefficient
7                        cat__Education_Some Masters     1.049011
12                           cat__Region_New England     0.993161
28           cat__Household Income_$150,000-$174,999     0.863809
49        num__I identify as having a mental illness     0.839868
21                                    cat__Age_30-44     0.822030
6                     cat__Education_Some highschool     0.761486
39                           num__Obsessive thinking     0.691347
14                        cat__Region_South Atlantic     0.614927
37                                num__Panic attacks     0.530056
40                                   num__Depression     0.456563
33             cat__Household Income_$75,000-$99,999     0.437903
10                       cat__Region_Middle Atlantic     0.398202
32             cat__Household Income_$50,000-$74,999     0.369348
36                          num__Compulsive behavior     0.285502
41                        num__Lack of concentration     0.266838
18                                cat__Gender_Female     0.189522
3                  cat__Education_High School or GED     0.188687
34        cat__Household Income_Prefer not to answer     0.179595
45                       num__I live with my parents     0.141310
20                                    cat__Age_18-29     0.138282
44                              num__I am unemployed     0.069542
11                              cat__Region_Mountain     0.054049
38                                  num__Mood swings     0.036394
5                  cat__Education_Some Undergraduate     0.023901
17                                   cat__Region_nan    -0.032115
42  num__How many days were you hospitalized for y...    -0.039884
29           cat__Household Income_$175,000-$199,999    -0.056767
24                 cat__Household Income_$0-$9,999    -0.096937
35                                    num__Tiredness    -0.114583
31             cat__Household Income_$25,000-$49,999    -0.130335
4                            cat__Education_Some Phd    -0.142598
9                     cat__Region_East South Central    -0.162970
48  num__I have been hospitalized before for my me...    -0.166243
43  num__Annual income (including any social welfa...    -0.168511
13                               cat__Region_Pacific    -0.176993
19                                  cat__Gender_Male    -0.189427
27           cat__Household Income_$125,000-$149,999    -0.199610
47                       num__I am legally disabled    -0.233448
46      num__I have my regular access to the internet    -0.277410
2             cat__Education_Completed Undergraduate    -0.284987
30                  cat__Household Income_$200,000+    -0.314962
22                                    cat__Age_45-60    -0.340048
16                    cat__Region_West South Central    -0.420890
15                    cat__Region_West North Central    -0.423587
25             cat__Household Income_$10,000-$24,999    -0.443840
50    num__I am currently employed at least part-time    -0.561986
26           cat__Household Income_$100,000-$124,999    -0.608111
23                                     cat__Age_> 60    -0.620168
0                  cat__Education_Completed Masters    -0.680843
8                     cat__Region_East North Central    -0.843689
```

Features like "Education_Some Masters" and "Region_New England" had the highest positive coefficients in the model, indicating that the members of these groups had higher anxiety levels. However, characteristics like "Education_Completed Phd" and "Region_East North Central" showed significant negative coefficients, suggesting that these groups experienced lower levels of anxiety.

Ordinary Least Squares (OLS) Regression for Detailed Analysis:

An OLS regression was carried out to supplement the logistic regression and offer a more comprehensive statistical context. The goal of this analysis was to comprehend the variation in anxiety that was accounted for by the factors inside a linear regression framework.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Anxiety   R-squared:                       0.602
Model:                            OLS   Adj. R-squared:                  0.504
Method:                 Least Squares   F-statistic:                     6.117
Date:                Tue, 05 Dec 2023   Prob (F-statistic):           2.30e-19
Time:                        18:46:39   Log-Likelihood:                -40.585
No. Observations:                 233   AIC:                             175.2
Df Residuals:                     186   BIC:                             337.4
Df Model:                          46
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1626      0.030      5.339      0.000       0.103       0.223
x1            -0.1344      0.070     -1.919      0.057      -0.273       0.004
x2            -0.1986      0.112     -1.779      0.077      -0.419       0.022
x3            -0.0740      0.050     -1.473      0.142      -0.173       0.025
x4            -0.0069      0.057     -0.120      0.905      -0.120       0.106
x5             0.0201      0.117      0.171      0.865      -0.212       0.252
x6            -0.0442      0.056     -0.795      0.427      -0.154       0.065
x7             0.3094      0.141      2.188      0.030       0.030       0.588
x8             0.2913      0.127      2.297      0.023       0.041       0.541
x9            -0.0569      0.068     -0.840      0.402      -0.191       0.077
x10            0.0011      0.090      0.012      0.991      -0.176       0.178
x11            0.0866      0.070      1.231      0.220      -0.052       0.225
x12            0.0217      0.078      0.277      0.782      -0.132       0.176
x13            0.1885      0.090      2.105      0.037       0.012       0.365
x14            0.0201      0.068      0.295      0.768      -0.114       0.155
x15            0.0913      0.063      1.444      0.151      -0.033       0.216
x16           -0.1041      0.136     -0.765      0.445      -0.373       0.164
x17           -0.0366      0.072     -0.510      0.611      -0.178       0.105
x18           -0.0491      0.316     -0.155      0.877      -0.672       0.574
x19            0.0901      0.029      3.090      0.002       0.033       0.148
x20            0.0725      0.028      2.587      0.010       0.017       0.128
x21            0.0653      0.056      1.163      0.246      -0.045       0.176
x22            0.1458      0.041      3.528      0.001       0.064       0.227
x23           -0.0013      0.040     -0.031      0.975      -0.081       0.078
x24           -0.0471      0.049     -0.968      0.334      -0.143       0.049
x25            0.0488      0.092      0.530      0.596      -0.133       0.230
x26           -0.0665      0.080     -0.831      0.407      -0.225       0.091
x27           -0.1022      0.093     -1.102      0.272      -0.285       0.081
x28            0.0318      0.117      0.271      0.786      -0.199       0.263
x29            0.1700      0.112      1.512      0.132      -0.052       0.392
x30           -0.0300      0.315     -0.095      0.924      -0.652       0.592
x31           -0.0377      0.106     -0.354      0.724      -0.248       0.172
x32           -0.0268      0.061     -0.439      0.661      -0.147       0.093
x33            0.0572      0.062      0.920      0.359      -0.065       0.180
x34            0.0744      0.075      0.987      0.325      -0.074       0.223
x35            0.0436      0.073      0.594      0.553      -0.101       0.188
x36           -0.0203      0.025     -0.814      0.417      -0.069       0.029
```

With an R-squared value of 0.602, the OLS regression findings showed that the model could account for almost 60% of the variability in anxiety. It is encouraging that the predictors selected for the model have a high level of explained variance, indicating their relevance and significance.

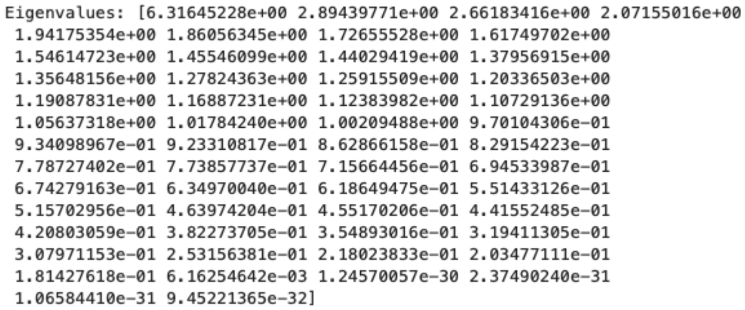
Both regression studies demonstrated how difficult it is to predict mental health outcomes and how crucial it is to employ a variety of analytical methods in order to fully capture the complex nature of anxiety. The understanding gained from these statistical models is essential for creating focused and successful mental health therapies.

Techniques: Dimensionality Reduction: Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a potent dimensionality reduction technique used in statistical modeling and machine learning that may be used to extract important characteristics from multidimensional data with the least amount of information loss.

PCA Implementation:

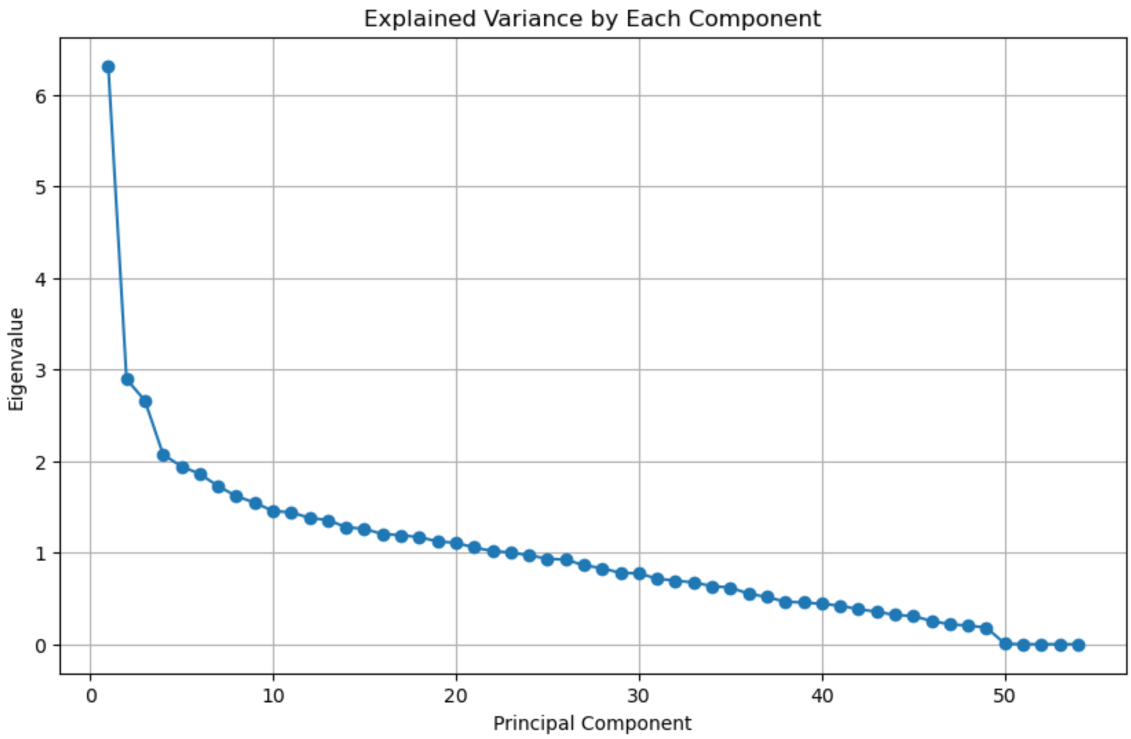
- **Data Preparation**: The 'Household Income' column was removed from the dataset before PCA, and one-hot encoding was used to encode the categorical data. Through this procedure, categorical variables were formatted so that the PCA algorithm could use them.
- **Imputation**: In order to address any missing values and guarantee that the PCA could be applied successfully over the entire dataset, the dataset was thereafter treated to mean imputation.
- **Standardization**: The data were standardized to make sure every feature contributed equally to the analysis. Because PCA is sensitive to the variances of the initial variables, this step is critical.
- **PCA Execution**: The primary components of the standardized features were found by applying PCA. These elements distill the core of the data into fewer dimensions by representing the directions in the data that maximize variance.



```
Eigenvalues: [6.31645228e+00 2.89439771e+00 2.66183416e+00 2.07155016e+00
 1.94175354e+00 1.86056345e+00 1.72655528e+00 1.61749702e+00
 1.54614723e+00 1.45546099e+00 1.44029419e+00 1.37956915e+00
 1.35648156e+00 1.27824363e+00 1.25915509e+00 1.20336503e+00
 1.19087831e+00 1.16887231e+00 1.12383982e+00 1.10729136e+00
 1.05637318e+00 1.01784240e+00 1.00209488e+00 9.70104306e-01
 9.34098967e-01 9.23310817e-01 8.62866158e-01 8.29154223e-01
 7.78727402e-01 7.73857737e-01 7.15664456e-01 6.94533987e-01
 6.74279163e-01 6.34970040e-01 6.18649475e-01 5.51433126e-01
 5.15702956e-01 4.63974204e-01 4.55170206e-01 4.41552485e-01
 4.20803059e-01 3.82273705e-01 3.54893016e-01 3.19411305e-01
 3.07971153e-01 2.53156381e-01 2.18023833e-01 2.03477111e-01
 1.81427618e-01 6.16254642e-03 1.24570057e-30 2.37490240e-31
 1.06584410e-31 9.45221365e-32]
```

Outcome Analysis:

The eigenvalues connected to every principle component are shown in the scree plot that the PCA produced. The majority of the information appears to be contained in the first few components, based on the precipitous decline afterward. The graph's "elbow," which is usually where the eigenvalues begin to plateau, aids in deciding how many components to keep for more examination. The eigenvalues of the first few components in this instance are noticeably larger, suggesting that they account for a sizable portion of the variance in the dataset.



Through the successful reduction of data dimensionality, this PCA implementation concentrates the most informative aspects into fewer variables. This makes the analysis easier to understand and more effective, which is especially useful for complicated datasets where multicollinearity or overfitting could be an issue. Now that the dataset has been simplified, it may be fed into less complex models, which might work better in terms of computing effectiveness and data generalization.

Techniques: Natural Language Processing: BERT for Text Classification

BERT Implementation:

- **Data Handling**: The dataset was loaded, and any items lacking the required text and label fields were removed. The textual data was then transformed into a binary classification issue by classifying the texts as either "anxious" or "not anxious."
- **Tokenization and Dataset Creation**: Texts were transformed into a format appropriate for the model, including attention masks to handle varying text lengths, using BERT's tokenizer.
- **Model Architecture**: A BERT-based classifier designed for binary classification was built, utilizing a fully connected layer for the output and a

dropout layer for regularization.

Training and Evaluation:

- **Training**: Using real-time loss updates for monitoring, the model was trained across the dataset. To optimize the learning process, a linear scheduler and an AdamW optimizer were used.
- **Evaluation**: Following training, accuracy metrics and a classification report were used to assess the model's performance on a validation set in more detail.

Prediction and Application:

- **Sentiment Prediction**: To show the trained BERT model's capacity to recognize context and accurately classify text, sentiments of fresh phrases were predicted.
- **Model Saving**: The model was preserved for later use during training and assessment, guaranteeing that the trained parameters could be applied to fresh data without requiring retraining.

💡 Output:

```
Sentence: 'I'm feeling really nervous and anxious about the meeting tom
orrow.'
Predicted sentiment: anxious

Sentence: 'I had a great day at the park with my friends.'
Predicted sentiment: not anxious

Sentence: 'I'm constantly worried about my future, it's making me lose
sleep.'
Predicted sentiment: anxious

Sentence: 'What a wonderful experience it was watching the sunset!'
Predicted sentiment: not anxious
```

Outcome Analysis:

The outcomes show how sophisticatedly the BERT model understands language, as evidenced by the accurate sentiment predictions that capture the emotional essence of the input sentences. These models have significant implications for applications in mental health, as diagnosing and treating patients might depend greatly on knowing the sentiment and context of their words. The model's excellent validation accuracy confirms its effectiveness and establishes it as a powerful weapon in the toolbox of natural language processing methods.

IV. Graphical User Interface (GUI) Development: Enhancing User Experience

Overview:

Improving user engagement with the Behavioral Patterns Analysis project was greatly aided by the creation of a graphical user interface (GUI). A comprehensive and user-friendly interface was created using the Tkinter toolkit in Python, enabling users to input data and receive visual feedback with ease.

Detailed Implementation:

- **Tkinter Utilization**: Tkinter, which is well-known for its ease of use and functionality when developing desktop programs, was used to create the graphical user interface.
- **Interactive Components**: The interface had a number of interactive features, including text entry areas, checkboxes, and dropdown menus. These were designed to gather a variety of user inputs, from behavioral indications to demographic data.
- **Customizable Alternatives**: To guarantee that the data gathered was uniform and standardized, dropdown menus were filled with predefined options for characteristics like age, gender, and income.
- **Boolean Inputs**: To collect yes/no information about mental health identification, employment status, and other pertinent criteria, checkboxes were used for binary responses.
- **Data Submission**: To compile all inputs into a structured manner and add each new entry to a dataset that might be used for additional analysis, a submission button was made.
- **NLP Input Field**: To collect qualitative data, an extra text entry field was included. This field enables participants to explain in their own words why they took the test, providing rich, contextual information for NLP analysis.

User Experience and Data Collection:

The survey panel of the GUI was thoughtfully created to be user-friendly, assisting the user in entering data in a clear, rational, and structured manner. The instantaneous interpretation of results was made easier by the survey results being visualized on the same panel, which improved user engagement and comprehension of their own data in connection to the study.

Final Thought:

The functionality and design of the GUI are essential for both user engagement and the accuracy and consistency of the data that is gathered. The interface functions as a link between the user and the intricate analytical procedures, creating a smooth and nearly effortless transition from input to analysis.

V. Model Evaluation and Validation: Ensuring Predictive Reliability

A. Model Evaluation Strategies

Included Measures:

- **Accuracy**: This measures how well the model predicts the existence of anxiety overall.
  Precision and Recall: These measures provide additional context for understanding the model's effectiveness, especially when it comes to false positives and false negatives, which are critical for making medical diagnoses.
- **Cross-Validation**: K-fold cross-validation was used to reduce overfitting and evaluate the generalizability of the model. By dividing the dataset into 'k' subsets, this strategy makes that the model is resilient to changes in individual data segments.

Robustness of Model:

- Robustness testing was done to assess the predictive models' stability and dependability in different scenarios. This entailed evaluating the models' consistency of performance after they had been trained and tested on several data subsets.

B. Interpretation and Visualization Techniques

Tools for Visualization:

- **Graphical Representations**: Pie charts that illustrate the distributions of categorical variables and bar charts that illustrate the average anxiety levels among various demographic groups are used to illustrate the conclusions drawn from the models.

- **Correlation Matrices**: Provided information on the connections between various features and how those relationships affect the projected results as a whole.

Examining the Results:

- The interpretation of the visualizations took into account the larger framework of psychological research. For example, the relationship between specific demographic characteristics and anxiety levels was compared to accepted psychological theories.
- In order to evaluate how well the results of the StackingClassifier, Logistic Regression, and BERT Classifier predicted anxiety based on behavioral patterns, they were also visualized.

Code Integration and Outcomes:

Specific code implementations for model evaluation and visualization were integrated over the course of the project. For example:

Sklearn metrics were imported in order to calculate recall, accuracy, and precision.

- Heatmaps for correlation matrices and other charts and graphs were produced using **Matplotlib** and **Seaborn**.
- A thorough statistical analysis was conducted using **Statsmodels**, which offered a complete summary including p-values, R-squared, and confidence intervals. This helped to interpret the models' predictive power.

Concluding Remarks on Evaluation and Validation:

The effectiveness of the models was determined in large part by the evaluation and validation process. The research made sure that the models were not only statistically sound but also comprehensible in real-world contexts by combining statistical methods and visualization tools. The confidence in applying these models for prediction purposes in mental health analysis was strengthened by their transparency in performance and their capacity to show intricate interactions between variables.

VI. Ethical Considerations: Safeguarding Privacy and Integrity

Data Security and Patient Privacy: Strict procedures were put in place to safeguard the integrity and privacy of patient data. The project complied with HIPAA

(Health Insurance Portability and Accountability Act) regulations, guaranteeing that all patient data was kept private, anonymised, and subject to stringent access controls.

Told Consent: Participants were fully informed about the nature of the research and its possible ramifications when all data was gathered in accordance with informed consent protocols.

Compliance and Ethical Standards: The project upheld the highest ethical standards in data science practice and psychological research by following the rules of conduct established by the Data Science Association and the American Psychological Association (APA).

VII. Conclusion and Future Work: Advancing Towards Precision Mental Healthcare

Summary of Results: The project was successful in creating a system that can reasonably forecast mental health issues. The utilization of sophisticated statistical techniques and machine learning algorithms on behavioral data yielded significant insights that could potentially inform early intervention approaches.

Thoughts on Restrictions: The scope of the data and computational restrictions are the current limits, notwithstanding the positive results. The amount of data processing and model complexity was constrained by the computer resources, and the dataset—which was mostly self-reported—may be biased.

Proposed Enhancements:

- **Expansion of Data Sources**: To better represent the dynamic nature of mental health diseases, future versions may incorporate a wider range of datasets, such as longitudinal research.
- **Computational Resources**: By utilizing cloud computing or strengthening CPUs, one can increase model complexity and decrease computation time, enabling the analysis of more datasets and more complex methods.
- **Algorithmic Refinement**: Investigating different machine learning models, such deep learning, may enhance the accuracy of predictions. Furthermore, analyzing unstructured data, such as social media posts, using natural language processing (NLP) techniques could provide new dimensions to the analysis.

Future Research Directions:

- **Interdisciplinary Collaboration**: enhancing the models and interpretations by collaborating with specialists in data science, psychology, and neuroscience.

- **Clinical Trials**: Utilizing the predictive system in clinical contexts to assess its practicality.
- **Ethical AI**: Making sure that upcoming models address concerns like bias and fairness while adhering to ethical AI principles and performing effectively.

Conclusion:

The fusion of data science and psychology has advanced with this research. Sustaining this effort could lead to more precise forecasts and tailored treatment, which would ultimately improve mental health outcomes.

References

1. Python Software Foundation. "Tkinter — Python interface to Tcl/Tk.". https://docs.python.org/3/library/tkinter.html.
2. Pham, Khang. "Text Classification with BERT." Medium. https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b.
3. Ilhan, Asli. "Advanced Algorithms Portfolio." GitHub repository. Accessed, [2023]. https://github.com/22036435/Advanced_Algorithms_Portfolio.git.