MentalHealth Dataset

Data Preprocessing

Training Data

Model Selection

Random Forest

Naïve Bayes

Decision Tree

KNN

Logistic Regression

SVM

Model Evaluation

F1-Score

Recall

Accuracy

Precision

Confusion Matrix

**Data Pre-Processing-**

The data is transformed to a specific format that allows for easy analysis by various algorithms. This involves several stages, including:

**A.Importing necessary packages-**

It is important to import the necessary libraries and modules into the working environment before starting. These libraries and modules, such as numpy, warnings, matplotlib, and pandas, are essential for improving functionality and streamlining workflows. They facilitate the creation of various types of graphical representations, simplify complex numerical computations, enable the generation of data frames, provide a user-friendly interface between the user and the system, and allow for notification handling.

**B.Loading the dataset-**

Obtaining and analyzing data are essential components of any research study. Therefore, it is crucial to conduct a thorough examination of the collected dataset, considering its features and entries. In this study, information was obtained from 1259 patients, which included 23 attributes. These attributes are composed of various factors such as - Timestamp,self\_employed,family\_history,treatment,work\_interfere,no\_employees,remote\_work,tech\_company,benefits,care\_options,wellness\_program,seek\_help,wellness\_program,,

leave,anonymity,mental\_health\_consequence,phys\_health\_consequence,coworkers,supervisor,mental\_health\_interview,phys\_health\_interview,mental\_vs\_physical,obs\_consequence‘chance of disorder’ which is classified as either HIGH or LOW. The "CHANCE OF DISORDER" feature is an independent variable and plays a significant role in determining whether a patient has a any mental disorder or not. The other attributes are assigned a value of 0 ,1,2…, representing "no" and "yes….," respectively.

**C.Dealing with Missing data-**

To ensure the accuracy of the data, it is essential to address any inaccuracies or missing information. There are several methods to handle missing or incorrect data, including imputing values with mean, median, or mode, using specific algorithms to fill in the gaps, removing the rows with missing data, or assigning them to a separate category. However, in this study, there were no missing values found in the collected data.

**D.Encoding Data-**

In order to streamline calculations and enhance efficiency during the research process, the non-numeric values transformed into numerical values using integer encoding. This approach involves assigning numerical values to categorical variables, which makes computations and analysis easier.

**E.Feature Selection-**

In order to streamline future calculations and avoid unnecessary complexity, the study involves identifying and selecting only the crucial features. This is achieved using the correlation matrix technique, which helps to determine the features that have a substantial impact on the output feature, thereby enhancing accuracy and hastening the training process. The correlation matrix technique produces results that fall within the range of -1 to 1, excluding 0. This makes it easier to identify the most important features based on their correlation values with other features.

**F.Data Splitting-**

To move forward, the collected data is partitioned into two separate sets, called the training set and the test set, in the appropriate ratios. The primary purpose of the test set is to build and train the model using multiple algorithms, and then verify its accuracy using the unused test data set, thereby enhancing the effectiveness of the algorithms used.