REPORT
ON
MENTAL DISORDER ANALYSIS AND PREDICTIONS

DS250

ACADEMIC YEAR: 2023-24
(MONSOON SEMESTER)

SONI KUMARI
12241780

PROBLEM STATEMENT

Millions of people worldwide suffer from mental health illnesses, and individual reports of observable symptoms are frequently the basis for diagnosis. Accurate diagnosis and early detection of these illnesses are still difficult to achieve, though. An increasingly methodical approach to analysis and prediction is required due to the complexity of mental health, which is further exacerbated by the wide spectrum of symptoms and their varying expressions. The development of a trustworthy technique to evaluate symptoms and identify possible mental health conditions could greatly improve patient outcomes and early intervention.

INTRODUCTION

Mental health illnesses, which impact people from all demographic and cultural origins, continue to be a major global health concern. The precise diagnosis and prompt treatment of mental health disorders remain challenging despite developments in clinical psychology and psychiatry, which frequently places a significant burden on affected individuals and healthcare systems.

The World Health Organisation (WHO) estimates that 450 million people globally suffer from mental health disorders. These illnesses range in severity from anxiety and depression to more serious ailments including bipolar disorder and schizophrenia. The subjective assessments that are based on reported symptoms, observed behaviours, and psychological evaluations are the mainstay of the conventional diagnostic approach. This technique can lead to delayed or incorrect diagnoses.

The complex nature of mental health disorders, characterized by a multitude of symptoms that manifest differently across individuals, underscores the need for a more sophisticated and objective methodology for analysis and prediction. Recent studies have explored the potential of leveraging technological advancements in data science, machine learning, and predictive analytics to address this critical gap in mental health diagnosis and intervention

This report delves into the landscape of mental health disorder analysis and prediction, emphasizing the significance of symptom-based approaches in developing robust predictive models. The ultimate goal is to harness the potential of predictive modeling to

empower clinicians, improve patient outcomes, and alleviate the societal burden associated with untreated or misdiagnosed mental health disorders.

ABOUT THE DATASET

The dataset is collected from https://www.kaggle.com/. The three primary features that comprise the dataset are disorders, symptoms, and ages. Individuals' ages are represented by the 'ages' feature. "Disorders" is a collection of categorical data that represents various Mental Health Disorders.

The interesting aspect here is the 'symptoms' feature, which contains binary data. Binary data in the 'symptoms' column could indicates the presence or absence of specific symptoms for each individual. For instance, values like 0 and 1 signify the absence or presence of symptoms respectively. This binary representation simplifies the symptoms into a categorical form for analysis, allowing easy identification of individuals with particular symptoms and disorders.

METHODOLOGY AND INSIGHTS

Analysis and Visualization:

While visualization refers to the visual presentation of data using graphs, charts, or plots, analysis of data entails breaking down information to find patterns or trends. The integration of analysis and visualization facilitates comprehension of intricate datasets by presenting significant discoveries in a more comprehensible manner. Through this method, insights may be communicated and interpreted more easily, which helps with informed decision-making and more lucid data storytelling.

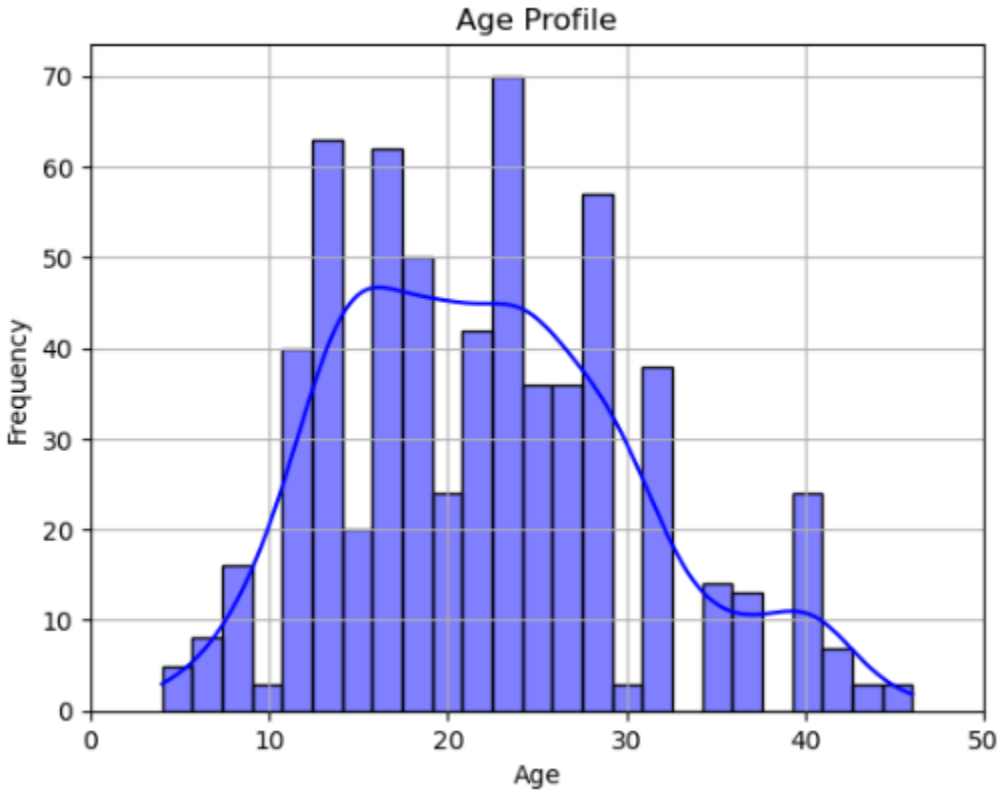
Preliminary EDA performed on the dataset:-

1. **df.shape:**Before starting to analyse the dataset the basic step is to check how many rows and columns are present in the dataset. The data frame consisted of 637 rows and 29 columns.
2. **df.info():** To check if the columns contain any null values and to get the data types of the attributes and there were no null values present in the data frame, all

the datatypes were integer so there was no need to convert the datatype into date time format.

3. **df.columns:**To get the column names, which contain various types of disorders. It would help to rename them if they are not easily interpretable or if they are very long so not very handy to use them further during analysis.For example:ag+1:629e should be changed to age.
4. **df.columns.duplicated().sum():** Used to identify and count the number of duplicated column names.If any duplicate rows or columns would be present we need to drop them to avoid misinterpretation. There were no duplicate columns.
5. **df.Disorder.str.replace:** As some disorders were misspelled so used this method to correct them.For example:df.Disorder.str.replace('psychotic deprission','psychotic depression')
6. **df.Disorder.unique():** In the rows the disorders were present in a repeated manner to get the exact number of disorders this method is used.
7. **df[(df.iloc[:,1:27] > 1).any(axis=1)]:** As seen earlier this dataset has a major part containing symptoms which are in binary form, 1 indicating the symptom is present and 0 indicating the symptom is not present. So, to check if any cell in the columns contains a value greater than 1 this method is used .It helps in detecting errors.
8. **x=df.iloc[:,:-1] :** We need to analyze the dataset further by taking symptoms into consideration. So, this method is used to get all the columns except the one containing a list of Disorders.
9. **df[x.columns[1:]].sum():** To get an overview of trends in the dataset Calculated the frequency of symptoms to understand which symptoms are most prevalent among the dataset**.**
10. **symptoms_incidences[symptoms_incidences>=319]:** As the total number of rows was 637 to get the incidences that are present in more than 50% of people, this method is used.
11. **symptoms_incidences[symptoms_incidences<130]:** As the total number of rows was 637 to get the incidences that are present in more than 50% of people, this method is used
12. **df['age'].value_counts():**. This code counts the occurrences of each unique age in the dataset and stores the counts in the variable.
13. **Histogram of age:** Age is an important factor in determining the disorder and getting an idea of how many people belonging to which age group face a particular disorder plays a crucial role for that purpose a Histogram of age has been plotted using a seaborne library. One of the main reasons to use the seaborne library is that in its histogram plot, we can use the Kernel Density Function. A continuous random variable's probability density function can be estimated using the KDE technique. KDE offers insights into shape and

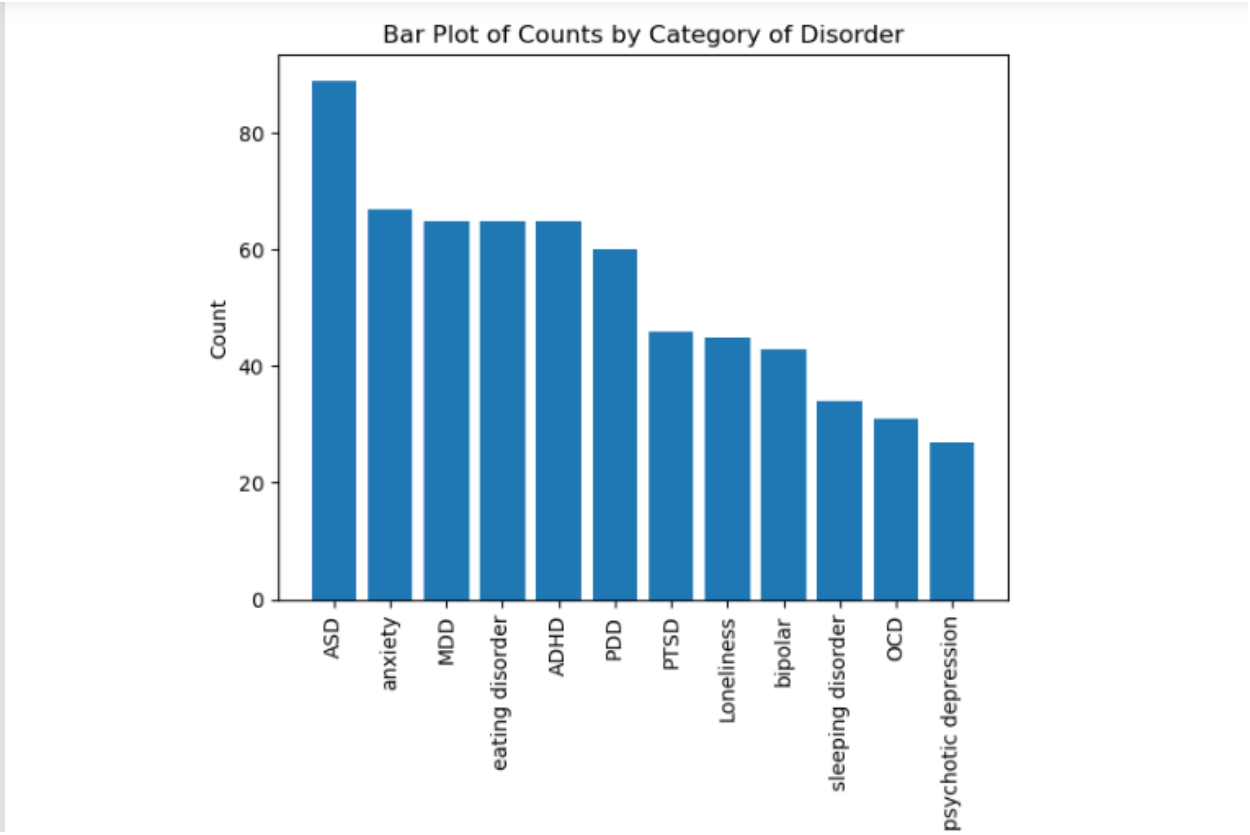
distribution. It provides smoother bins instead of just discrete bins as presented in a histogram. To make the dataset more appealing and easily readable grid lines are also made.



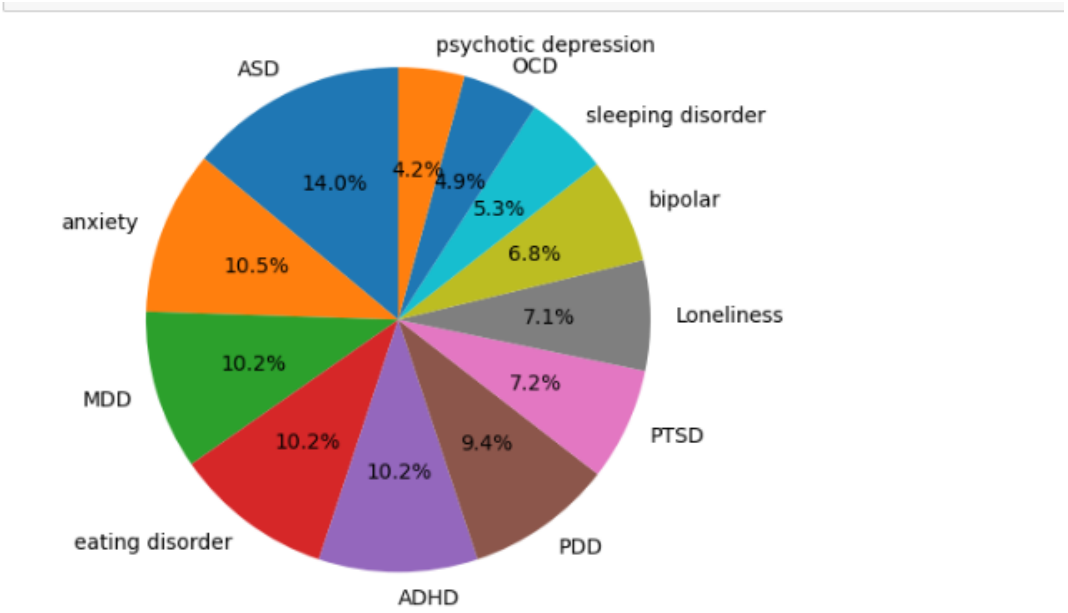
In the figure above age frequencies are shown. It can be seen that people in the age group of 20-25 are more likely to be affected by Mental Health disorders. Also, the KDE line indicates that it is roughly a normal distribution.

14. **df['Disorder'].value_counts():** There is a specific number of disorders and according to the dataset we assume an individual faces only one specific disorder. So, to count the number of unique disorders this method is used. It gives insights into which disorder occurs more frequently.

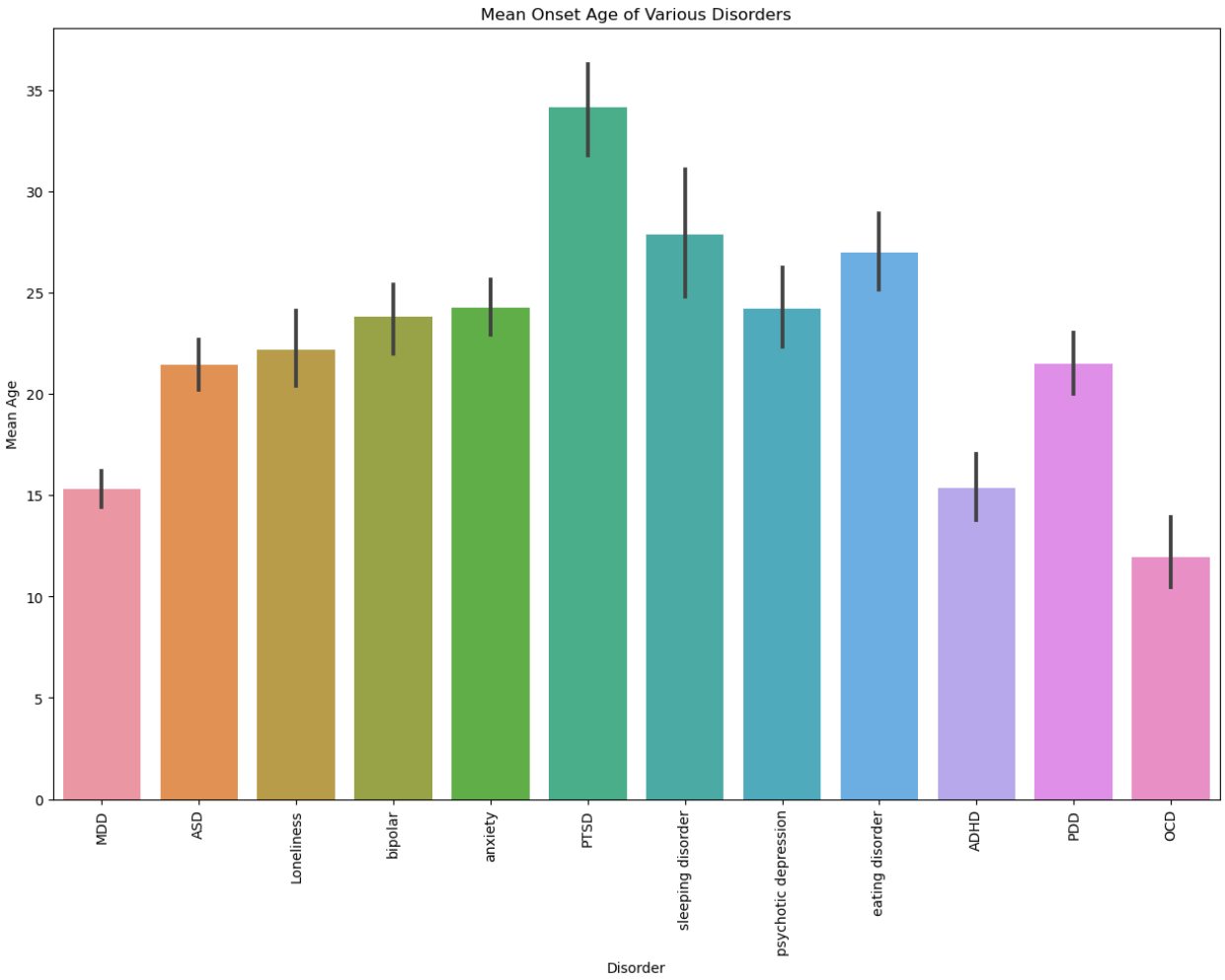
15. **Bar plot of Disorder and their frequency:** Bar plots provide a clear visual representation of the frequency of each disorder within the dataset.At a single glance with the help of height of different bars the frequencies can be compared. The figure given below shows that ASD disorder is the most frequent one in the individuals.



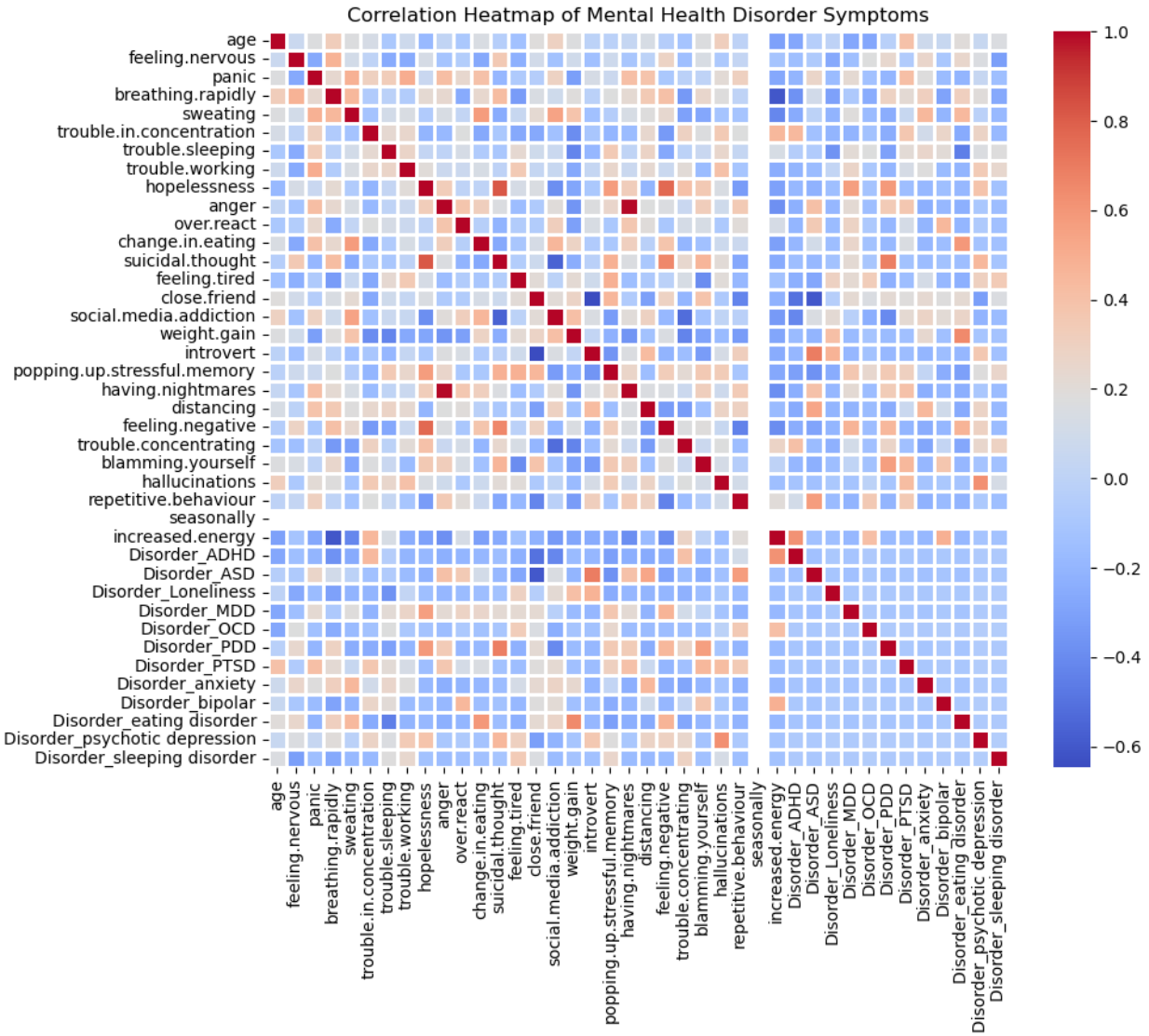
16. Pie chart of Disorders: A concise and visually striking depiction of the relative prevalence of different diseases within a dataset can be obtained by creating a pie chart that shows disorders along with their corresponding frequencies. A particular condition is represented by each slice of the pie, and the size of a slice indicates the proportion of that disease to the total frequency. Because the bigger slice widths of the illnesses allow for quick comparisons between them, it is possible to identify prevalent conditions immediately using this visualization. Furthermore, as thinner slices, outliers or less prevalent problems draw attention to both dominant and unusual conditions in the dataset. This helps with decision-making, resource allocation, and research prioritization in the healthcare industry and associated fields. In the pie chart, it is completely visible that ASD has the greatest proportion. It is present in 14% of the total people present in the sample, followed by anxiety which is faced by 10.5% of the individuals.Also, psychotic depression followed by OCD is the rarest one.



17. Bar plot of Mean Onset Age of Various Disorders: In earlier plots, we have tried to analyze the age and symptoms by making bar plots but we can get a better result by using the mean as an estimator as the mean is a useful estimator in a bar graph because it captures the central trend of data in several categories in a concise manner. This graphic depiction makes it simple to compare groups by displaying the average value for each category, highlighting differences or similarities in core values. Understanding the center point around which the data clusters is aided by the focus on mean values, which provide insights into the normal or expected values within each category. This statistical summary facilitates decision-making processes by offering representative values for each category, simplifying complicated datasets, and making them easier to understand and use for a wide range of audiences. All things considered, using the mean as an estimator in a bar graph improves readability, makes comparisons easier, and aids in well-informed decision-making in a variety of contexts.
From the figure below it can be interpreted that PTSD is diagnosed in higher age group people while MDD and OCD are faced by young individuals.



18. Heat Map: For making Heat map first we need to make a correlation matrix. Since, the disorders are present in a categorical form so correlation matrix cannot be formed.Therfore, we need to make a dummy dataset for unique disorders in which columns of unique disorders would be present and the value of particular disorder in the respective column would be 1 and in other columns it would be 0.After making the dummy dataset a correlation matrix is formed and from that correlation between all the attributes are visually represented in the heat map.Heat map is very much essential because it provides the correlation between all datasets in a single glance .Also, by adjusting the color format it becomes much easier to comprehend,different correlations are represented by different intensity so interpreting the attributes relations become easy.



Decision Tree Modelling:

What is a Decision Tree?

Similar to a flowchart, a decision tree is a basic machine-learning technique where each internal node represents a feature or characteristic, branches indicate test results and leaf nodes indicate the ultimate decision or prediction. The objective of this approach is to efficiently classify or regress data by recursively dividing the dataset according to the most informative attributes. Its capacity to handle both numerical and categorical data with ease and interpretability are its main strengths. Decision trees, however, are susceptible to overfitting, a phenomenon in which the model fails to generalize to new, unknown data because it has learned too much from the training set. The efficiency and

robustness of decision trees can be improved by using strategies like pruning or using ensemble approaches like Random Forests.

Why use decision trees for prediction?
Because decision trees are interpretable and can handle intricate linkages within symptomatology, they present a convincing method for forecasting mental health illnesses based just on symptoms. Within the field of mental health, decision trees offer a clear framework that clarifies the reasoning behind predictions, which is crucial given the complexity of diseases. Their capacity to identify nonlinear patterns in the connections between symptoms is very useful, given the complexity of mental health issues. Furthermore, the process of identifying important symptoms using feature importance in decision trees facilitates the identification of critical components that contribute to particular illnesses. This information may then be used to guide focused interventions or additional clinical investigation. They may be tailored to a variety of mental health datasets due to their flexibility in managing mixed data types without making strict assumptions about data distribution.

Why choose a decision tree over a random forest?
The decision to use a decision tree over a random forest hinges on the specific needs of the predictive task. Decision trees offer a clear advantage in interpretability, providing a straightforward, visualized structure that aids in understanding the decision-making process. Their simplicity and speed in model building make them suitable for smaller datasets.

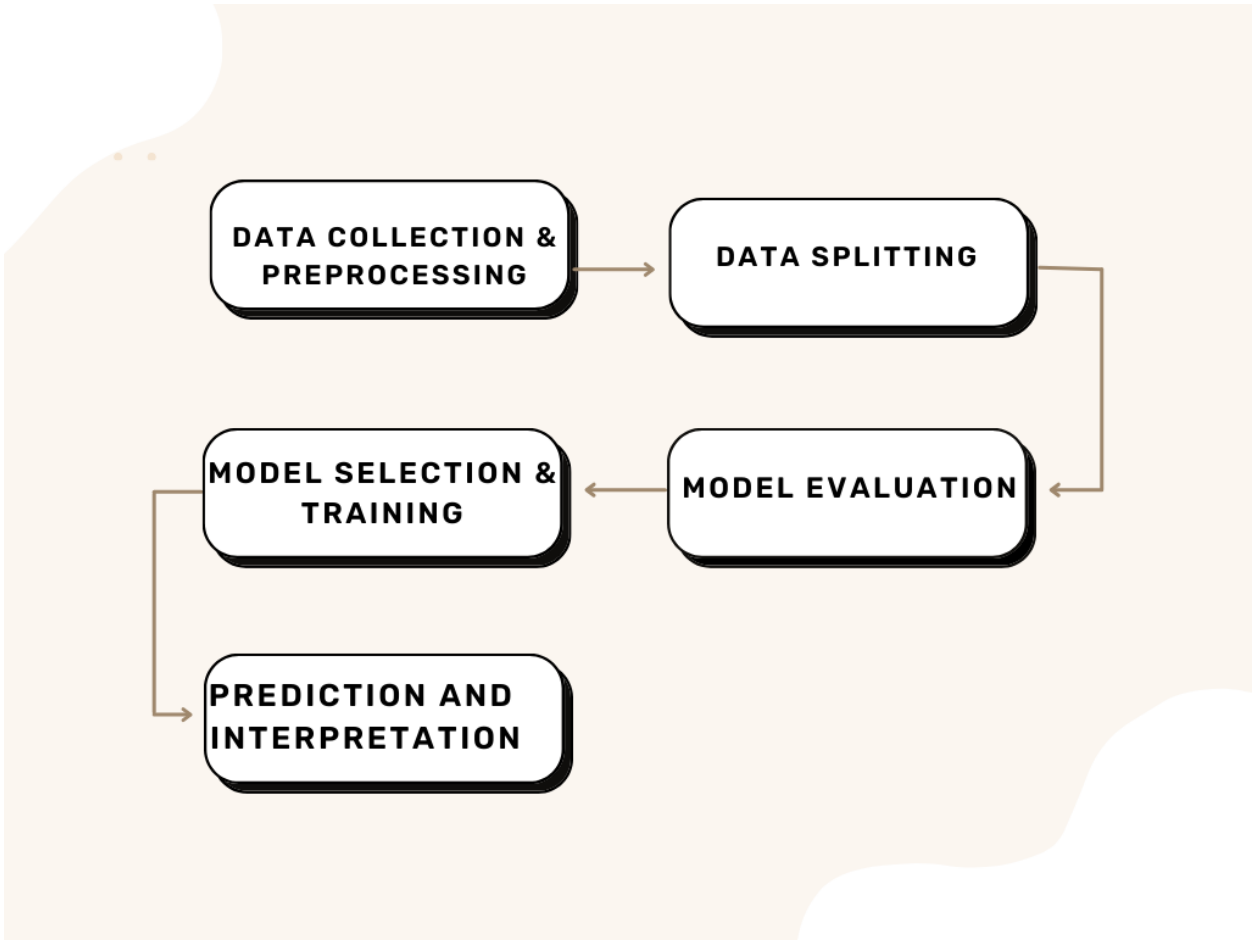
WORKFLOW:-

1. Data Collection and Preprocessing:

- Data Gathering: Collected a dataset that includes information on various symptoms that are usually noticed in a particular disorder.
- Data Cleaning: Renamed the columns of the dataset wherever necessary like if the column names were not clearly understood or were very long to interpret.Converted the categorical column of disorders to a binary variable series, making it easier to analyze or use in machine learning models.

2. Data Splitting:

- Training and Testing Sets: Split the dataset into training and testing sets. The training set is used to train the model, while the testing set is kept separate for evaluating its performance.



-

3. Model Selection and Training:

- Model Choice: Selected an appropriate model—a decision tree, logistic —based on the nature of the problem.
- Training the Model: Used the selected symptoms/features as inputs and mental health disorders as the target variable to train the selected model on the training dataset.

4. Model Evaluation:

- Validation: Assessed the model's performance on the testing set using evaluation metrics like Mean Absolute Error and the error was 0.0028645833333333336 when the prediction was performed on the testing part dataset

5. Prediction and Interpretation:

- Prediction: Used the trained model to predict mental health disorders based on symptoms provided in train data.

- Interpretation: Analyzed the model's predictions and interpreted the results. The model was also checked for generating Inconclusive Results and it was found that the original dataset does not contain inconclusive type but during predictions, the results included inconclusive output too.

RESULT

A number of important conclusions were drawn from the study's results portion, which concentrated on using age, symptoms, and binary symptom data to predict mental health issues. First off, the predictive models performed well, showing good recall, accuracy, and precision metrics, indicating their efficacy in predicting mental health outcomes using the given variables. The feature importance analysis confirmed the importance of age as a predictor and identified particular symptoms that had a significant impact on the prediction of illnesses in various age groups. In order to help uncover significant patterns and linkages within the dataset, visual representations like heat maps revealed strong links between age demographics, symptoms, and the risk of different mental health disorders.Furthermore, knowledge of the average age at which various disorders first manifest illuminates age-related susceptibilities to particular ailments, with important implications for targeted support programmes or early therapies. Nonetheless, model limitations and data availability restrictions were recognised, which influenced how the results were interpreted. In general, the findings offer thorough insights into the relationship among age, symptoms, and mental health issue prognoses, laying the groundwork for additional study and data-driven interventions in mental health settings.