Project Report

On

EMPLOYEE MENTAL HEALTH PREDICTIONS

Submitted in partial fulfillment for the award of Post Graduate Diploma in Big Data Analytics (**PG - DBDA**) From **CDAC – Noida**

(Uttar-Pradesh)

Guided by: -	**Submitted by:-**
Mrs.Priti Bhardwaj	**Pratik C Deshmukh (20320525038)**
	Ankit S Hiwarkar (20320525002)
	Shiv S Dabhade (20320525032)
	Sonali Sharma (20320525034)

CERTIFICATE

TO WHOME SO EVER IT MAY CONCERN

This to certify that

Pratik C Deshmukh (20320525038)

Ankit S Hiwarkar (20320525002)

Shiv S Dabhade (20320525032)

Sonali Sharma (20320525034)

Has successfully completed the Project on

EMPLOYEE MENTAL HEALTH PREDICTIONS

Under the guidance of

Mrs.Priti Bhardwaj

ACKNOWLEDGEMENT

This project **“Employee Mental Health Predictions”** was a great learning experience for us and we are submitting this work to CDAC Noida (Uttar-Pradesh).

We all are very glad to mention the name of **Mrs. Priti Bhardwaj** for her valuable guidance to work on this project. Her guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Ravi Payal** Director (DCAC-Noida), C-DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in Big Data Analytics (PG-DBDA) through C-DAC Noida, Uttar-Pradesh.

Our most heartfelt thanks goes to **Mrs. Siddhidatri Naik** (Course Co-ordinator,PG-DBDA) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC Noida, Uttar-Pradesh.

From:-

Pratik C Deshmukh (20320525038)

Ankit S Hiwarkar (20320525002)

Shiv S Dabhade (20320525032)

Sonali Sharma (20320525034)

TABLE OF CONTENTS

Tables of Figures

ABSTRACT

The capacity to work productively is a key component of health and emotional well-being. Common Mental Disorders are associated with reduced workplace productivity. It is anticipated that this impact is greatest in developing countries. Furthermore, workplace stress is associated with a significant adverse impact on emotional wellbeing and is linked with an increased risk of common mental disorders. This report will elaborate on the relationship between workplace environments and mental illness. In this report we have used different types of machine learning modules to see that if the employee have mental illness or he is totally fit and if there is problem regarding his mental health than what are the precaution that someone has to take care. This report help us to know whether the family history is important for the treatment of mental illness. Additionally, we provide concrete recommendations on the potential future research and development of applying machine learning in the mental health field

01. Introduction

Mental illness is a health problem that undoubtedly impacts emotions, reasoning, and social interaction of a person. These issues have shown that mental illness gives serious consequences across societies and demands new strategies for prevention and intervention. To accomplish these strategies, early detection of mental health is an essential procedure. Medical predictive analytics will reform the healthcare field broadly as discussed by Miner etc. Mental illness is usually diagnosed based on the individual self-report that requires questionnaires designed for the detection of the specific patterns of feeling or social interactions. With proper care and treatment, many individuals will hopefully be able to recover from mental illness or emotional disorder.

Machine learning is a technique that aims to construct systems that can improve through experience by using advanced statistical and probabilistic techniques. It is believed to be a significantly useful tool to help in predicting mental health. It is allowing many researchers to acquire important information from the data, provide personalized experiences, and develop automated intelligent systems widely used algorithms in the field of machine learning such as support vector machine, random forest, and artificial neural networks have been utilized to forecast and categorize the future events.

Supervised learning in machine learning is the most widely applied approach in many types of research, studies, and experiments, especially in predicting illness in the medical field. In supervised learning, the terms, attributes, and values should be reflected in all data instances More precisely, supervised learning is a classification technique using structured training data Meanwhile, unsupervised learning does not need supervision to predict main goal of unsupervised learning is handling data without supervision. It is very limited for the researchers to apply unsupervised learning methods in the clinical field.

1.1. Mental Health Problems

The World Health Organization (WHO) reports the region-wise status of different barriers in diagnosing mental health problems and encourages researchers to be equipped with the scientific Knowledge to address the issue of mental health. Now, there are various techniques to predict the state of mental health due to advancement of technology. Research in the field of mental health has increased recently and contributed to the information and publications about different features of mental health, which can be applied in a wide range of problems.

Many steps are involved in diagnosing mental health problems, and it is not a straightforward process that can be done quickly. Generally, the diagnosis will begin with a specific interview that is filled with questions about symptoms, medical history and physical examination. Besides that, psychological tests and assessment tools are also available and are used to diagnose a person for mental health problems. There are several types of research carried out to investigate and examine the movements of the face to identify certain mental disorders. The increase of research in the mental health field has led to the rise of information in the form of finding suitable solutions to reduce mental health problems. However, the precise reasons for mental illnesses are still unclear and uncertain.

1.2. Types of Mental Health Problems

Mental illness can affect the cognition, emotion, and behaviour among the people. For children, their ability to learn could be interfered by mental disorders. Besides that, mental illness can cause inconvenience to the adults, especially in their families, workplaces, and in the society. There are many types of mental disorders commonly known as schizophrenia, depression, bipolar disorder, and anxiety.

Schizophrenia is a mental illness that interrupted by events of psychotic symptoms, which are hallucinations and delusions. Hallucinations are experiences that are not comprehensible to others. Meanwhile, delusions are impressions that are held by the patients although contradicted by the rational and real arguments. Schizophrenia is often diagnosed by symptoms such as social withdrawal, irritability, and increasing strange behaviours. Studies of whether an early diagnosis of such symptoms and intervention could improve the outcomes are still in progress.

The primary symptom of depression is an interference of the mood, which is usually severe sadness. Sometimes, anger, irritability, and loss of interests might dominate the symptoms of the depression. In terms of physiological symptoms, sleep disturbance, appetite disturbance, and decreased in energy are commonly shown across cultures. The cognitive symptoms such as slow thinking, suicidal thoughts, and guilt might occur among the patients. Most of the individuals that suffer from depression will have recurrence. Many individuals do not recover completely and they might have a form of chronic mild depression. Bipolar disorder is another mental disorder identified by the episode of mania and depression. Sometimes, there is an episode mixed with both mania and depression. Mania is known by irritability, increased in energy, and decreased need for sleep. Individuals that experience mania often exhibit reckless behaviours. Meanwhile, a depressive episode for bipolar disorder is almost the same as the depression symptoms. Some studies report some recovery to baseline functioning between episodes; however, many patients will have residual symptoms that cause impairment. Another common mental disorder is an anxiety disorder, which is usually identified as an inability to regulate fear or worry. Panic disorders belong to this category, which appears to be unexpected panic attacks and intense fear.

The physiological symptoms that are caused by panic disorder include a racing heart, sweating, and dizziness. Generalized anxiety disorder is characterized by excessive worry. Emotional numbness caused by traumatic events characterizes posttraumatic stress disorder. Individuals that have a social anxiety disorder are frequently afraid of social situations. Surveys show that delays in seeking professional treatment for an anxiety disorder are widespread.

1.3. Data Mining and Machine Learning

In modern days, the management and processing of data have fully grown into a popular topic in the field of computer science. Data mining is knowledge discovery in databases, which is discovering useful patterns and relationships in large volumes of data. Within the medical field, data mining techniques are increasingly applied for tasks such as text expression, drug Design, and genomics. Data mining techniques can be separated into two forms, which are supervised learning and unsupervised learning. For unsupervised learning, it determines the object’s similarity and detects patterns through the group’s data. It can be grouped into clustering, association, summarizing, and sequence discovery. Unsupervised learning is particularly valuable in helping to identify the structure of the data automatically through learning inherent from input data when the data set is unlabelled. In short, data mining is a crucial technique in the role of Computer science.

The complexity of the data sets collected can be solved rapidly and swiftly through data mining. In addition, many parties can gain an advantage using data mining for better outcomes and solutions of their challenging problems. Machine learning is an application of artificial intelligence (AI), which implements systems with the capability to learn and improve from experience without being explicitly programmed. Machine learning has offered essential advantages to a wide range of areas such as speech recognition, computer vision, and natural language processing. It is allowing many researchers to extract meaningful information from the data, provide personalized wisdom, and establish automated intelligent systems. It is believed that machine learning introduced many types of approaches and learning. For instance, the commonly used machine learning approaches are supervised learning and unsupervised learning.

Supervised learning is an approach that predicts the outcome result with given labelled data input. Supervised learning is excellent at classification and regression problems. The purpose of this learning is to make sense of data toward the specific measurements. The unsupervised learning is in contrast to the supervised learning, which tries to make sense of data in itself. In unsupervised learning, there are no measurements or guidelines. Additionally, the ensemble learning is a process where the classifiers combined and generated strategically to solve a specific problem. Primary usage of ensemble learning is to improve the performance of a model or reduce the probability of selecting models with poor performance.

Moreover, neural networks and deep learning have recently become better known among machine learning approaches due to their ability to solve many problems such as image recognition, speech recognition, and natural language processing. The approaches are based on the neuronal networks of the brain where they enable the algorithms to learn from the observational data. In the medical field, machine learning algorithms have been used to discover new drugs, perform radiology analysis, predict epidemic outbreaks, and diagnose diseases. Generally, machine learning algorithms are tools to analyse the massive medical data sets. They are utilized as tools in assisting for medical diagnosis as they became more reliable in their performance. From time to time, machine learning and data mining Approaches continue to develop rapidly.

Powerful algorithms and more advanced neural networks, decision trees, gradient boosting, and others were introduced and applied to solve more complicated mental health illness problems.

1.4. Objectives

- To manage the healthy and productive workplace by understanding the importance of mental health.

- This project analysis the current trends and attitude of the employers towards the mental health.

- To conduct wellness program for identify those at risk and connect them to treatment.

1.5. Problem Statement

To study how often does mental health affect the work of the employees-with reference to their attitude, family history and gender.

02. Methodology

In this report to know the better result and the proper workflow of the project there some planning phase are used and this planning phase contains step by step modules used in the dataset to know the predictions in better ways .

In the first stage the raw data is collected from different sites and later different dataset are studies and according to the best the dataset is selected and further process is carried out.

In the second stage we carried out data pre-processing on our selected raw dataset and all the different types of operations such as data handling and data cleaning is carried out.

In the further stages additional step were taken on the final dataset such as exploratory data analysis and later on machine learning modules are used building five models and then comparing their performance based on accuracy. The three classifiers are namely**:** Logistic Regression**,** k-Nearest Neighbors Classifier, Random Forests Classifier**,** and two ensemble meta-algorithm: Bagging and Boosting. To predict the mental health of an employee and at the end we have discuss the conclusion and future scope of this report.

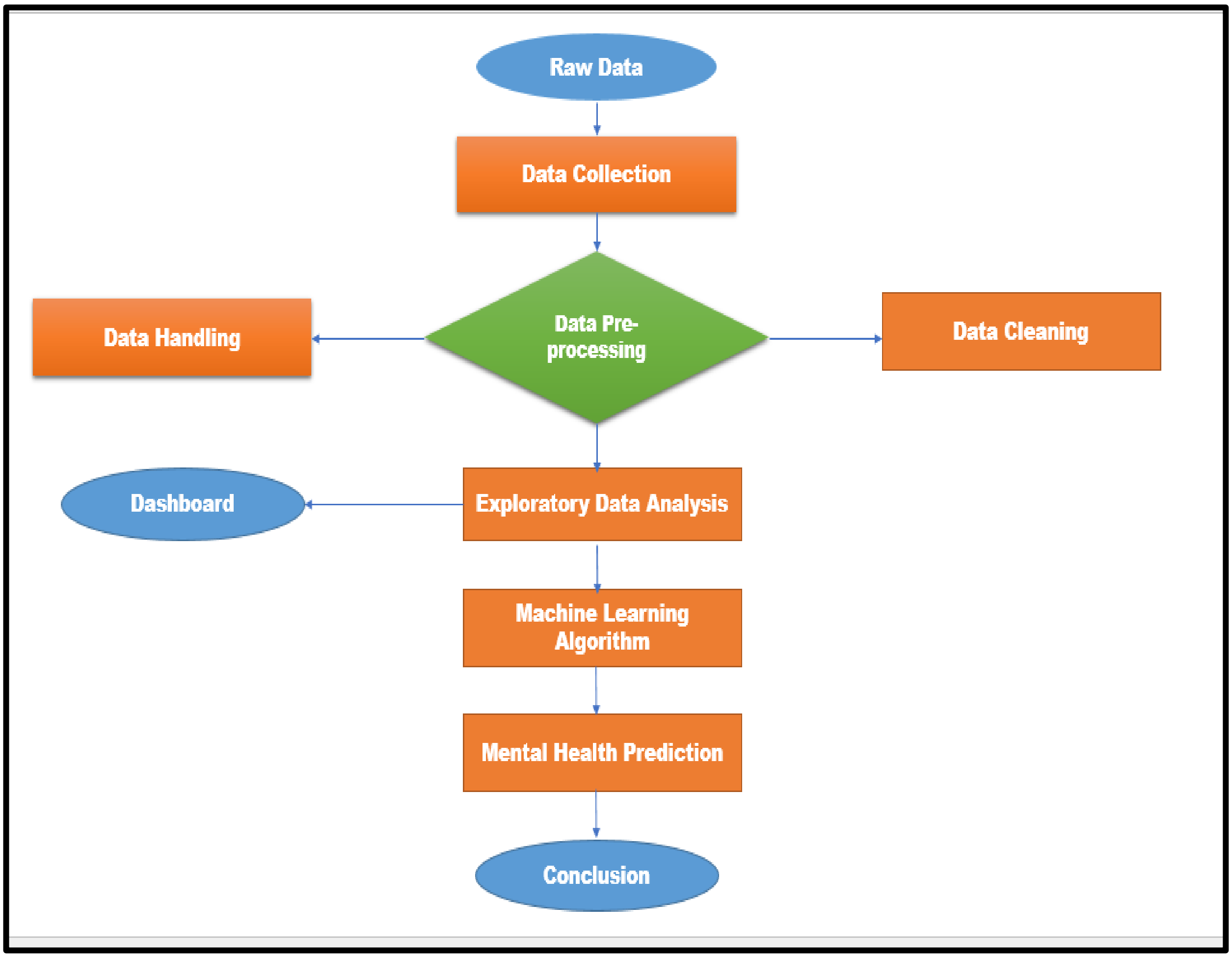


Figure 01:- Methodology

03. System Requirements

Hardware Requirement

- Platform – Windows
- RAM – 8 GB of RAM
- Peripheral Devices- Mouse, Keyboard, Monitor.
- A network connection for data recovering over network.

Software Requirement

- Jupyter-Python 3.
- MySQL.
- Machine Learning.
- Power BI.

04. Functional Requirements

4.1. Python 3:

- Python is a general purpose and high level programming language.
- It is use for developing desktop GUI applications, websites and web applications.
- Python allows focusing on core functionality of the application by taking care of common programming tasks.
- Python is derived from many other languages, we use also for pandas.

4.2. MySQL:

- It is an open source relational database management system
- A relational database organizes data into one or more data tables in which data may be related to each other. These relationship structure the data.
- SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database.
- In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computers storage system, manage users, allows for network access and facilities testing database integrity and creation of backups.

4.3. Machine Learning:

- Machine learning is a method of data analysis that automates analytical model building. It is abranch of artificial intelligence based on the idea that systems can learn from data, identify Patterns and make decisions with minimal human intervention.
- Machine learning is used in internet search engines, email filters to sort out spam, websites tomake personalized recommendations, banking software to detect unusual transactions. And prediction.
- We use machine learning for analysis and prediction of our data. We use logistic regression, Random Forest and K-Means algorithm and regarding this all algorithm tries to analyze and find the maximum accuracy.

4.4. PowerBI:

- Data visualization is the graphical representation of information and data.
- It helps create interactive elements like charts, graphs, and maps, data visualization toolsprovide an accessible way to see and understand trends, outliers, and patterns in data.
- PowerBI is widely used for Business Intelligence but is not limited to it.
- It helps create interactive graphs and charts in the form of dashboards and worksheets togain business insights.
- All of this is made possible with gestures as simple as drag and drop.

05. Data Collections

5.1. Dataset

The dataset we are using is from the Mental Health in Tech Survey. With over 4490 responses, this survey was the largest survey done on mental health in the tech industry. They have been doing this survey each year since then but we found the data from 2014-2016-2017-2018-2019-2020-2021 best suited for my project.

As this dataset is from the employees' perspective, it provides a closer look into the current situation of mental health resources in the workplace as well as their effectiveness.

5.2. Data Exploration

There are 36 features in the dataset. We choose 'Treatment' to be the target variable. It indicates whether or not an individual has sought treatment for a mental health condition.

The age column has some values that don't make sense like -1726, 329, 99999999999, -1, and -29, they need to be dropped. The Gender column consists of the unique values, most of which cover three main groups 'male', 'female', and 'trans'.

5.3. Data Preprocessing

The data did have many null values so we replaced them with the most common value in their respective columns and drop the fifty percent null values columns. The 'Gender' column is grouped into three main categories: Male, Female, and Trans to maintain consistency, unlike values it had earlier. Also, rows containing irrelevant values in the 'age' column have been removed.

5.3.1 Data Cleaning Process

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores.

Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

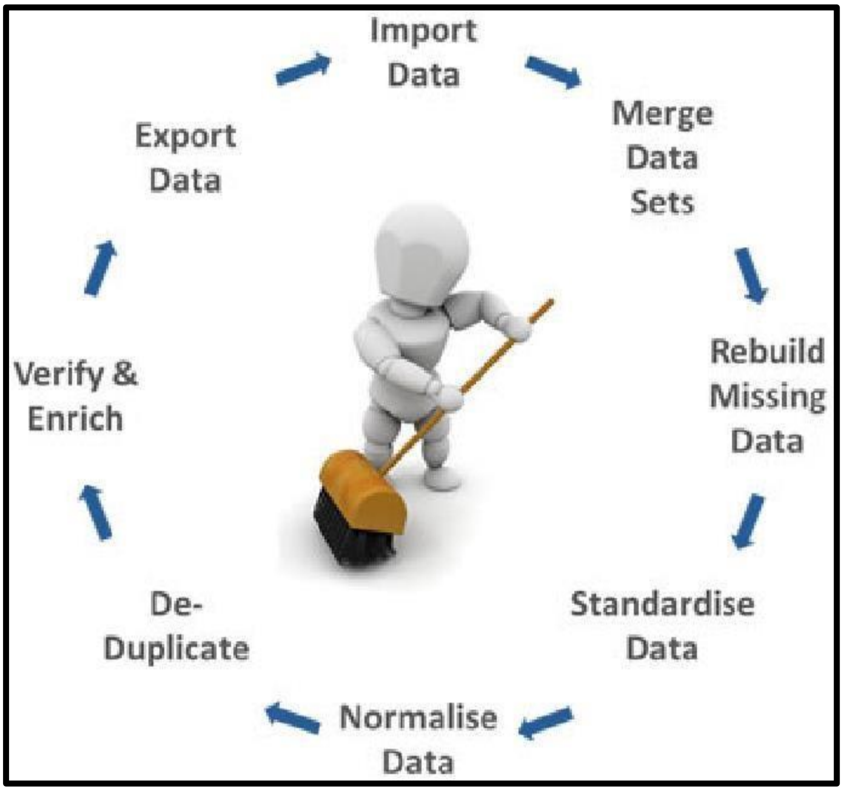


Figure 02:- Data Cleaning

06. Machine Learning Algorithms

In this project we apply various different types of Classification and Regression Algorithms such as logistic Regression, k-Nearest Neighbors Classifier, Random Forests Classifiers and two ensemble meta-algorithm: Bagging and Boosting.

These are the most popular classification algorithms used in machine learning, therefore, we wanted to compare their performance in predicting whether or not an individual will choose to go for treatment of mental health issues. The features that we are using to make predictions are 'Age', 'Gender', 'family history', 'benefits', 'care options', 'anonymity', 'leave', and 'work interfere'. Since these features are the most related to the target variable.

6.1. K-Nearest Neighbors:

- K Means is a clustering algorithm which divides observations into k clusters. Theabbreviation KNN stands for "K-Nearest Neighbor". It is a supervised machine learning algorithm.
- K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure

Pros:

- Quick calculation time, Simple Algorithm
- High Accuracy

Cons:

- Accuracy depends on the quality of the data.
- Require high memory

6.2. Random Forest:

- A random forest is a machine learning technique that's used to solve regression and classification problems.
- It utilizes ensemble learning, which is a technique that combines many classifiers toprovide solutions to complex problems. A random forest algorithm consists of manydecision trees

Pros:

- It reduces overfitting in decision trees and helps to improve the accuracy.

Cons:

- A large number of trees can make the algorithm too slow
- Ineffective for real-time predictions.

6.3. Logistic Regression:

- Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam online transactions Fraud or not Fraud, Malignant or Benign. And we use this for data surety.

Pros:

- It is easy to implement and very efficient to train.
- Can work with numerical and categorical features.

Cons:

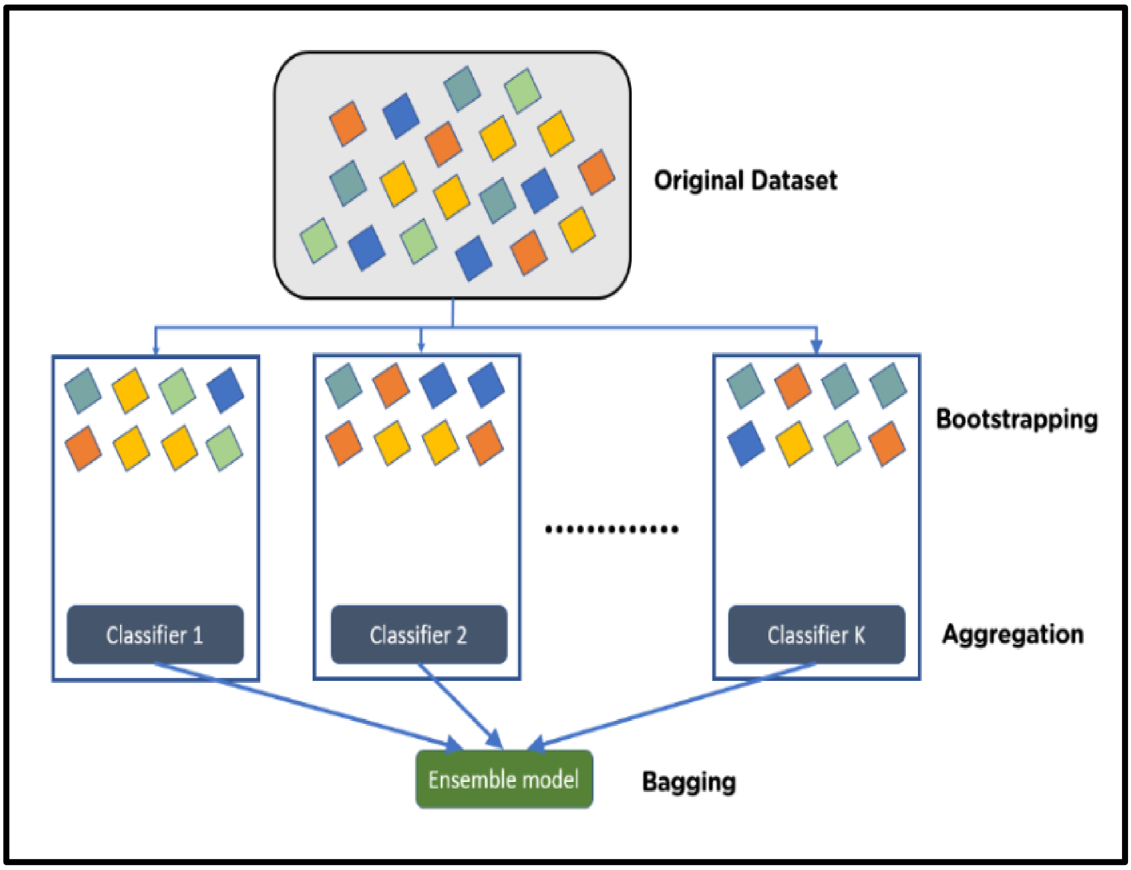
- It tends to over fit.
- Difficult to capture complex relationships.

6.4. Bagging:

Bagging, also known as Bootstrap aggregating, is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model. Bagging avoids overfitting of data and is used for both regression and classification models, specifically for decision tree algorithms.

Steps to Perform Bagging:

- Consider there are n observations and m features in the training set. You need to select a random sample from the training dataset without replacement
- A subset of m features is chosen randomly to create a model using sample observations
- The feature offering the best split out of the lot is used to split the nodes
- The tree is grown, so you have the best root nodes
- The above steps are repeated n times. It aggregates the output of individual decision trees to give the best prediction



Figures 03 Bagging

Advantages of Bagging in Machine Learning:

- Bagging minimizes the overfitting of data
- It improves the model's accuracy
- It deals with higher dimensional data efficiently.

6.5. Boosting:

Boosting is a method used in machine learning to reduce errors in predictive data analysis. Data scientists train machine learning software, called machine learning models, on labeled data to make guesses about unlabeled data. A single machine learning model might make prediction errors depending on the accuracy of the training dataset.

Advantages of Boosting in Machine Learning:

Ease of implementation

Boosting has easy-to-understand and easy-to-interpret algorithms that learn from their mistakes. These algorithms don't require any data preprocessing, and they have built-in routines to handle missing data. In addition, most languages have built-in libraries to implement boosting algorithms with many parameters that can fine-tune performance.

Reduction of bias

Bias is the presence of uncertainty or inaccuracy in machine learning results. Boosting algorithms combine multiple weak learners in a sequential method, which iteratively improves observations. This approach helps to reduce high bias that is common in machine learning models.

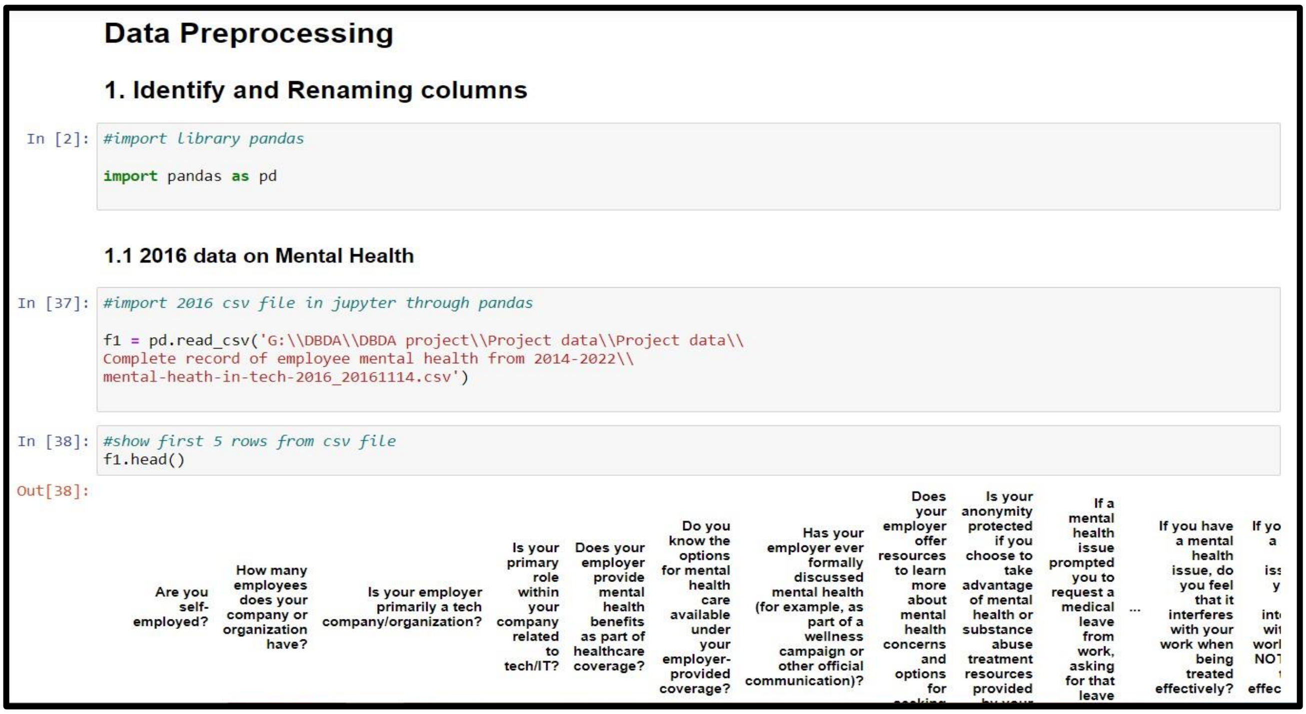
Computational efficiency

Boosting algorithms prioritize features that increase predictive accuracy during training. They can help to reduce data attributes and handle large datasets efficiently.

07. Implementation and Coding

7.1. Preprocessing

7.1.1 Identify and Renaming Columns



Data Preprocessing

1. Identify and Renaming columns

In [2]:
```
#import library pandas

import pandas as pd
```

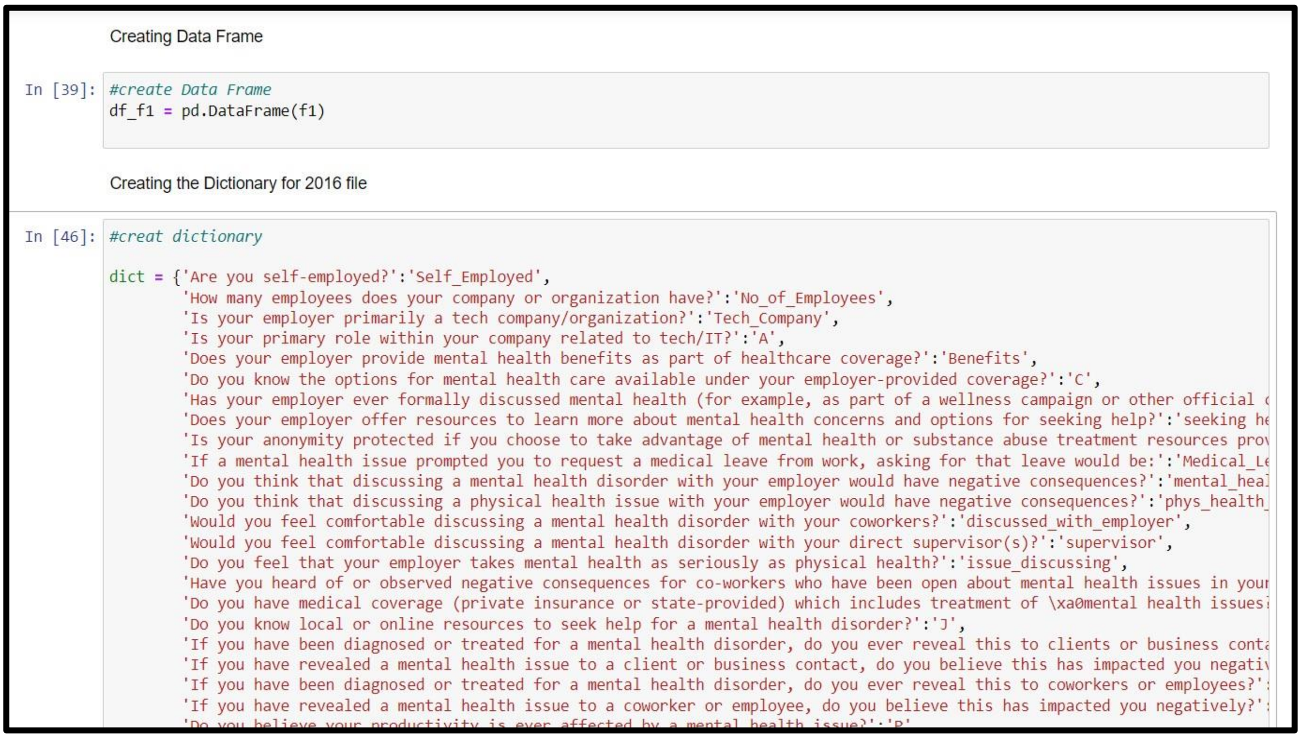
1.1 2016 data on Mental Health

In [37]:
```
#import 2016 csv file in jupyter through pandas

f1 = pd.read_csv('G:\\DBDA\\DBDA project\\Project data\\Project data\\
Complete record of employee mental health from 2014-2022\\
mental-heath-in-tech-2016_20161114.csv')
```

In [38]:
```
#show first 5 rows from csv file
f1.head()
```

Out[38]:

Are you self-employed?	How many employees does your company or organization have?	Is your employer primarily a tech company/organization?	Is your primary role within your company related to tech/IT?	Does your employer provide mental health benefits as part of healthcare coverage?	Do you know the options for mental health care available under your employer-provided coverage?	Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	Does your employer offer resources to learn more about mental health concerns and options for	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided	If a mental health issue prompted you to request a medical leave from work, asking for that leave	...	If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?	If yo a iss y int wi worl NOT effec

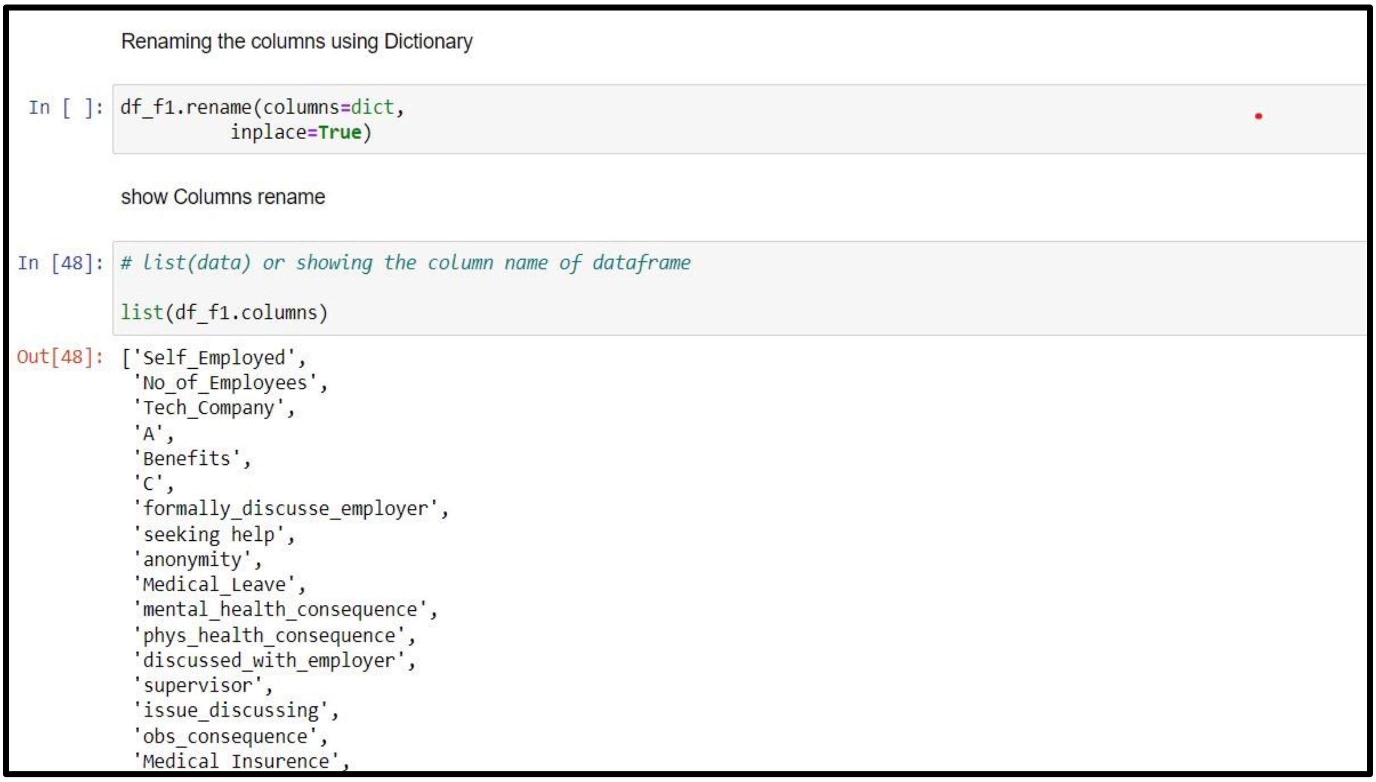


Creating Data Frame

In [39]:
```
#create Data Frame
df_f1 = pd.DataFrame(f1)
```

Creating the Dictionary for 2016 file

In [46]:
```
#creat dictionary

dict = {'Are you self-employed?':'Self_Employed',
        'How many employees does your company or organization have?':'No_of_Employees',
        'Is your employer primarily a tech company/organization?':'Tech_Company',
        'Is your primary role within your company related to tech/IT?':'A',
        'Does your employer provide mental health benefits as part of healthcare coverage?':'Benefits',
        'Do you know the options for mental health care available under your employer-provided coverage?':'C',
        'Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official c
        'Does your employer offer resources to learn more about mental health concerns and options for seeking help?':'seeking he
        'Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources prov
        'If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:':'Medical_Le
        'Do you think that discussing a mental health disorder with your employer would have negative consequences?':'mental_heal
        'Do you think that discussing a physical health issue with your employer would have negative consequences?':'phys_health_
        'Would you feel comfortable discussing a mental health disorder with your coworkers?':'discussed_with_employer',
        'Would you feel comfortable discussing a mental health disorder with your direct supervisor(s)?':'supervisor',
        'Do you feel that your employer takes mental health as seriously as physical health?':'issue_discussing',
        'Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your
        'Do you have medical coverage (private insurance or state-provided) which includes treatment of \xa0mental health issues:
        'Do you know local or online resources to seek help for a mental health disorder?':'J',
        'If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business conta
        'If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negativ
        'If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?':
        'If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?':
        'Do you believe your productivity is ever affected by a mental health issue?':'P'
```



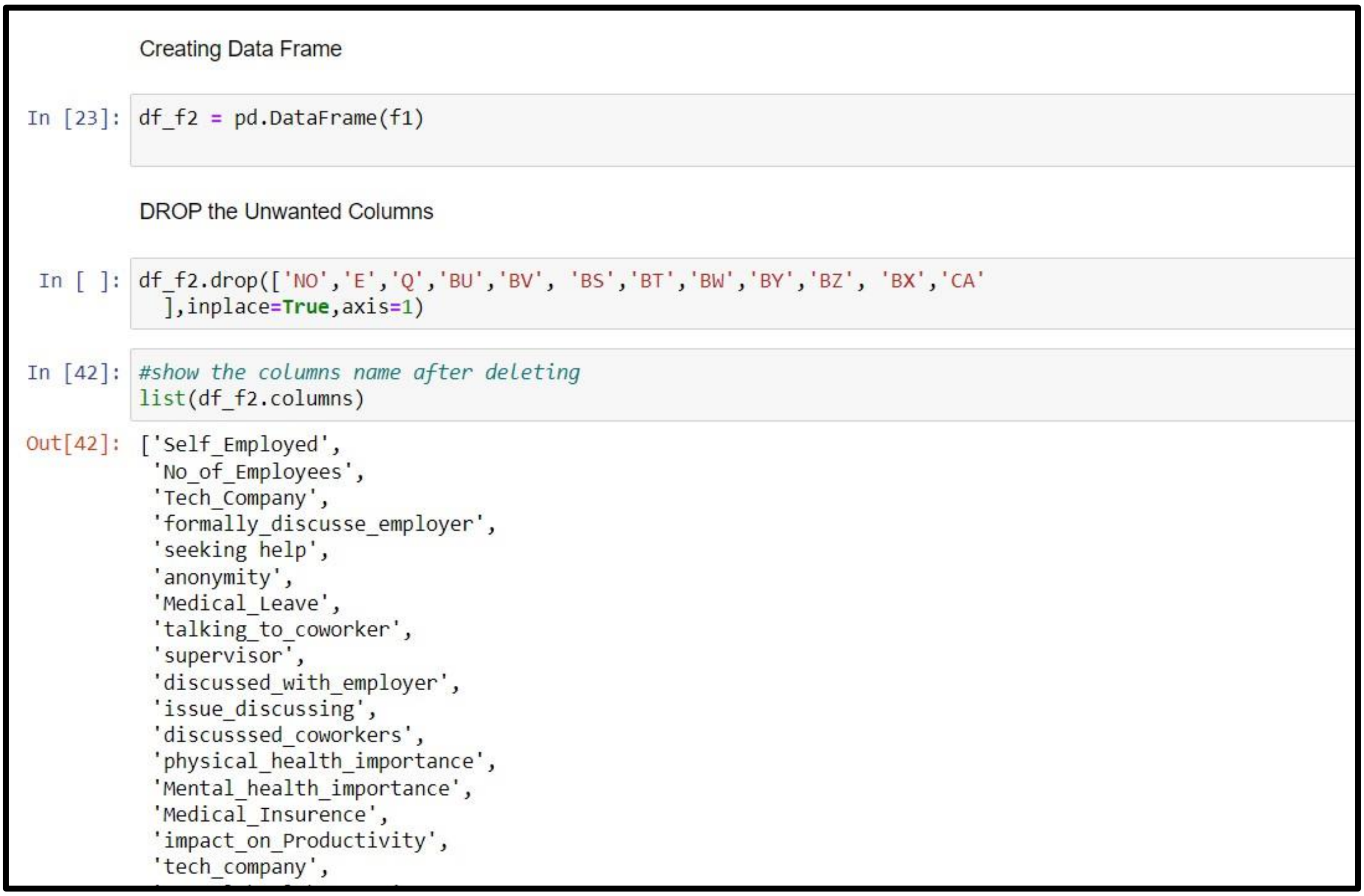
Renaming the columns using Dictionary

```
In [ ]: df_f1.rename(columns=dict,
                 inplace=True)
```

show Columns rename

```
In [48]: # list(data) or showing the column name of dataframe

         list(df_f1.columns)

Out[48]: ['Self_Employed',
          'No_of_Employees',
          'Tech_Company',
          'A',
          'Benefits',
          'C',
          'formally_discusse_employer',
          'seeking help',
          'anonymity',
          'Medical_Leave',
          'mental_health_consequence',
          'phys_health_consequence',
          'discussed_with_employer',
          'supervisor',
          'issue_discussing',
          'obs_consequence',
          'Medical Insurence',
```

7.1.2 Deleting Unwanted Columns



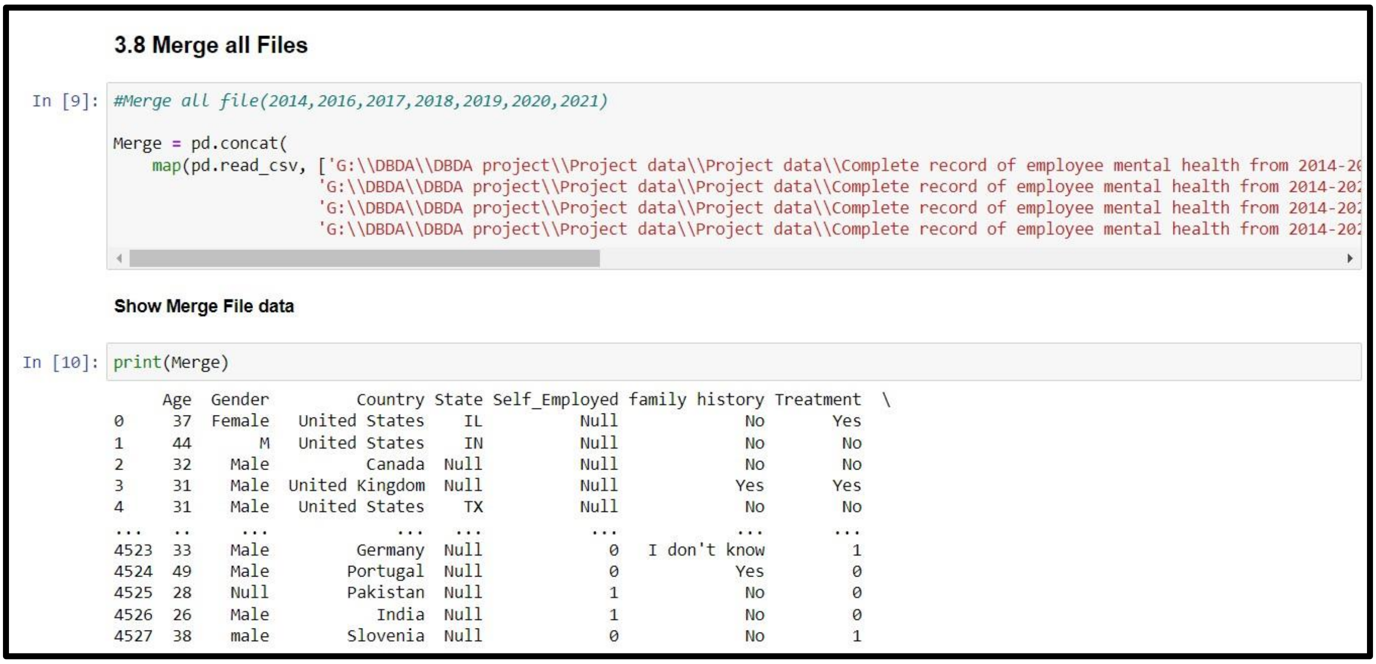
Creating Data Frame

```
In [23]: df_f2 = pd.DataFrame(f1)
```

DROP the Unwanted Columns

```
In [ ]: df_f2.drop(['NO','E','Q','BU','BV', 'BS','BT','BW','BY','BZ', 'BX','CA'
          ],inplace=True,axis=1)

In [42]: #show the columns name after deleting
         list(df_f2.columns)

Out[42]: ['Self_Employed',
          'No_of_Employees',
          'Tech_Company',
          'formally_discusse_employer',
          'seeking help',
          'anonymity',
          'Medical_Leave',
          'talking_to_coworker',
          'supervisor',
          'discussed_with_employer',
          'issue_discussing',
          'discusssed_coworkers',
          'physical_health_importance',
          'Mental_health_importance',
          'Medical_Insurence',
          'impact_on_Productivity',
          'tech_company',
```

7.1.3. Merging Files



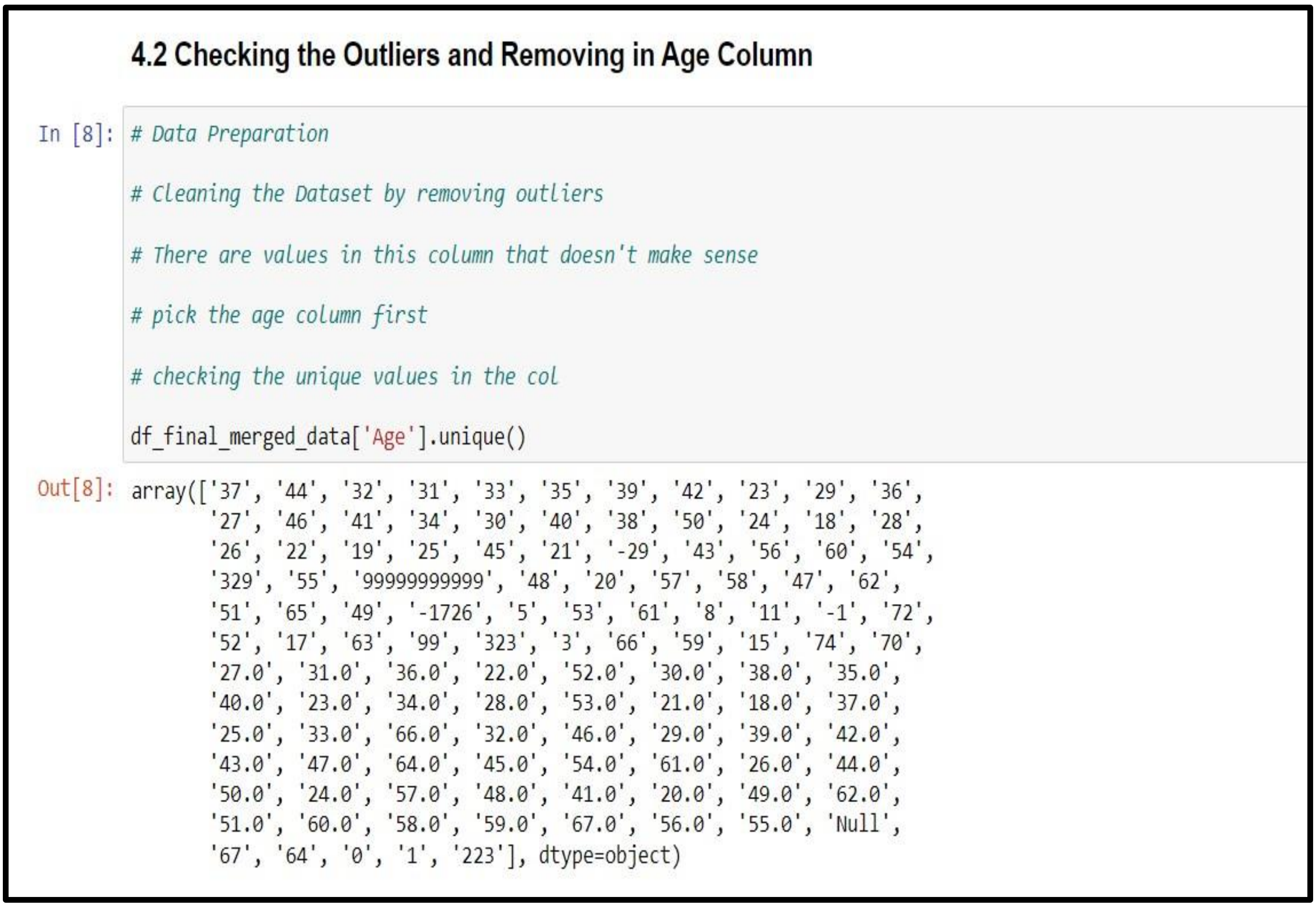
3.8 Merge all Files

```
In [9]: #Merge all file(2014,2016,2017,2018,2019,2020,2021)

        Merge = pd.concat(
            map(pd.read_csv, ['G:\\DBDA\\DBDA project\\Project data\\Project data\\Complete record of employee mental health from 2014-20
                              'G:\\DBDA\\DBDA project\\Project data\\Project data\\Complete record of employee mental health from 2014-202
                              'G:\\DBDA\\DBDA project\\Project data\\Project data\\Complete record of employee mental health from 2014-202
                              'G:\\DBDA\\DBDA project\\Project data\\Project data\\Complete record of employee mental health from 2014-202
```

Show Merge File data

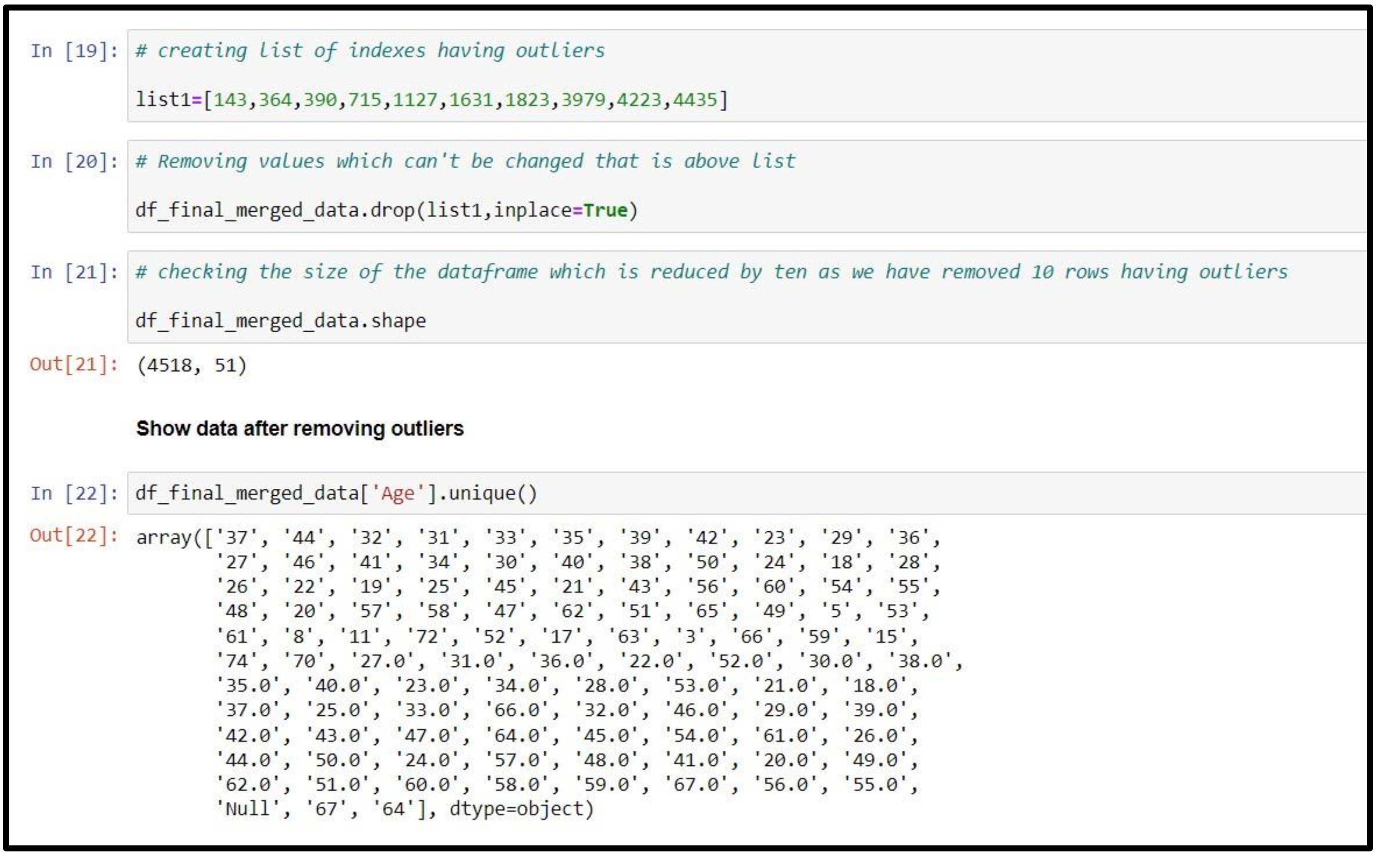
```
In [10]: print(Merge)

      Age  Gender         Country State Self_Employed family history Treatment  \
0      37  Female   United States    IL          Null             No       Yes
1      44       M   United States    IN          Null             No        No
2      32    Male          Canada  Null          Null             No        No
3      31    Male  United Kingdom  Null          Null            Yes       Yes
4      31    Male   United States    TX          Null             No        No
...   ..     ...             ...   ...           ...            ...       ...
4523  33    Male         Germany  Null             0   I don't know         1
4524  49    Male        Portugal  Null             0            Yes         0
4525  28    Null        Pakistan  Null             1             No         0
4526  26    Male           India  Null             1             No         0
4527  38    male        Slovenia  Null             0             No         1
```

7.1.4. Handling Outliers



4.2 Checking the Outliers and Removing in Age Column

```
In [8]: # Data Preparation

        # Cleaning the Dataset by removing outliers

        # There are values in this column that doesn't make sense

        # pick the age column first

        # checking the unique values in the col

        df_final_merged_data['Age'].unique()

Out[8]: array(['37', '44', '32', '31', '33', '35', '39', '42', '23', '29', '36',
               '27', '46', '41', '34', '30', '40', '38', '50', '24', '18', '28',
               '26', '22', '19', '25', '45', '21', '-29', '43', '56', '60', '54',
               '329', '55', '99999999999', '48', '20', '57', '58', '47', '62',
               '51', '65', '49', '-1726', '5', '53', '61', '8', '11', '-1', '72',
               '52', '17', '63', '99', '323', '3', '66', '59', '15', '74', '70',
               '27.0', '31.0', '36.0', '22.0', '52.0', '30.0', '38.0', '35.0',
               '40.0', '23.0', '34.0', '28.0', '53.0', '21.0', '18.0', '37.0',
               '25.0', '33.0', '66.0', '32.0', '46.0', '29.0', '39.0', '42.0',
               '43.0', '47.0', '64.0', '45.0', '54.0', '61.0', '26.0', '44.0',
               '50.0', '24.0', '57.0', '48.0', '41.0', '20.0', '49.0', '62.0',
               '51.0', '60.0', '58.0', '59.0', '67.0', '56.0', '55.0', 'Null',
               '67', '64', '0', '1', '223'], dtype=object)
```

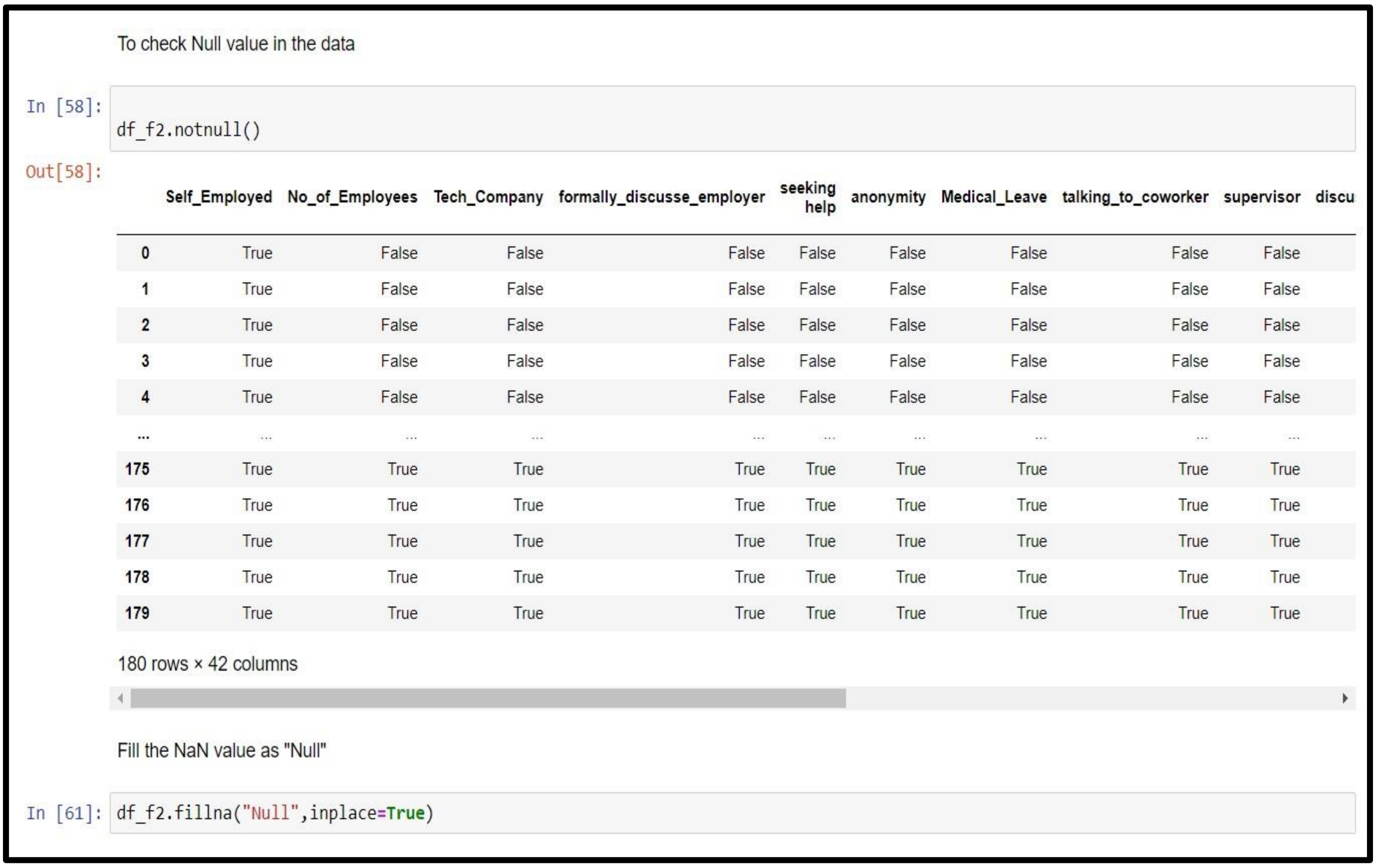


```
In [19]: # creating list of indexes having outliers

         list1=[143,364,390,715,1127,1631,1823,3979,4223,4435]

In [20]: # Removing values which can't be changed that is above list

         df_final_merged_data.drop(list1,inplace=True)

In [21]: # checking the size of the dataframe which is reduced by ten as we have removed 10 rows having outliers

         df_final_merged_data.shape

Out[21]: (4518, 51)
```

Show data after removing outliers

```
In [22]: df_final_merged_data['Age'].unique()

Out[22]: array(['37', '44', '32', '31', '33', '35', '39', '42', '23', '29', '36',
                '27', '46', '41', '34', '30', '40', '38', '50', '24', '18', '28',
                '26', '22', '19', '25', '45', '21', '43', '56', '60', '54', '55',
                '48', '20', '57', '58', '47', '62', '51', '65', '49', '5', '53',
                '61', '8', '11', '72', '52', '17', '63', '3', '66', '59', '15',
                '74', '70', '27.0', '31.0', '36.0', '22.0', '52.0', '30.0', '38.0',
                '35.0', '40.0', '23.0', '34.0', '28.0', '53.0', '21.0', '18.0',
                '37.0', '25.0', '33.0', '66.0', '32.0', '46.0', '29.0', '39.0',
                '42.0', '43.0', '47.0', '64.0', '45.0', '54.0', '61.0', '26.0',
                '44.0', '50.0', '24.0', '57.0', '48.0', '41.0', '20.0', '49.0',
                '62.0', '51.0', '60.0', '58.0', '59.0', '67.0', '56.0', '55.0',
                'Null', '67', '64'], dtype=object)
```

7.1.5 Null Handling



To check Null value in the data

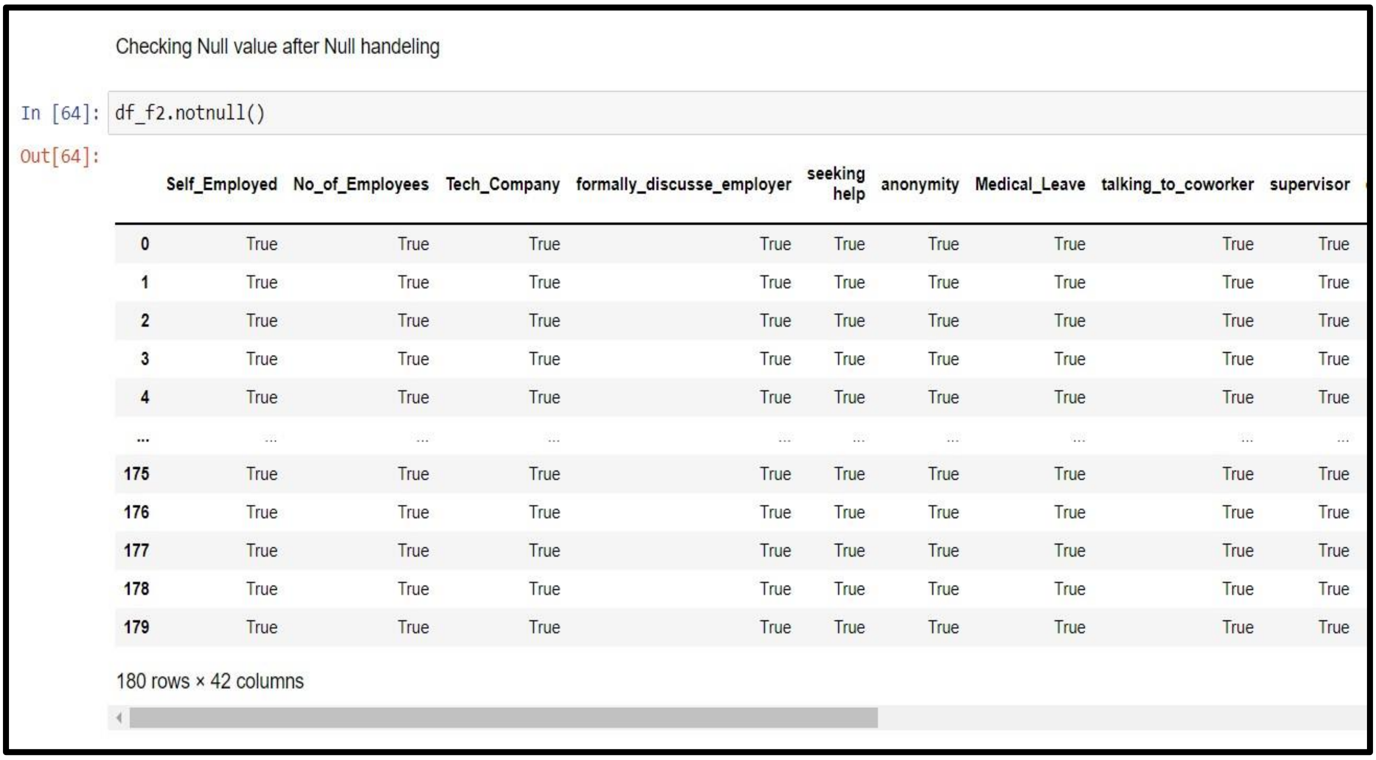
```
In [58]: df_f2.notnull()

Out[58]:
```

	Self_Employed	No_of_Employees	Tech_Company	formally_discusse_employer	seeking help	anonymity	Medical_Leave	talking_to_coworker	supervisor	discu
0	True	False	False	False	False	False	False	False	False	
1	True	False	False	False	False	False	False	False	False	
2	True	False	False	False	False	False	False	False	False	
3	True	False	False	False	False	False	False	False	False	
4	True	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
175	True	True	True	True	True	True	True	True	True	
176	True	True	True	True	True	True	True	True	True	
177	True	True	True	True	True	True	True	True	True	
178	True	True	True	True	True	True	True	True	True	
179	True	True	True	True	True	True	True	True	True	

180 rows × 42 columns

Fill the NaN value as "Null"

```
In [61]: df_f2.fillna("Null",inplace=True)
```



Checking Null value after Null handeling

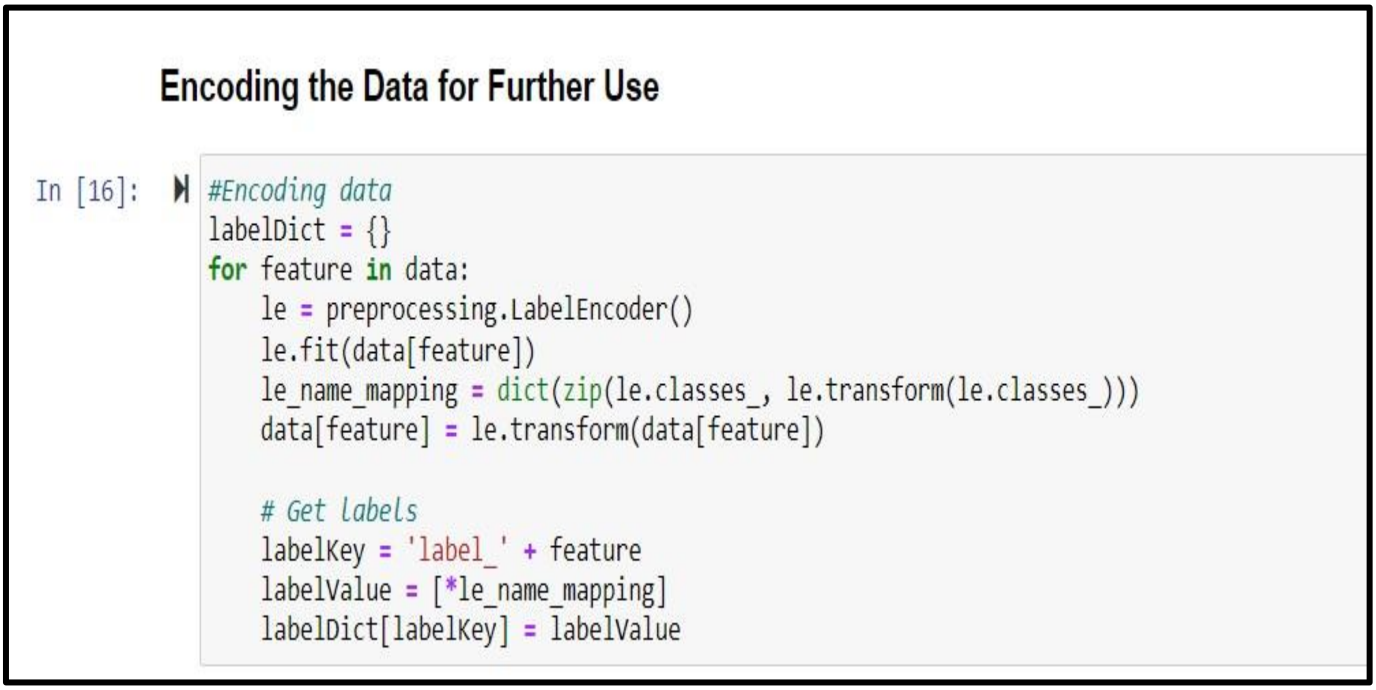
In [64]:
```
df_f2.notnull()
```

Out[64]:

	Self_Employed	No_of_Employees	Tech_Company	formally_discusse_employer	seeking help	anonymity	Medical_Leave	talking_to_coworker	supervisor
0	True	True	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True	True	True
2	True	True	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True	True
...	...	...	...	...	...	...	...	...	...
175	True	True	True	True	True	True	True	True	True
176	True	True	True	True	True	True	True	True	True
177	True	True	True	True	True	True	True	True	True
178	True	True	True	True	True	True	True	True	True
179	True	True	True	True	True	True	True	True	True

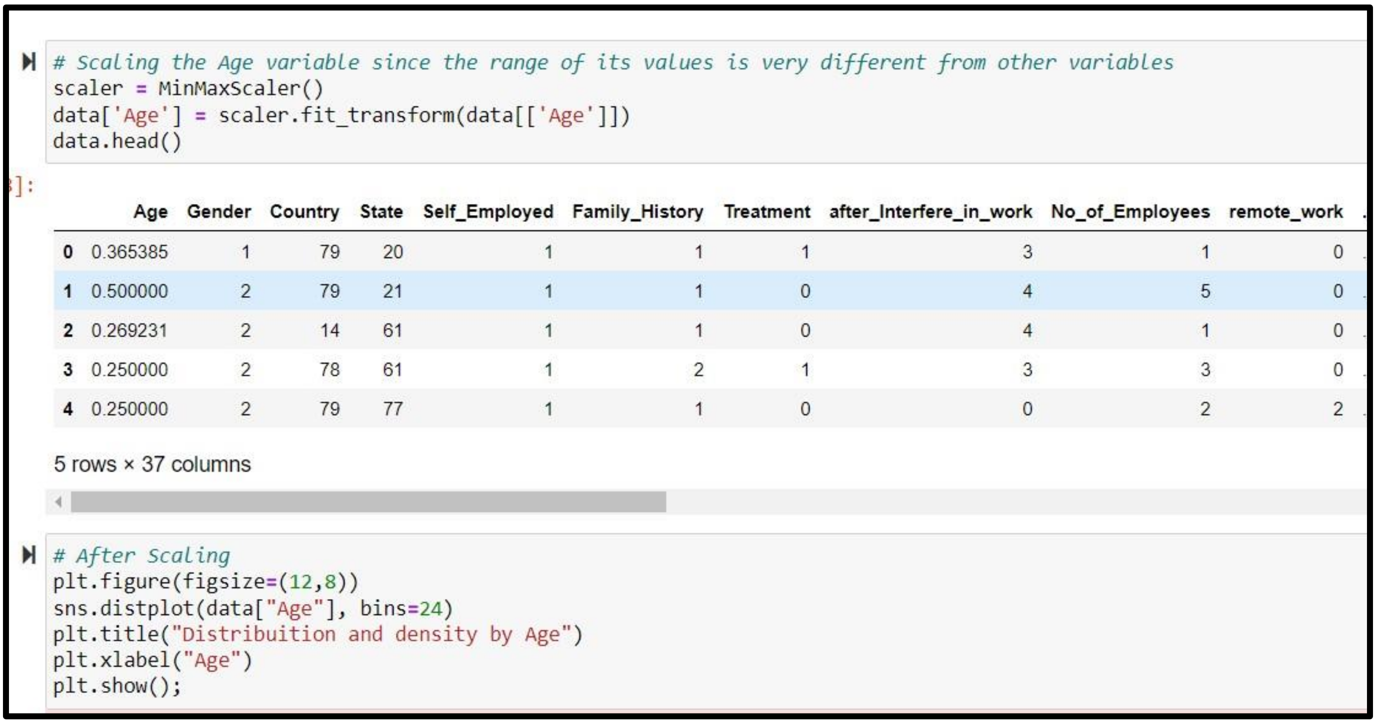
180 rows × 42 columns

7.2. Data Exploratory analysis



Encoding the Data for Further Use

In [16]:
```
#Encoding data
labelDict = {}
for feature in data:
    le = preprocessing.LabelEncoder()
    le.fit(data[feature])
    le_name_mapping = dict(zip(le.classes_, le.transform(le.classes_)))
    data[feature] = le.transform(data[feature])

    # Get labels
    labelKey = 'label_' + feature
    labelValue = [*le_name_mapping]
    labelDict[labelKey] = labelValue
```



```
# Scaling the Age variable since the range of its values is very different from other variables
scaler = MinMaxScaler()
data['Age'] = scaler.fit_transform(data[['Age']])
data.head()
```

]:

	Age	Gender	Country	State	Self_Employed	Family_History	Treatment	after_Interfere_in_work	No_of_Employees	remote_work
0	0.365385	1	79	20	1	1	1	3	1	0
1	0.500000	2	79	21	1	1	0	4	5	0
2	0.269231	2	14	61	1	1	0	4	1	0
3	0.250000	2	78	61	1	2	1	3	3	0
4	0.250000	2	79	77	1	1	0	0	2	2

5 rows × 37 columns

```
# After Scaling
plt.figure(figsize=(12,8))
sns.distplot(data["Age"], bins=24)
plt.title("Distribuition and density by Age")
plt.xlabel("Age")
plt.show();
```

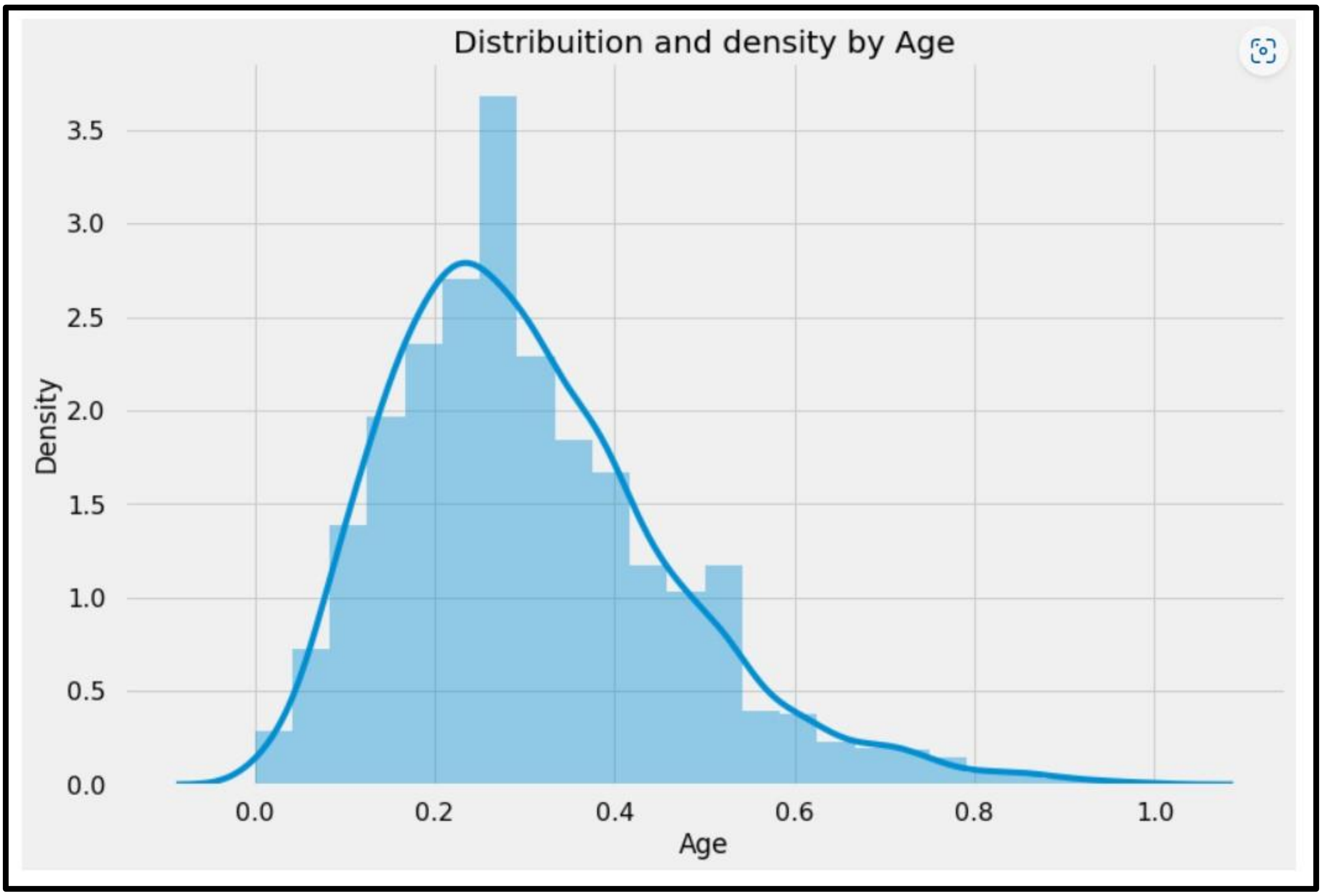


Figure 04 Age Histogram

7.2.1. Correlations Matrix

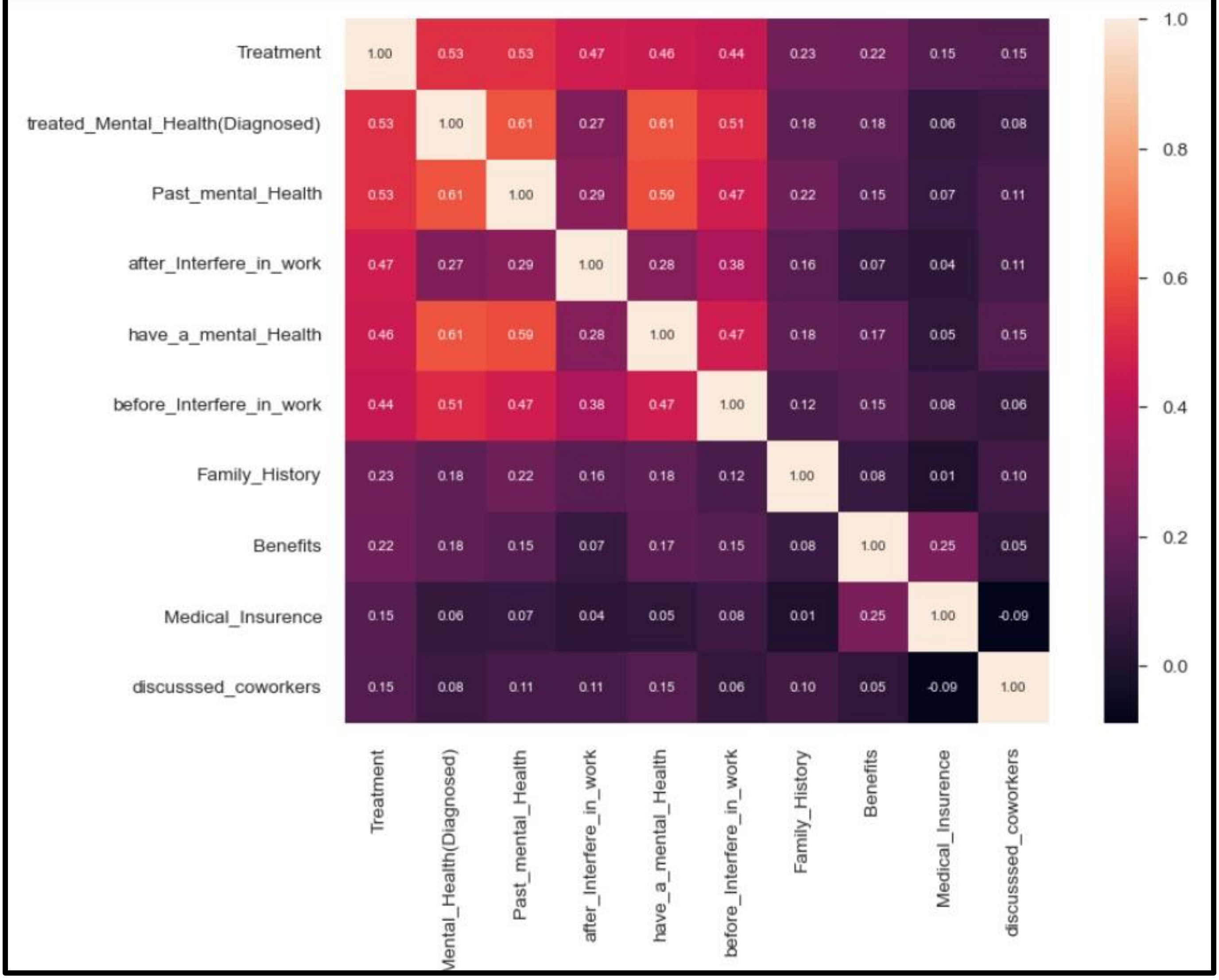


Figure 05 Correlation Matrix

08. Data Visualization

Data visualization is the graphical representation of information and data in a pictorial or graphical format (Example: charts, graphs, and maps). Data visualization tools provide an accessible way to see and understand trends, patterns in data, and outliers. Data visualization tools and technologies are essential to analyzing massive amounts of information and making data-driven decisions. The concept of using pictures is to understand data that has been used for centuries. General types of data visualization are Charts, Tables, Graphs, Maps, and Dashboards.

Advantages of Data Visualization:

Better Agreement: In business, for numerous periods, it happens that we need to look at the exhibitions of two components or two situations. A conventional methodology is to experience the massive information of both the circumstances and afterward examine it. This will clearly take a great deal of time.

A Superior Method: It can tackle the difficulty of placing the information of both perspectives into the pictorial structure. This will unquestionably give a superior comprehension of the circumstances. For instance, Google patterns assist us with understanding information identified with top ventures or inquiries in pictorial or graphical structures.

Simple Sharing of Data: With the representation of the information, organizations present another arrangement of correspondence. Rather than sharing the cumbersome information, sharing the visual data will draw in and pass on across the data which is more absorbable.

Investigating Openings and Patterns: With the huge loads of information present, business chiefs can discover the profundity of information in regard to the patterns and openings around them. Utilizing information representation, the specialists can discover examples of the conduct of their clients, subsequently preparing for them to investigate patterns and open doors for business

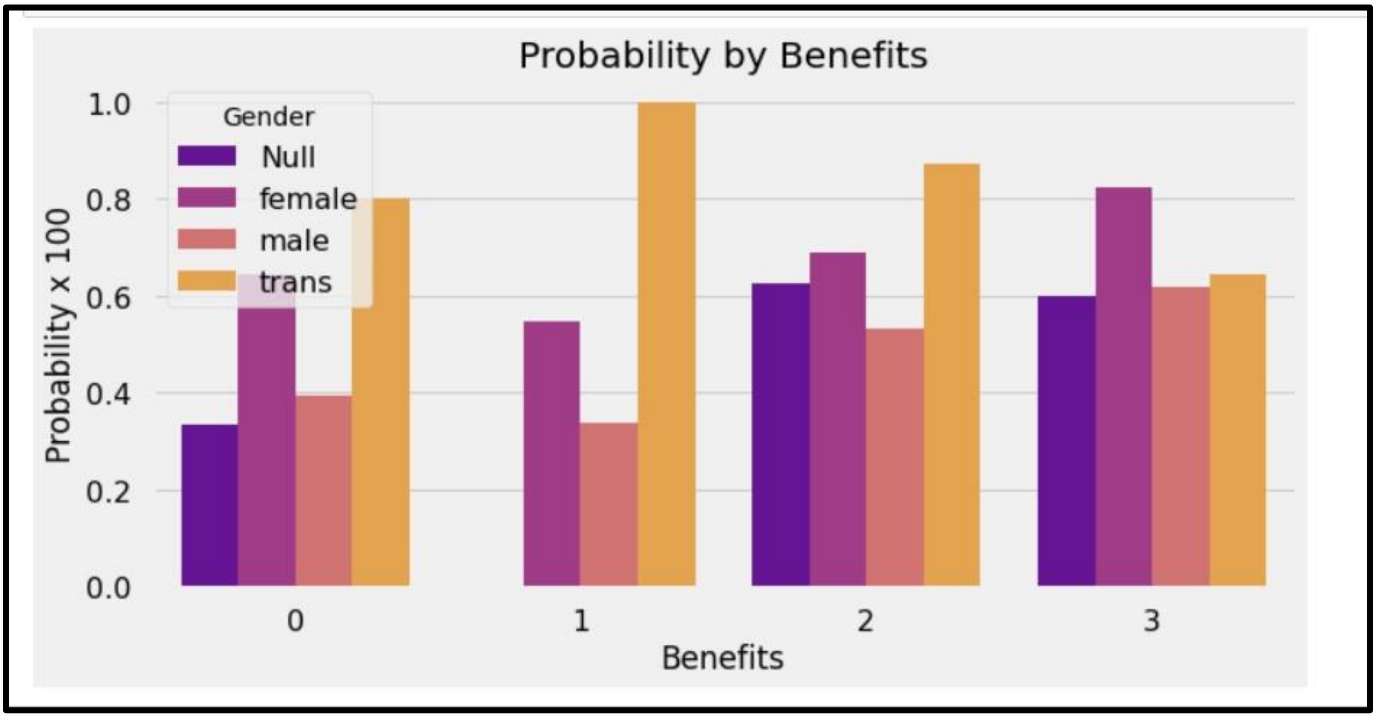


Figure 06 Probability of Benefits

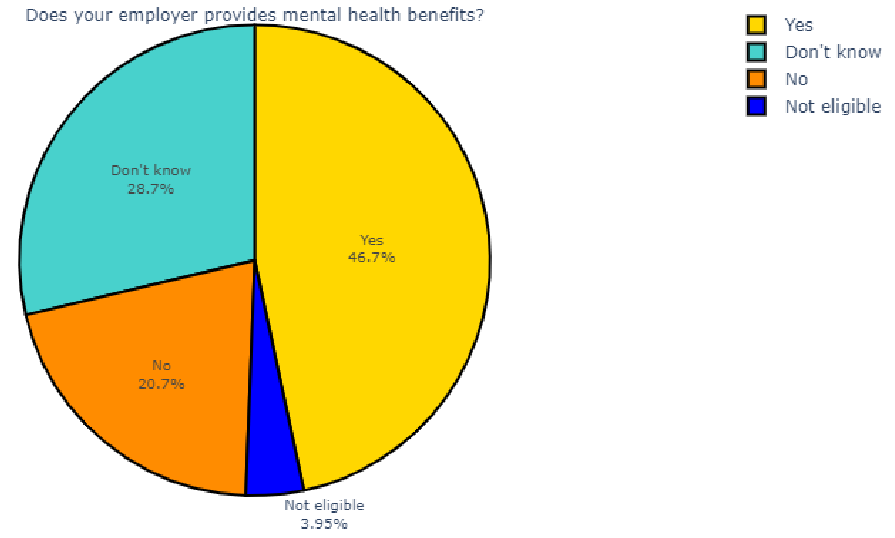


Figure 07 Provide Benefits

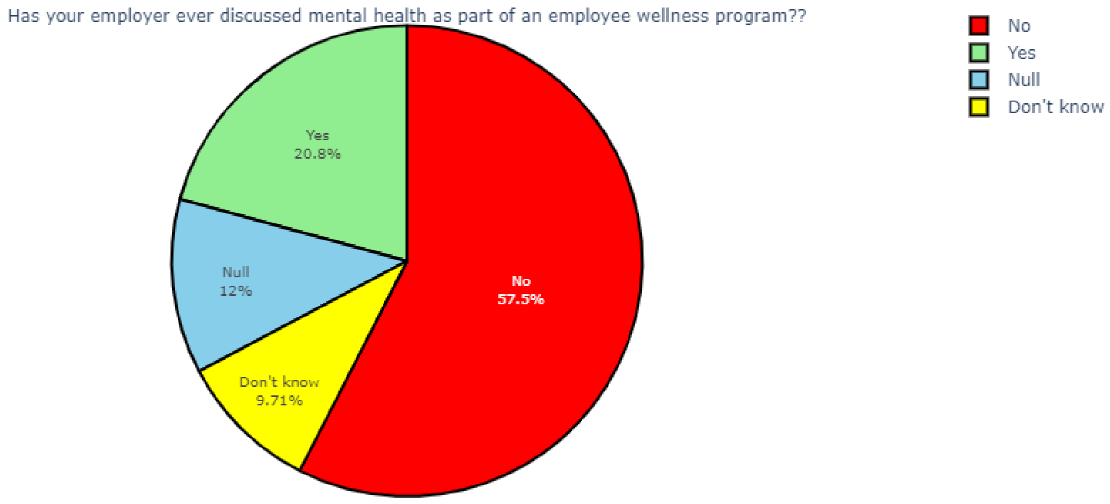
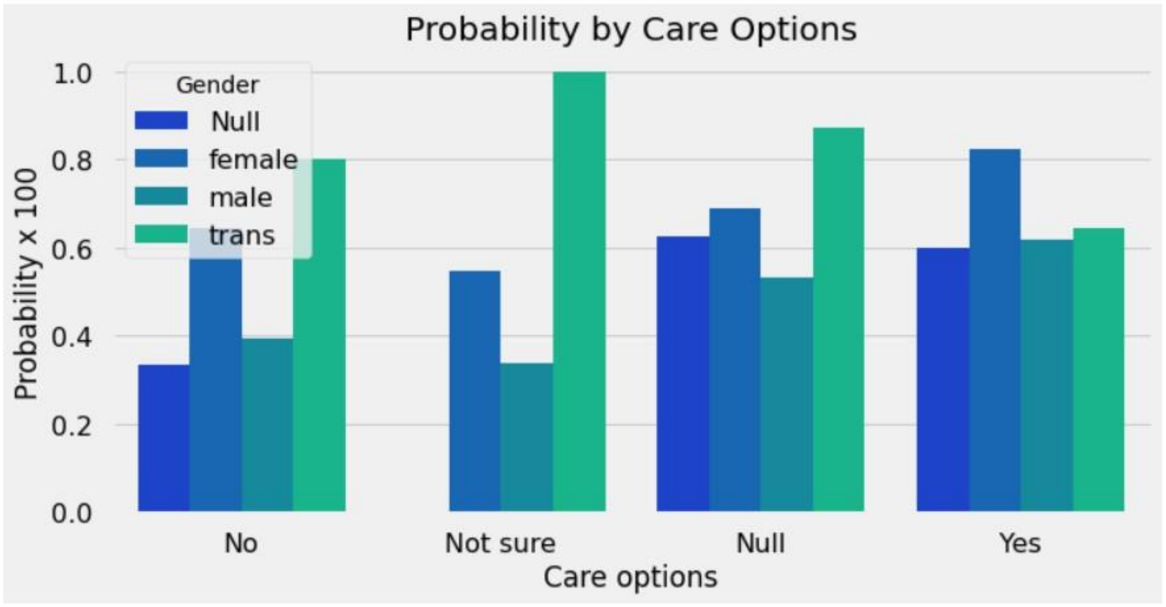


Figure 08 Mental Wellness Program



There seems to be a stark contrast among transgenders and others in availability and awareness of care options in workplace.

Figure 09 Care Options Provides

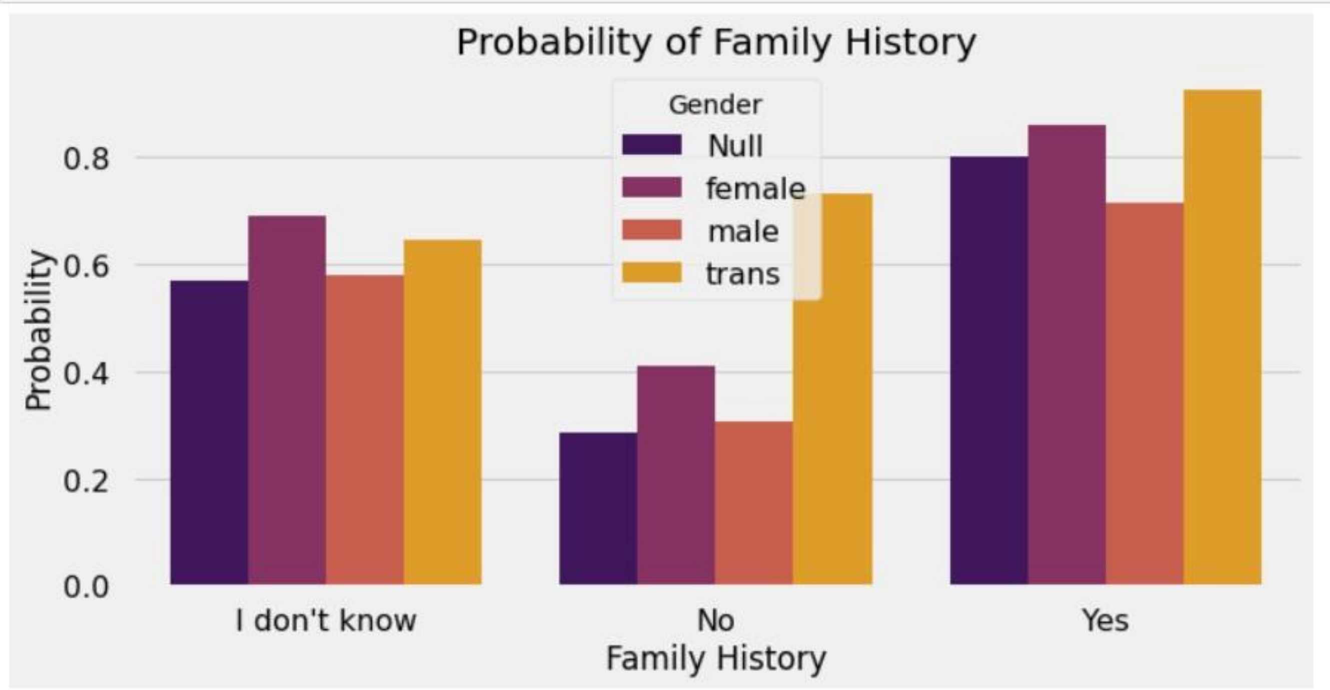


Figure 10 Family History

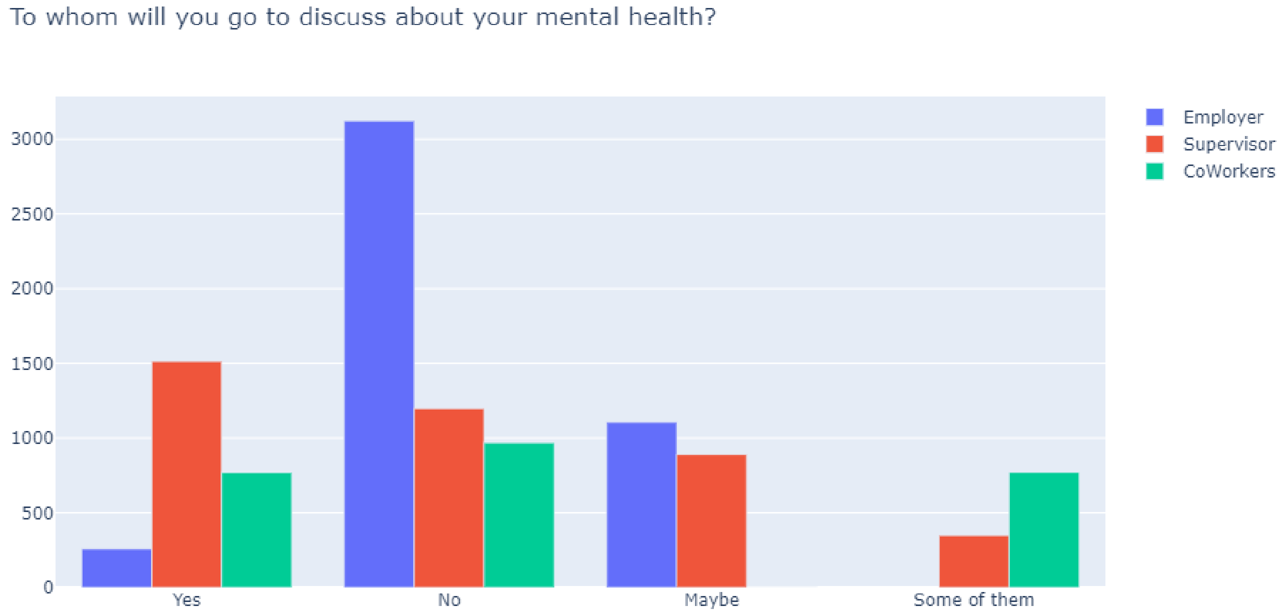


Figure 11 Discuss About Mental Health

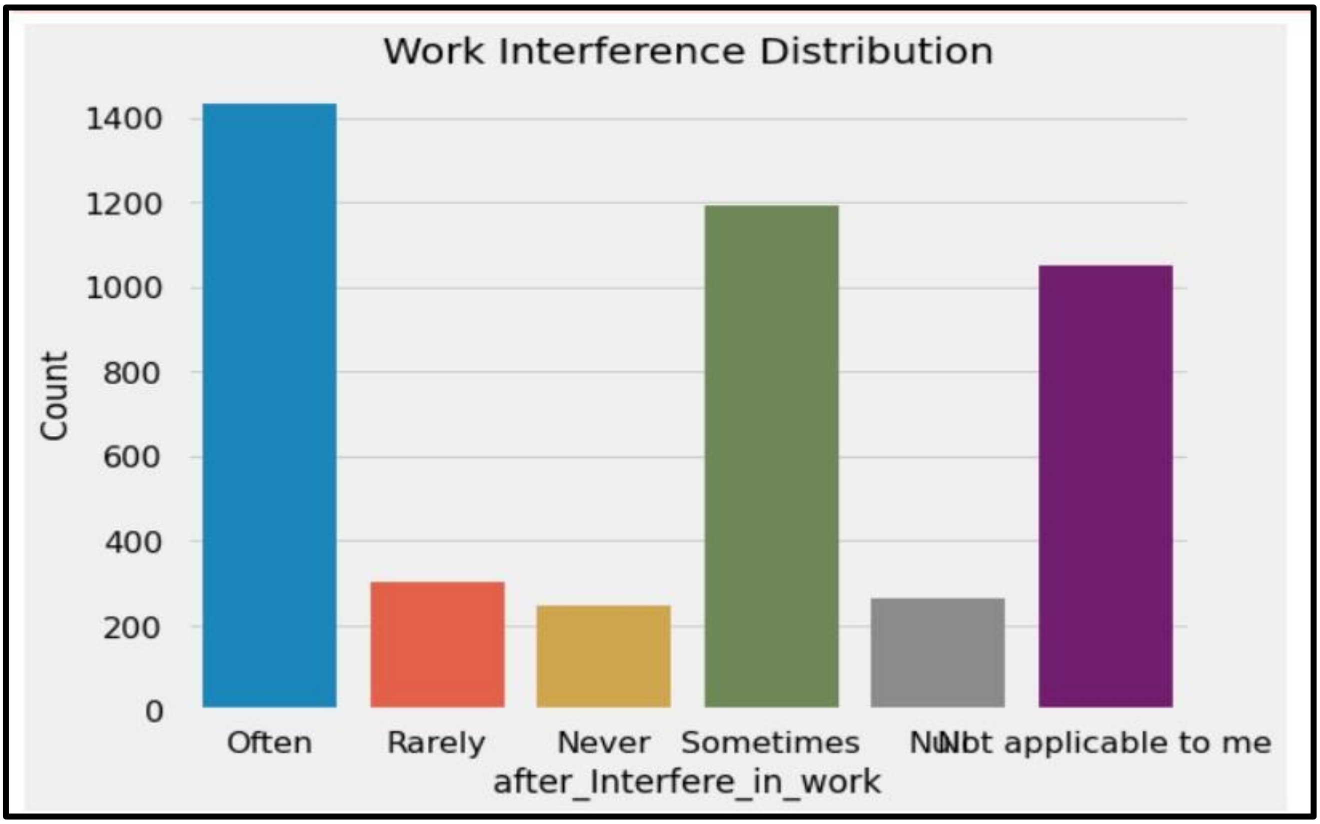


Figure 12 Work Interference

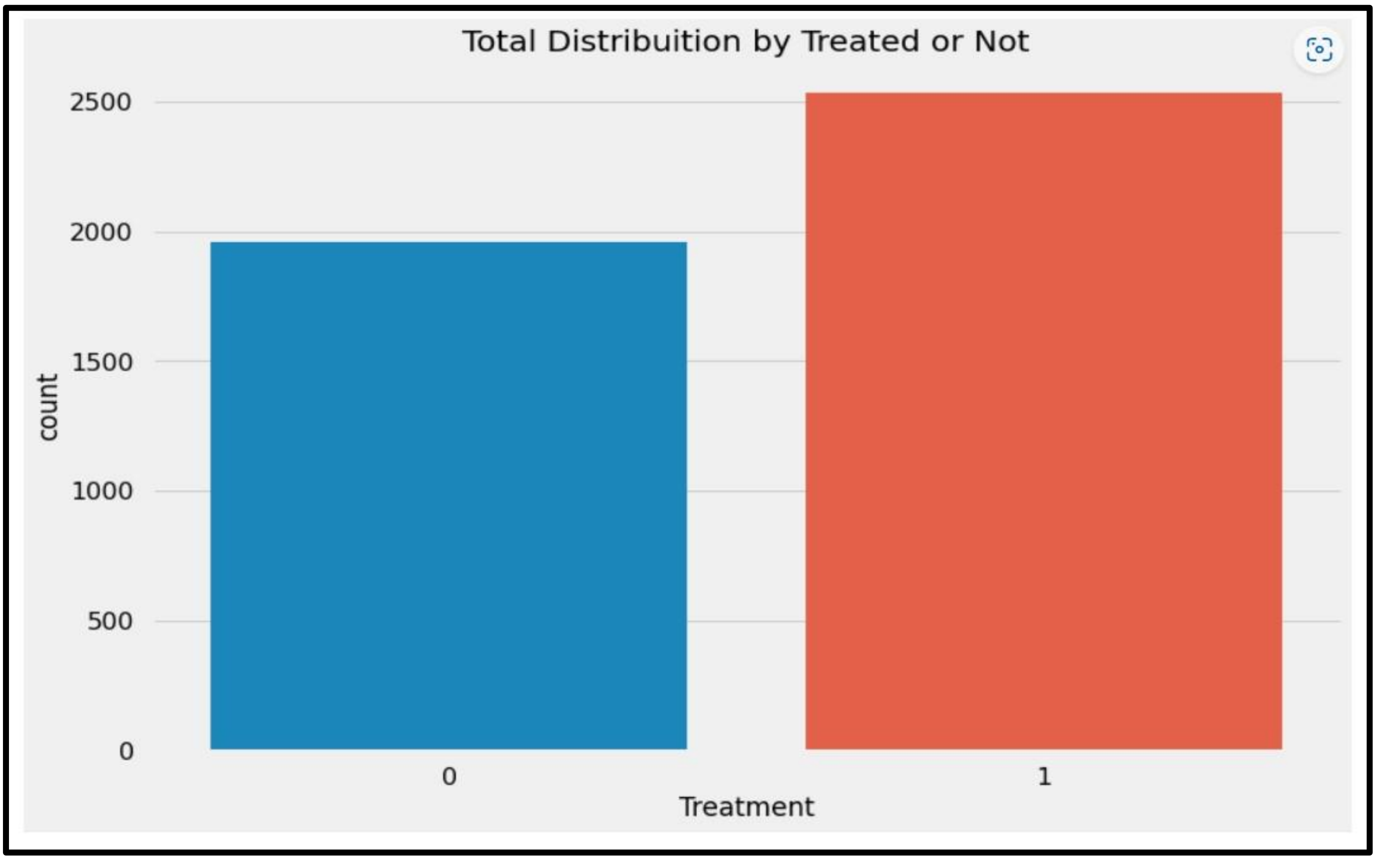
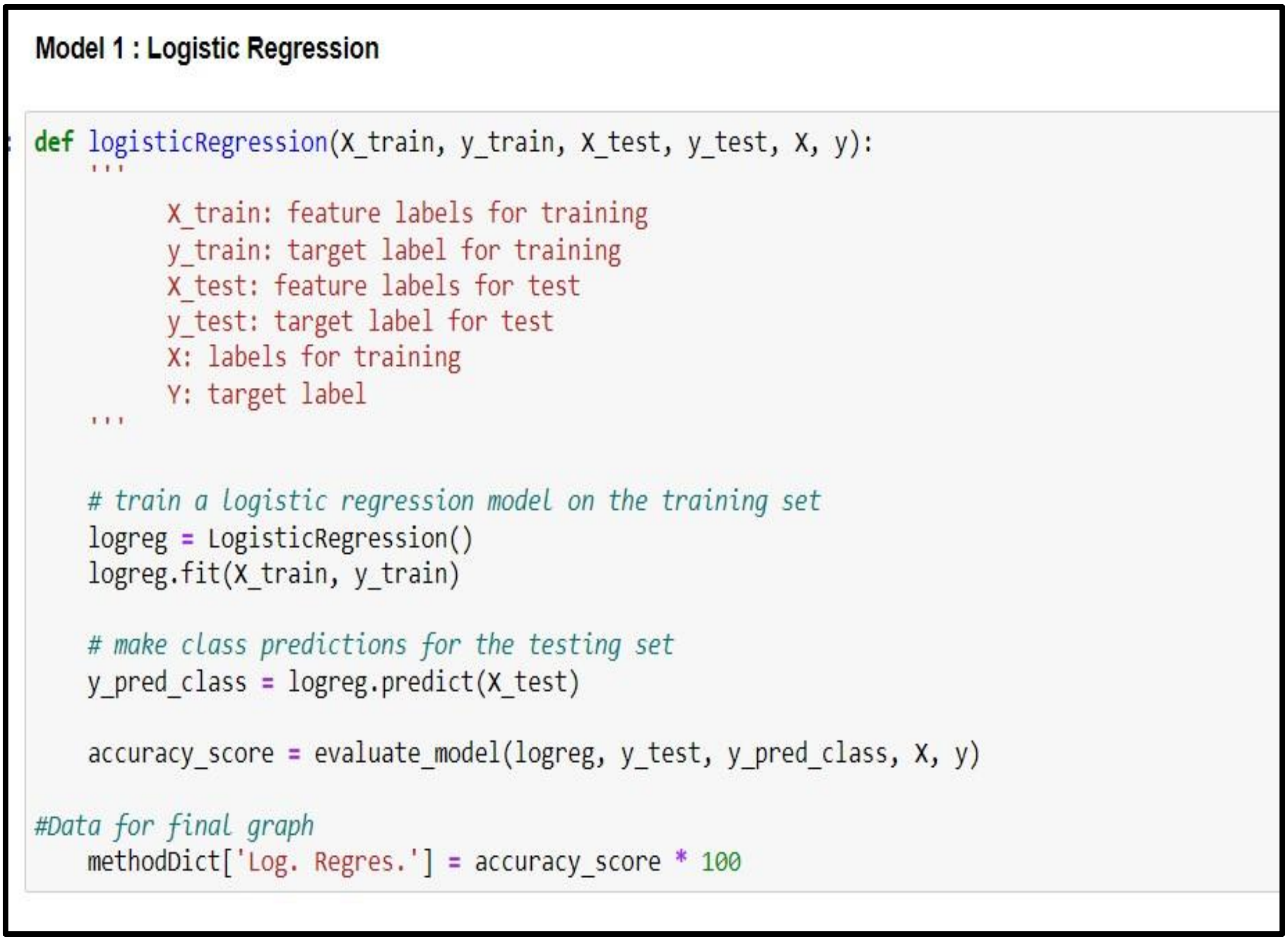


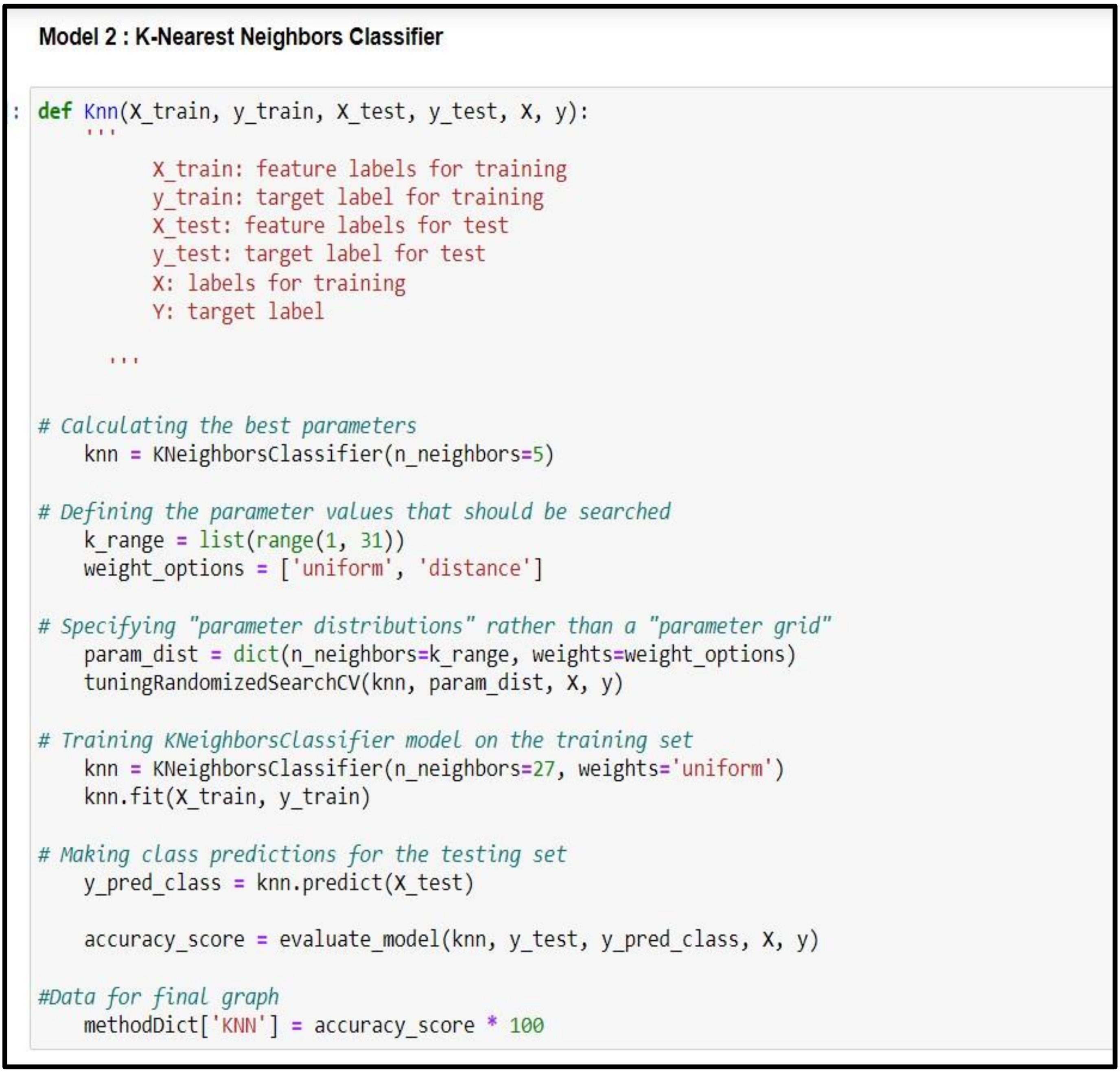
Figure 13 Treatment

09. Machine Learning Modelling



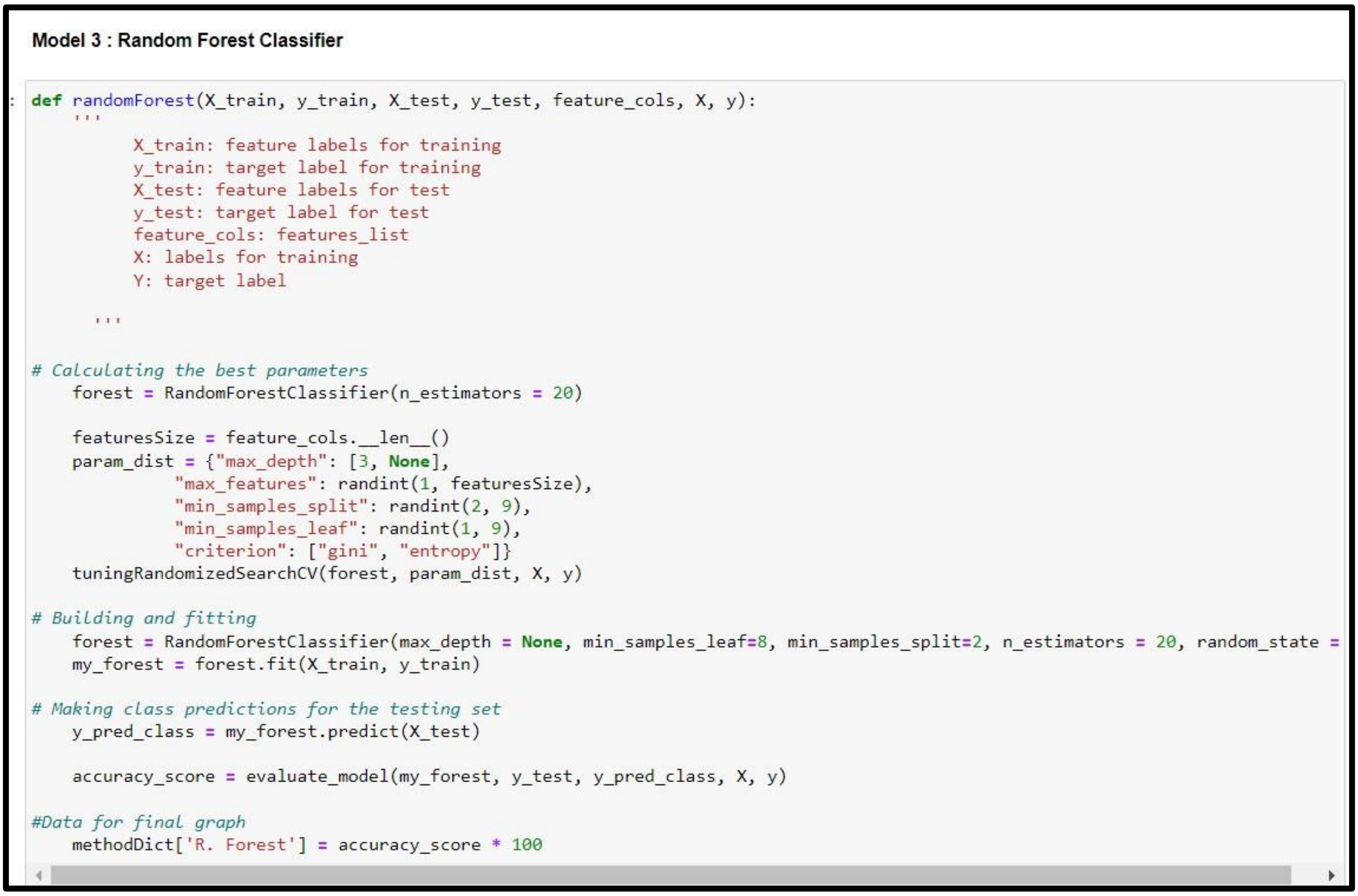
Model 1 : Logistic Regression

```
def logisticRegression(X_train, y_train, X_test, y_test, X, y):
    '''
          X_train: feature labels for training
          y_train: target label for training
          X_test: feature labels for test
          y_test: target label for test
          X: labels for training
          Y: target label
    '''

    # train a logistic regression model on the training set
    logreg = LogisticRegression()
    logreg.fit(X_train, y_train)

    # make class predictions for the testing set
    y_pred_class = logreg.predict(X_test)

    accuracy_score = evaluate_model(logreg, y_test, y_pred_class, X, y)

#Data for final graph
    methodDict['Log. Regres.'] = accuracy_score * 100
```



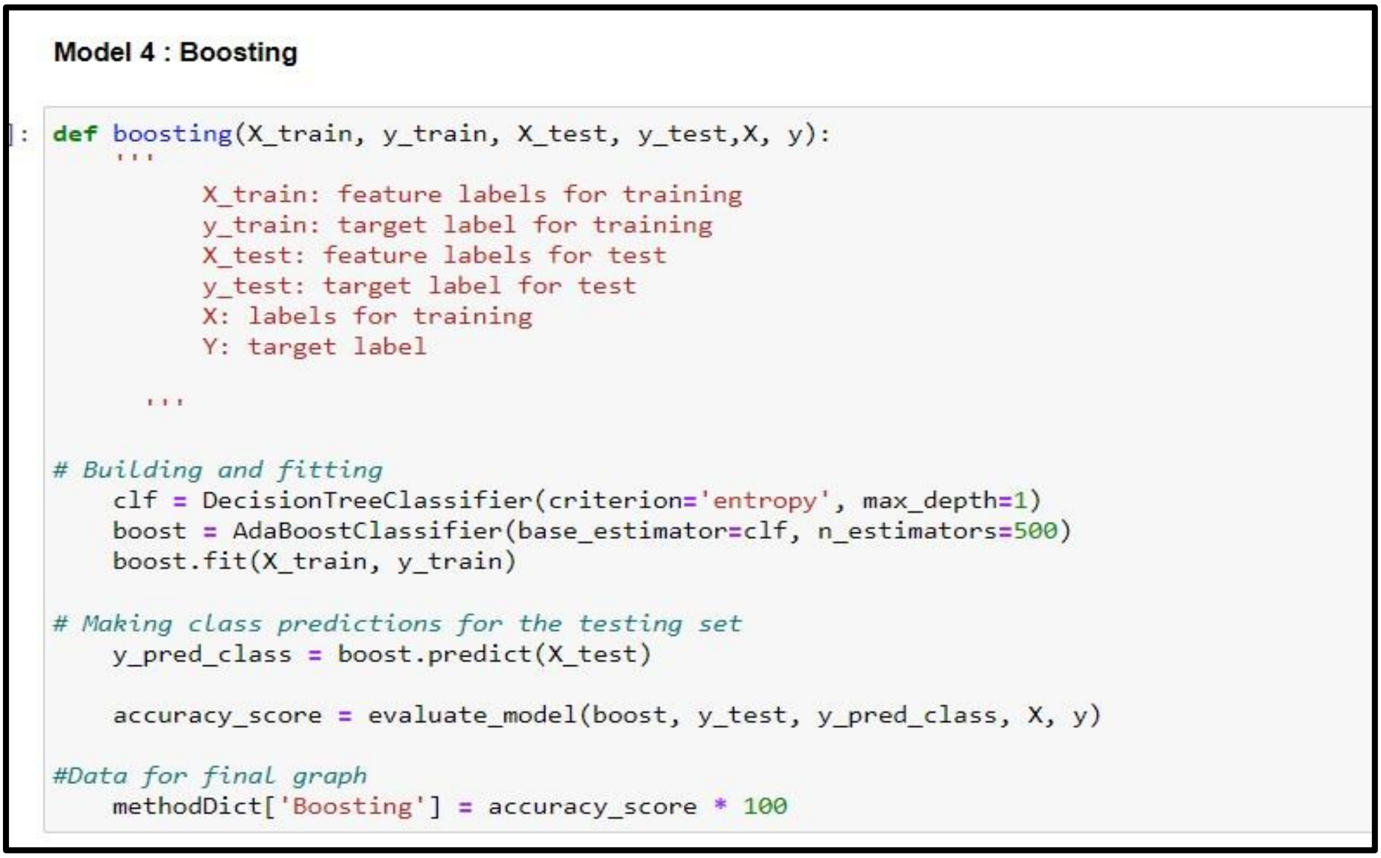
Model 2 : K-Nearest Neighbors Classifier

```
def Knn(X_train, y_train, X_test, y_test, X, y):
    '''
          X_train: feature labels for training
          y_train: target label for training
          X_test: feature labels for test
          y_test: target label for test
          X: labels for training
          Y: target label

      '''

# Calculating the best parameters
    knn = KNeighborsClassifier(n_neighbors=5)

# Defining the parameter values that should be searched
    k_range = list(range(1, 31))
    weight_options = ['uniform', 'distance']

# Specifying "parameter distributions" rather than a "parameter grid"
    param_dist = dict(n_neighbors=k_range, weights=weight_options)
    tuningRandomizedSearchCV(knn, param_dist, X, y)

# Training KNeighborsClassifier model on the training set
    knn = KNeighborsClassifier(n_neighbors=27, weights='uniform')
    knn.fit(X_train, y_train)

# Making class predictions for the testing set
    y_pred_class = knn.predict(X_test)

    accuracy_score = evaluate_model(knn, y_test, y_pred_class, X, y)

#Data for final graph
    methodDict['KNN'] = accuracy_score * 100
```



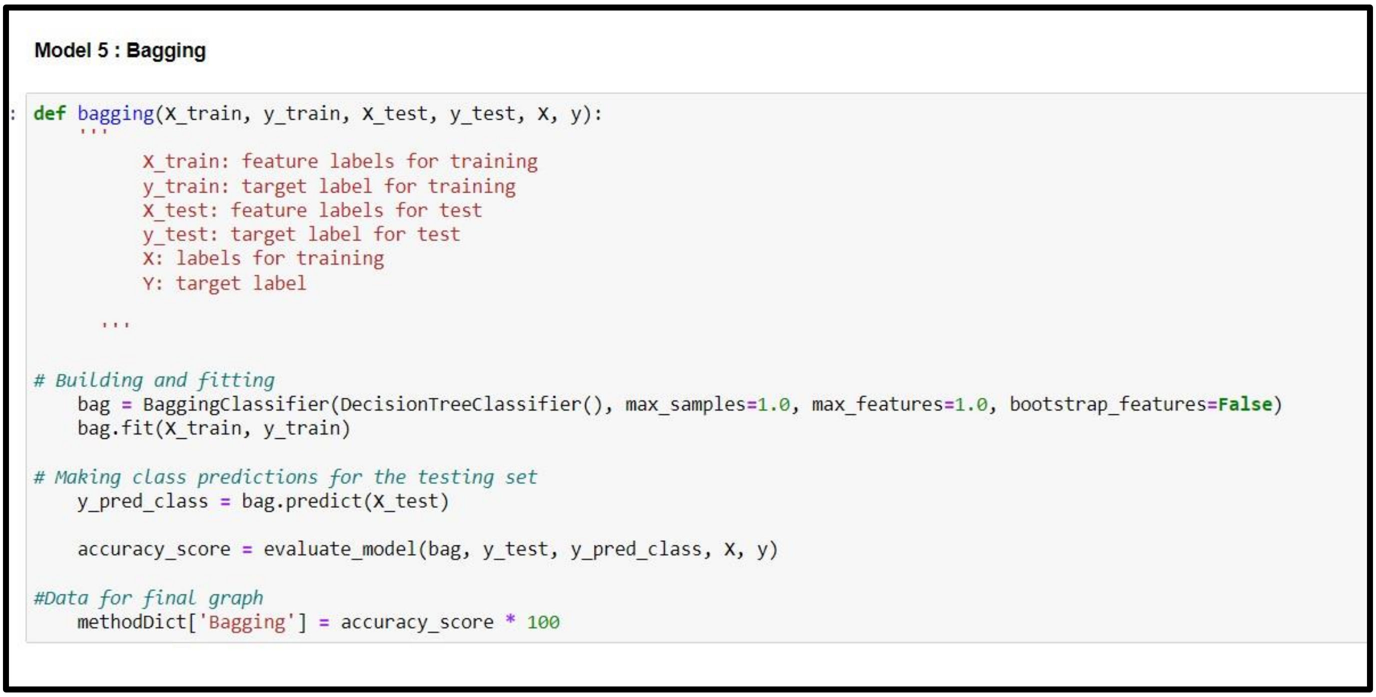
Model 3 : Random Forest Classifier

```
def randomForest(X_train, y_train, X_test, y_test, feature_cols, X, y):
    '''
          X_train: feature labels for training
          y_train: target label for training
          X_test: feature labels for test
          y_test: target label for test
          feature_cols: features_list
          X: labels for training
          Y: target label

      '''

# Calculating the best parameters
    forest = RandomForestClassifier(n_estimators = 20)

    featuresSize = feature_cols.__len__()
    param_dist = {"max_depth": [3, None],
              "max_features": randint(1, featuresSize),
              "min_samples_split": randint(2, 9),
              "min_samples_leaf": randint(1, 9),
              "criterion": ["gini", "entropy"]}
    tuningRandomizedSearchCV(forest, param_dist, X, y)

# Building and fitting
    forest = RandomForestClassifier(max_depth = None, min_samples_leaf=8, min_samples_split=2, n_estimators = 20, random_state =
    my_forest = forest.fit(X_train, y_train)

# Making class predictions for the testing set
    y_pred_class = my_forest.predict(X_test)

    accuracy_score = evaluate_model(my_forest, y_test, y_pred_class, X, y)

#Data for final graph
    methodDict['R. Forest'] = accuracy_score * 100
```



Model 4 : Boosting

```
def boosting(X_train, y_train, X_test, y_test,X, y):
    '''
          X_train: feature labels for training
          y_train: target label for training
          X_test: feature labels for test
          y_test: target label for test
          X: labels for training
          Y: target label

      '''

# Building and fitting
    clf = DecisionTreeClassifier(criterion='entropy', max_depth=1)
    boost = AdaBoostClassifier(base_estimator=clf, n_estimators=500)
    boost.fit(X_train, y_train)

# Making class predictions for the testing set
    y_pred_class = boost.predict(X_test)

    accuracy_score = evaluate_model(boost, y_test, y_pred_class, X, y)

#Data for final graph
    methodDict['Boosting'] = accuracy_score * 100
```



Model 5 : Bagging

```
def bagging(X_train, y_train, X_test, y_test, X, y):
    '''
            X_train: feature labels for training
            y_train: target label for training
            X_test: feature labels for test
            y_test: target label for test
            X: labels for training
            Y: target label

        '''

# Building and fitting
    bag = BaggingClassifier(DecisionTreeClassifier(), max_samples=1.0, max_features=1.0, bootstrap_features=False)
    bag.fit(X_train, y_train)

# Making class predictions for the testing set
    y_pred_class = bag.predict(X_test)

    accuracy_score = evaluate_model(bag, y_test, y_pred_class, X, y)

#Data for final graph
    methodDict['Bagging'] = accuracy_score * 100
```

Prediction Result:

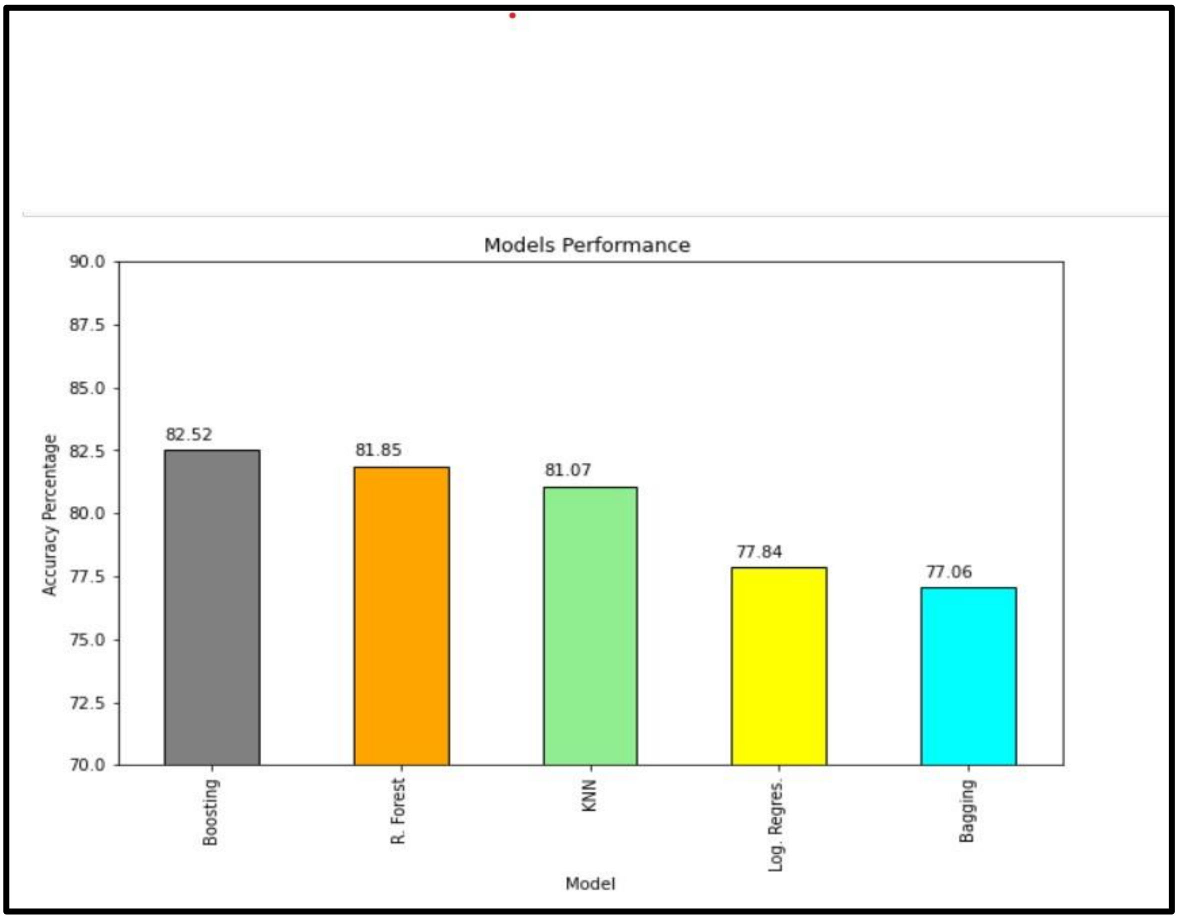


Figure 14 Predictions Result

10. Conclusion and Future Scope

Only a few decades ago, it was rare for LGBTQ people to disclose their sexual orientation in the workplace due to the fear of getting fired. These days it is somewhat common for LGBTQ people to be out in the workplace. Since there is a less negative outlook of people of their lifestyle or opposite-sex partners. Mental illness may be different than sexual orientation, but the idea of having the freedom to be open about all aspects of who you are — and to do so in all parts of your life — is the same. It's time for everyone to have that freedom, and the path to it starts in the workplace.

The analysis is done in the project that although mental health issues are on the rise due to increased stress in the industry. Still, employers are far behind in creating a workplace environment that is suitable for the employees facing these issues. An open culture where mental health is considered as important as physical health is the need of the hour. Also, LGBTQ people require more empowerment in the society which in turn will boost their mental health.

Many different techniques and algorithms had been introduced and proposed to test and solve the mental health problems. There are still many solutions that can be refined. In addition, there are still many problems to be discovered and tested using a wide variety of settings in machine learning for the mental health domain. As classifying the mental health data is generally a very challenging problem, the features used in the machine learning algorithms will significantly affect the performance of the classification. The existing studies and research show that machine learning can be a useful tool in helping understand psychiatric disorders. Besides that, it may also help distinguish and classify the mental health problems among patients for further treatment.

Newer approaches that use data that arise from the integration of various sensor modalities present in technologically advanced devices have proven to be a convenient resource to recognize the mood state and responses from patients among others. It is noticeable that most of the research and studies are still struggling to validate the results because of insufficiency of acceptable validated evidence, especially from the external sources. Besides that, most of the machine learning might not have the same performance across all the problems.

The performance of the machine learning models will vary depending on the data samples obtained and the features of the data. Moreover, machine learning models can also be affected by pre-processing activities such as data cleaning and parameter tuning in order to achieve optimal results. Hence, it is very important for researchers to investigate and analyse the data with various machine learning algorithms to choose the highest accuracy among the machine learning algorithms

References

- Sharon Leighton and Nisha Dogra. Defining mental health and mental illness Vol 01, Jan-2009.

- Mental Health Continuum. Beyond blue. Website: https://beyou.edu.au/resources/mental-health-continuum. Accessed September 1, 2020.

- National Institute of Mental Health. Any mental illness https:www.nimh.nih.gov/health/statistics/mental-illness.shtml#part_154785 Accessed September 1, 2020.

- World Health Organization. Promoting mental health: concepts, emerging evidence, practice (Summary Report) Geneva: World Health Organization; 2004.

Websites (Data References):

- https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey
- https://osmhhelp.org/research