PREDICTIVE ANALYTICS OF MENTAL HEALTH DISORDERS

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

By

KANDI BALA SAI NITHIN	**2103A51049**
BINGI ROHAN	**2103A51082**
PENDYALA AASHRITH	**2103A51207**

Under the guidance of

Mr. S NARESH KUMAR

Assistant Professor, Department of CSE.

Department of Computer Science and Artificial Intelligence

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **"PREDICTIVE ANALYTICS OF MENTAL HEALTH DISORDERS"** is the bona fide work carried out by **KANDI BALA SAI NITHIN, BINGI ROHAN, PENDYALA AASHRITH** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

Mr. S Naresh Kumar
Asst. Professor,
S R University,
Ananthasagar, Warangal.

Dr. M. Sheshikala
Assoc. Prof. & HOD (CSE),
S R University,
Ananthasagar, Warangal.

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr. S Naresh Kumar, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M. Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

Mental Health Disorders are of a great concern. But we can't know very much about these as they are not physically perceptible. The busy lives of today's world increased cases of anxiety, stress, etc. All these lead to increased mental health issues like anxiety disorder, depression, bipolar disorder, etc. This is a global problem. In this project, we are going to analyze a lot of data about different mental health disorders. By this, we are willing to bring more awareness on this global problem and contribute in improving mental health of the people. By this project, we can answer questions like:

What are mental health disorder types?

How many people suffer these disorder types?

How many people in your country have depression?

What percentage?

Which age group has more depression?

Which work group has more depression?

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

This project involves using machine learning to analyze data related to mental health disorders, with the aim of identifying patterns and predicting the risk of developing mental health disorders in individuals. The project requires an understanding of mental health disorders and their risk factors, as well as the use of machine learning techniques.

The first step in undertaking a mental health disorder ML project is to gather reliable data from sources such as clinical studies, surveys, or patient records. This data needs to be cleaned, processed, and analyzed using appropriate machine learning algorithms to identify patterns and relationships between different variables. The machine learning algorithms used in the project can help in identifying high-risk individuals and developing effective prevention and intervention strategies. For instance, regression algorithms can be used to predict the likelihood of developing mental health disorders based on various independent variables.

The project's results can also contribute to a better understanding of the risk factors associated with mental health disorders, thereby informing the development of effective preventive and intervention strategies.

In summary, this mental health disorder ML project involves using machine learning algorithms to analyze data related to mental health disorders, with the aim of identifying patterns, predicting the risk of developing mental health disorders, and developing effective prevention and intervention strategies. Undertaking such a project can provide valuable experience in data analysis, problem-solving, and interdisciplinary collaboration while contributing to the development of effective solutions for a significant public health concern.

1.2 PROBLEM STATEMENT

Mental health disorders have become a significant public health concern, with the World Health Organization estimating that around 1 in 4 people in the world will be affected by mental or neurological disorders at some point in their lives. The prevalence of mental health disorders is influenced by a complex interplay of factors, including genetic, environmental, and social determinants.

Identifying the risk factors associated with mental health disorders and predicting the likelihood of developing these disorders can aid in early intervention and prevention. Machine learning techniques can play a crucial role in analyzing vast amounts of data and identifying patterns that may not be evident through traditional statistical methods. The goal of this project is to develop a machine learning model that can accurately predict the risk of developing specific mental health disorders based on a set of independent variables such as demographic information, medical history, and other relevant factors. The project aims to contribute to the development of effective interventions and prevention strategies for mental health disorders.

1.3 EXISTING SYSTEM

There are many systems to predict the risk of mental health disorder. Most of the systems rely solely on medical diagnosis which isn't accessible to everyone. The data is mostly private and secured. So, it cannot be used for spreading awareness. The systems also don't have very good model, as they may predict wrong values sometimes.

1.4 PROPOSED SYSTEM

In this project, we have built a very good model based on the dataset with correct partition of data for training and testing. This system demonstrates the relationship between different mental health disorders. It is very useful to spread awareness on Mental Health Disorders. Our project is a regression model which fits the data and predicts the values accurately.

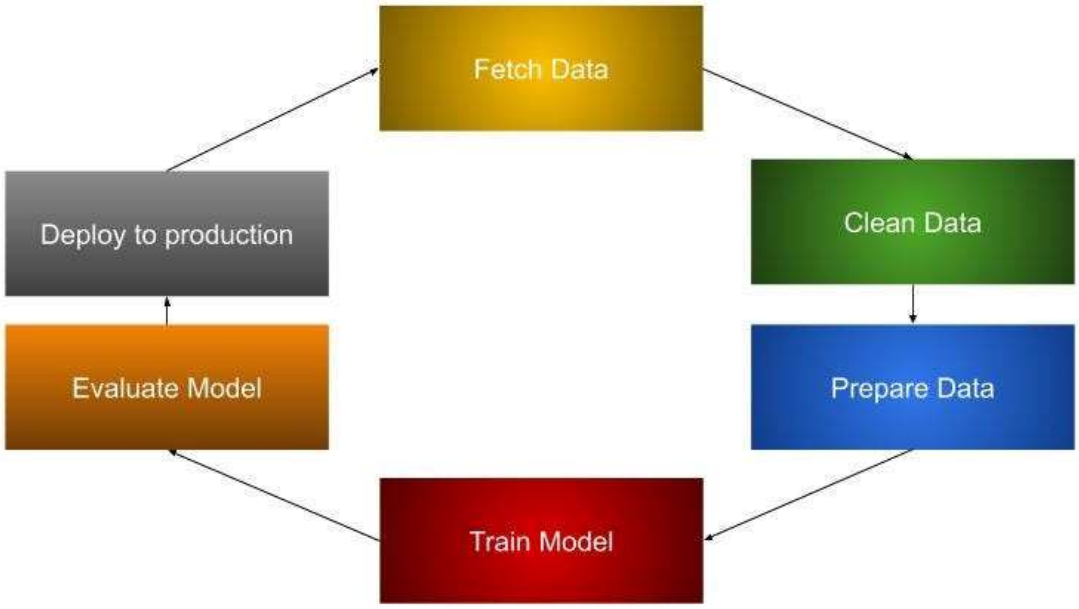
1.5 OBJECTIVES

Objectives for a mental health disorder ML

1. Develop an ML model that can accurately predict the likelihood of an individual developing a specific mental health disorder based on their demographic information, medical history, and other relevant factors.
2. Identify new risk factors for mental health disorders by analyzing large sets of data from various sources, including clinical studies, surveys, and patient records.
3. Spread awareness on Mental Health Disorders, which aids in prevention of those disorders.

1.6 ARCHITECTURE

Just like every other machine learning system, we have also followed the basic steps to implement the model.



CHAPTER 2

LITERATURE SURVEY

2.1 SURVEY SUMMARY

Mental health is a neglected area in health care in Ghana. With few clinicians and trained researchers in the field, research has been limited both in quantity and quality. A search of the available literature revealed 98 articles published between 1955 and 2009. Sixty-six are reviewed in this paper. Topics covered included hospital and community-based prevalence studies, psychosis, depression, substance misuse, self-harm, and help-seeking. Much of the research was small in scale and thus largely speculative in its conclusions. Epidemiological data is scarce and unreliable and no large-scale studies have been published. There are very few studies of clinical practice in mental health. The existing literature suggests several important areas for future research to inform the development of targeted and effective interventions in mental health care in Ghana.

Mental health is a state of wellbeing where people and societies function at their best. That is, where people can cope with the stresses of everyday life, work productively, and contribute to their communities.

Mental health conditions and neurological disorders are conditions that affect thoughts, emotions, behaviours, and relationships. These include disorders that cause a high burden of disease such as depression, bipolar affective disorder, schizophrenia, anxiety disorders, dementia, substance use disorders, among many others. These conditions can be experienced in isolation; however, they often occur alongside other noncommunicable diseases such as cardiovascular disease, diabetes, respiratory diseases and cancers.

They also share many NCD risk factors, such as tobacco use, alcohol use, unhealthy diets and physical inactivity.

Mental health and wellbeing are central to reducing the global burden of NCDs and were included as part of the 5x5 approach to tackling NCDs by the World Health Organization in 2018.

Depression

Depression is when a person experiences depressed mood (feeling sad, irritable, empty) or a loss of pleasure or interest in activities for more than two weeks. Depression can severely impact a person's ability to function and interact with people and society. It is estimated that around 264 million people are affected by depression, or around 5% of adults.

Women are more affected by depression than men. Worldwide about 10% of pregnant women and 13% of women who have just given birth experience a mental disorder, primarily depression. In developing countries, this is even higher – 15.6% during pregnancy and 19.8% after.

Depression can impede personal development, health, education, and employment. Severe cases of depression can lead to suicide. 75% of suicides occur in low- and middle-income countries, but effective strategies such as early detection, treatment and ongoing support mean suicide can be prevented.

CHAPTER 3

DATA PRE-PROCESSING

3.1 DATASET DESCRIPTION

Our dataset has been collected from DataWorld website which has many datasets for machine learning implementation.

Our dataset has the following columns:

Year: It shows the year in which the data is collected.

Country: It shows the country to which the data belongs to.

Schizophrenia %: It shows the percentage of people having Schizophrenia.

Bipolar Disorder %: It shows the percentage of people having Bipolar Disorder.

Eating Disorders %: It shows the percentage of people having Eating Disorders.

Anxiety Disorders %: It shows the percentage of people having Anxiety Disorders.

Drug Use Disorder %: It shows the percentage of people having Drug Use Disorder.

Depression %: It shows the percentage of people having Depression.

Alcohol Use Disorders %: It shows the percentage of people having Alcohol Use Disorder.

Our dataset also has other sheets which have data on depression by level of education and depression by age.

	Entity	Code	Year	Schizophrenia (%)	Bipolar disorder (%)	Eating disorders (%)	Anxiety disorders (%)	Drug use disorders (%)	Depression (%)	Alcohol use disorders (%)
0	Afghanistan	AFG	1990	0.160560	0.697779	0.101855	4.828830	1.677082	4.071831	0.672404
1	Afghanistan	AFG	1991	0.160312	0.697961	0.099313	4.829740	1.684746	4.079531	0.671768
2	Afghanistan	AFG	1992	0.160135	0.698107	0.096692	4.831108	1.694334	4.088358	0.670644
3	Afghanistan	AFG	1993	0.160037	0.698257	0.094336	4.830864	1.705320	4.096190	0.669738
4	Afghanistan	AFG	1994	0.160022	0.698469	0.092439	4.829423	1.716069	4.099582	0.669260
...	...	...	...	...	...	...	...	...	...	...
6463	Zimbabwe	ZWE	2013	0.155670	0.607993	0.117248	3.090168	0.766280	3.128192	1.515641
6464	Zimbabwe	ZWE	2014	0.155993	0.608610	0.118073	3.093964	0.768914	3.140290	1.515470
6465	Zimbabwe	ZWE	2015	0.156465	0.609363	0.119470	3.098687	0.771802	3.155710	1.514751
6466	Zimbabwe	ZWE	2016	0.157111	0.610234	0.121456	3.104294	0.772275	3.174134	1.513269
6467	Zimbabwe	ZWE	2017	0.157963	0.611242	0.124443	3.110926	0.772648	3.192789	1.510943

6468 rows × 10 columns

3.2 DATA CLEANING

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

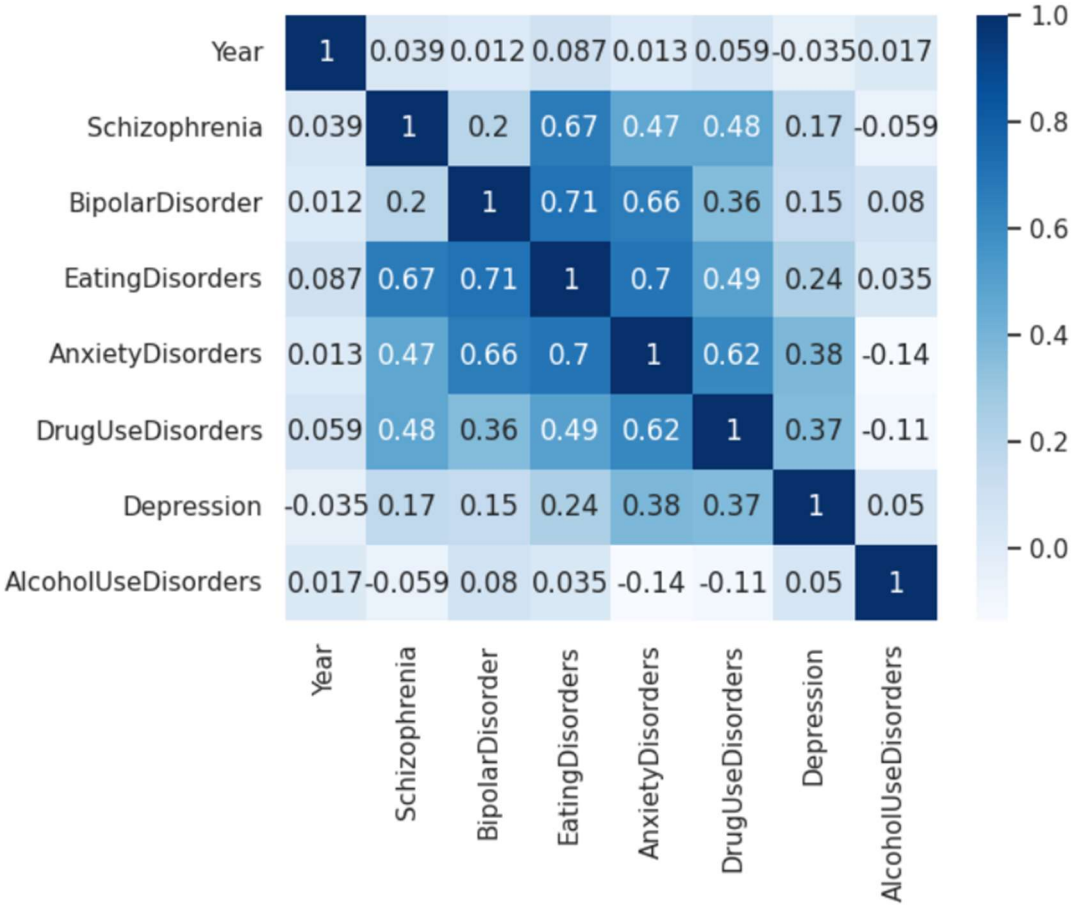
```
disorders.isna().sum()

    Entity                0
    Year                  0
    Schizophrenia         0
    BipolarDisorder       0
    EatingDisorders       0
    AnxietyDisorders      0
    DrugUseDisorders      0
    Depression            0
    AlcoholUseDisorders   0
    dtype: int64


disorders.isna().all()

    Entity                False
    Year                  False
    Schizophrenia         False
    BipolarDisorder       False
    EatingDisorders       False
    AnxietyDisorders      False
    DrugUseDisorders      False
    Depression            False
    AlcoholUseDisorders   False
    dtype: bool
```

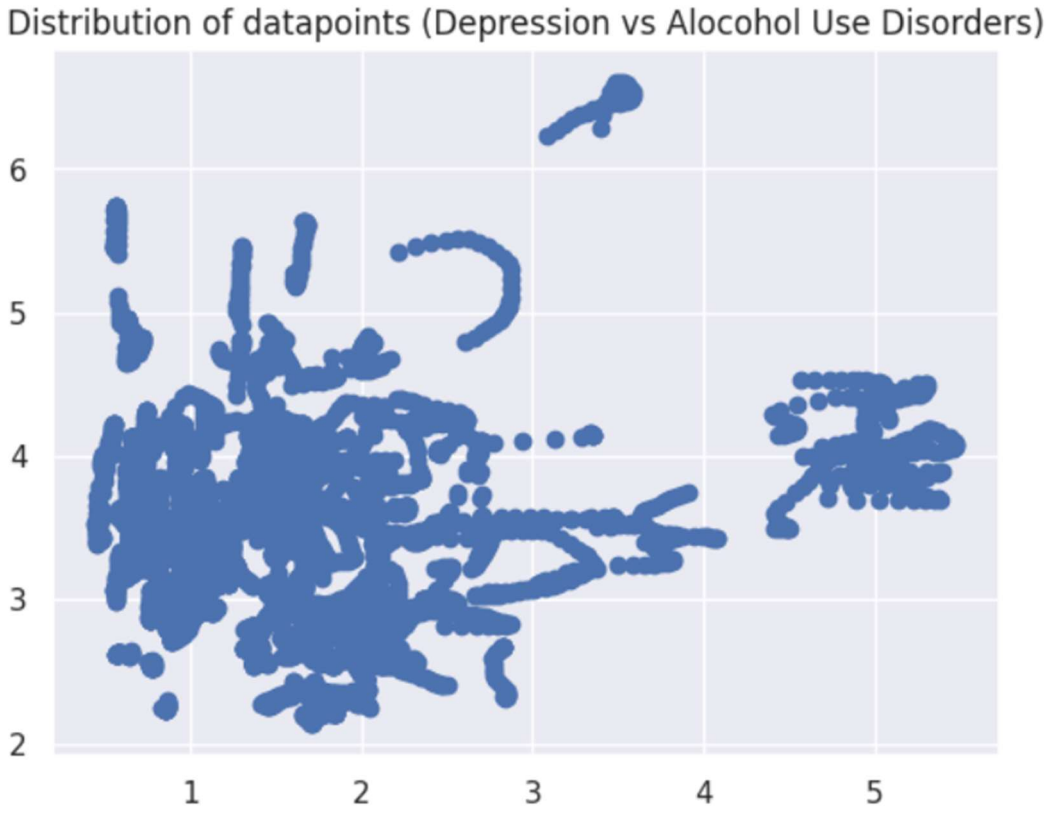
```
disorders.duplicated()

0       False
1       False
2       False
3       False
4       False
        ...
6463    False
6464    False
6465    False
6466    False
6467    False
Length: 6468, dtype: bool
```

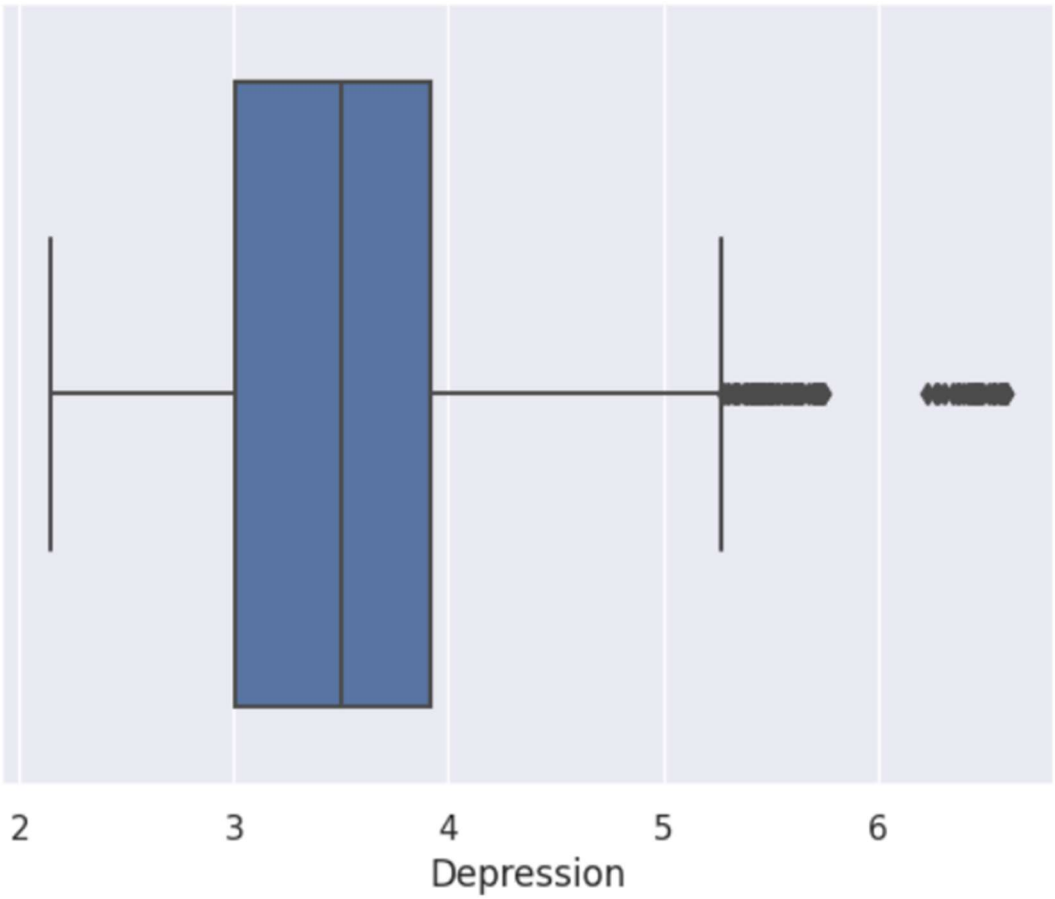
Our dataset has no NULL values or duplicates. So, there is no need to fill the null values or deal with the duplicates.

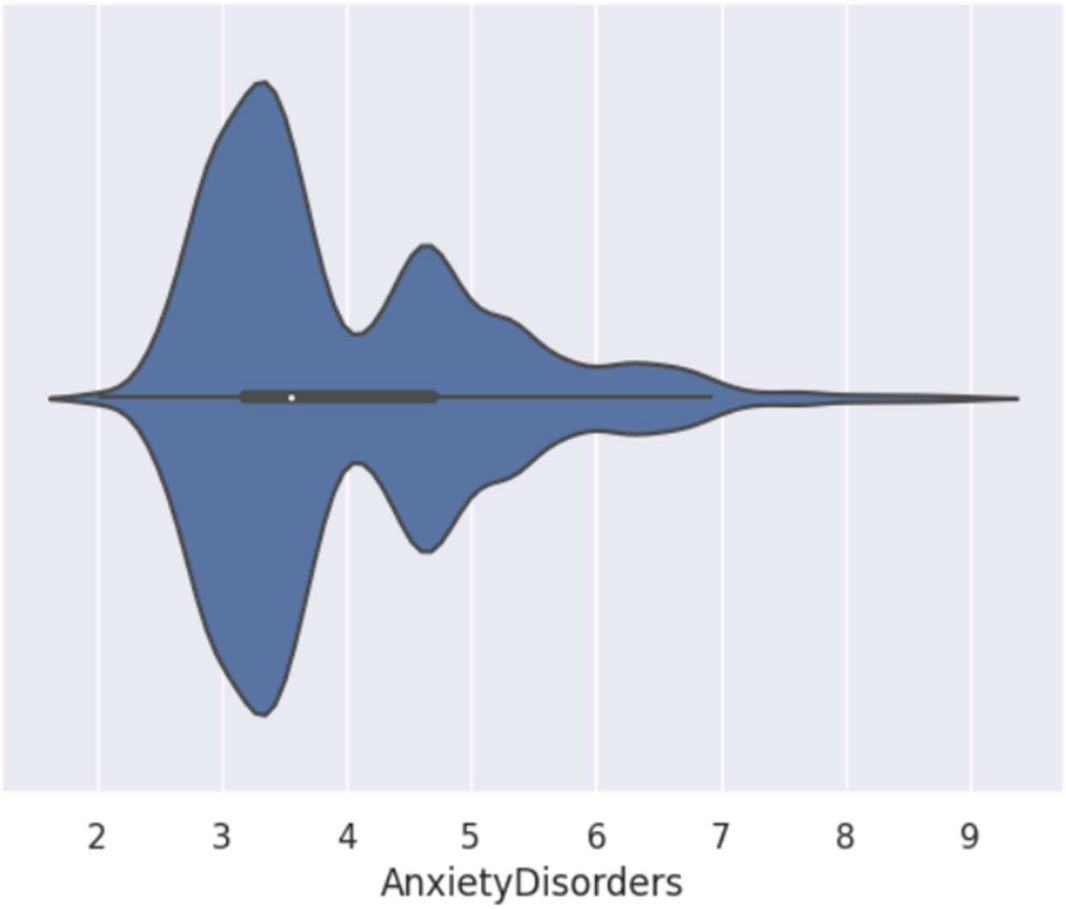
3.3 DATA VISUALIZATION

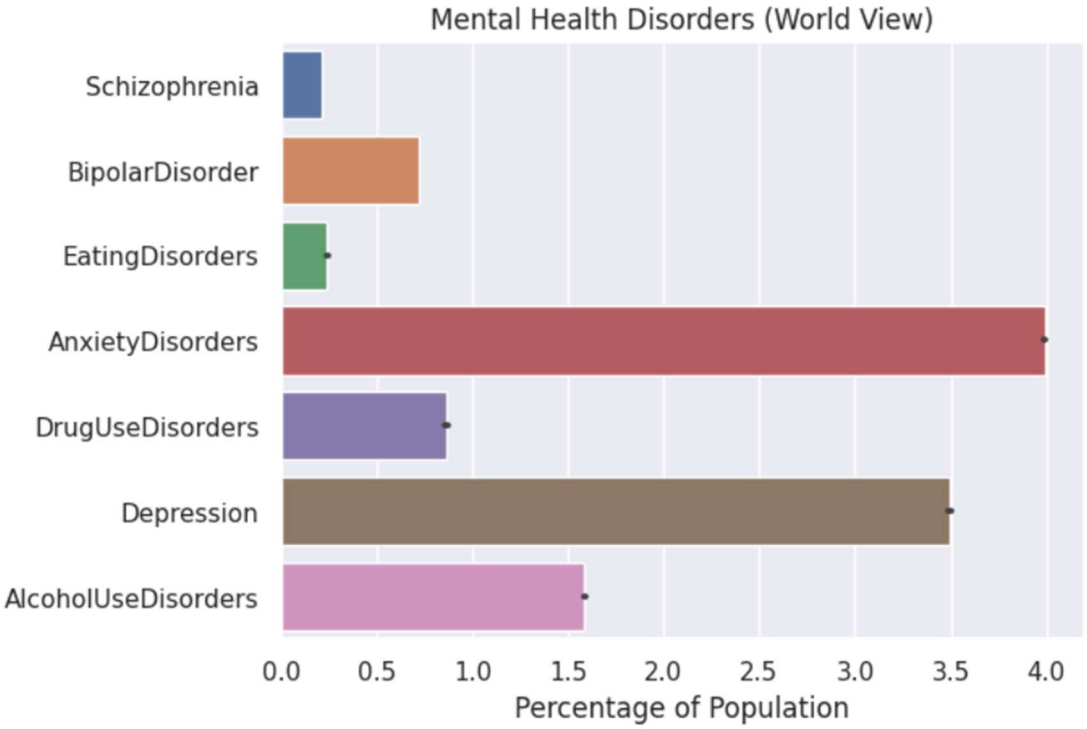
We have visualized the dataset we chose in meaningful ways using different plotting techniques.

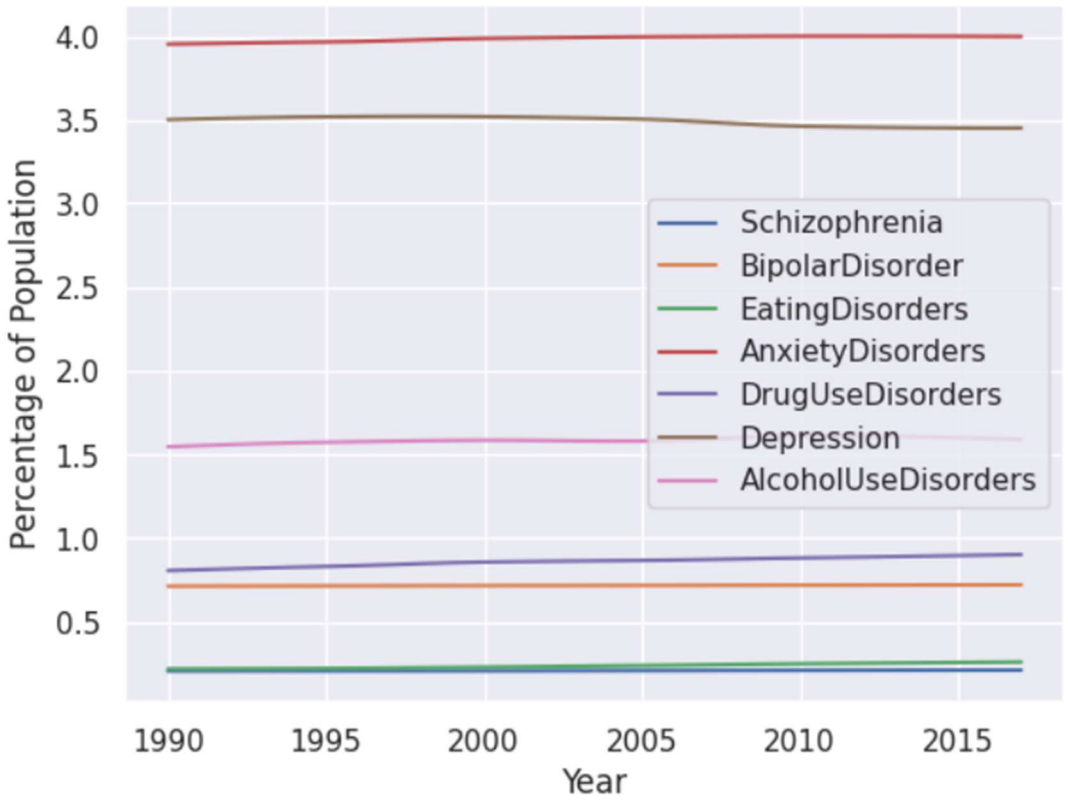

Year 1 0.039 0.012 0.087 0.013 0.059 -0.035 0.017
Schizophrenia 0.039 1 0.2 0.67 0.47 0.48 0.17 -0.059
BipolarDisorder 0.012 0.2 1 0.71 0.66 0.36 0.15 0.08
EatingDisorders 0.087 0.67 0.71 1 0.7 0.49 0.24 0.035
AnxietyDisorders 0.013 0.47 0.66 0.7 1 0.62 0.38 -0.14
DrugUseDisorders 0.059 0.48 0.36 0.49 0.62 1 0.37 -0.11
Depression -0.035 0.17 0.15 0.24 0.38 0.37 1 0.05
AlcoholUseDisorders 0.017 -0.059 0.08 0.035 -0.14 -0.11 0.05 1
Year
Schizophrenia
BipolarDisorder
EatingDisorders
AnxietyDisorders
DrugUseDisorders
Depression
AlcoholUseDisorders
1.0
0.8
0.6
0.4
0.2
0.0


Distribution of datapoints (Depression vs Alocohol Use Disorders)
6
5
4
3
2
1
2
3
4
5

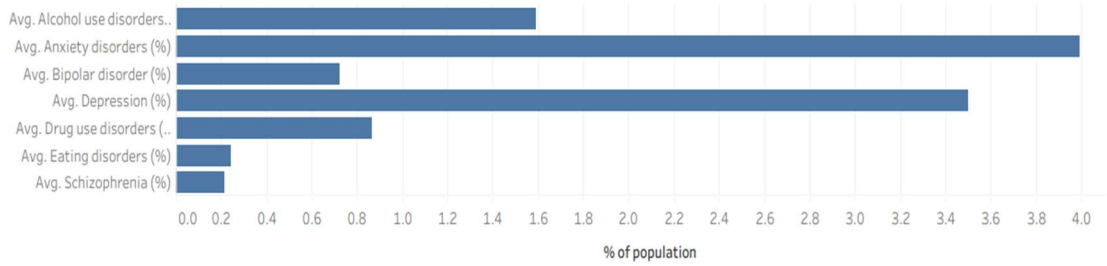

2
3
4
5
6
Depression


2
3
4
5
6
7
8
9
AnxietyDisorders


Mental Health Disorders (World View)
Schizophrenia
BipolarDisorder
EatingDisorders
AnxietyDisorders
DrugUseDisorders
Depression
AlcoholUseDisorders
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Percentage of Population

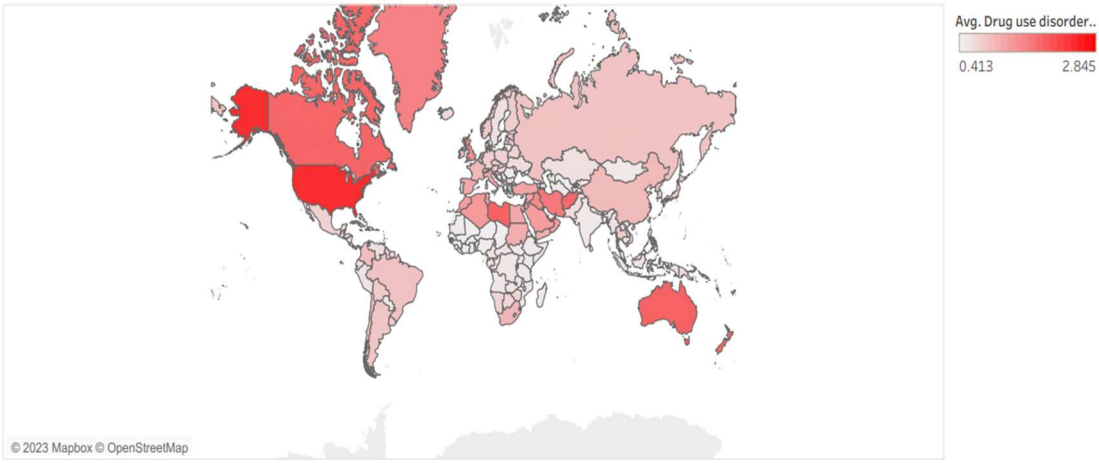

Percentage of Population
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Schizophrenia
BipolarDisorder
EatingDisorders
AnxietyDisorders
DrugUseDisorders
Depression
AlcoholUseDisorders
1990
1995
2000
2005
2010
2015
Year

Various Mental Health Disorders (Average of Different Countries)

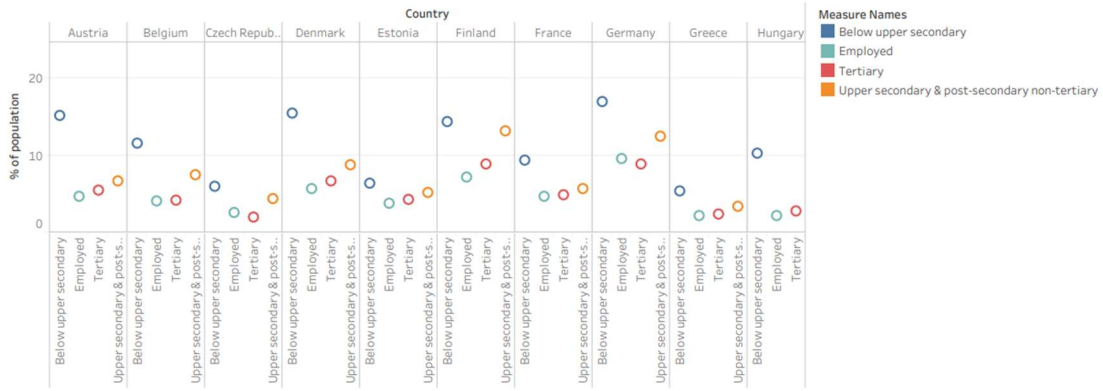


Depression in Different Countries (i.e., % of population having Depression)

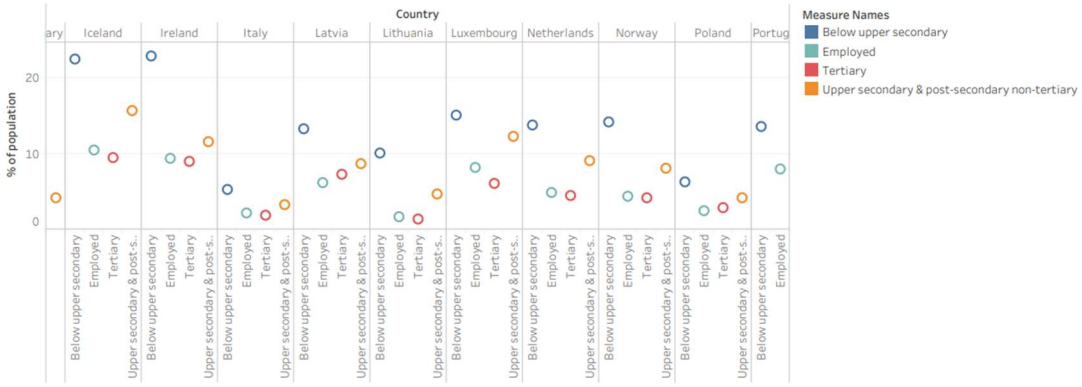
Drug Use in Different Countries (2)



Depression by Education Level in Different Countries

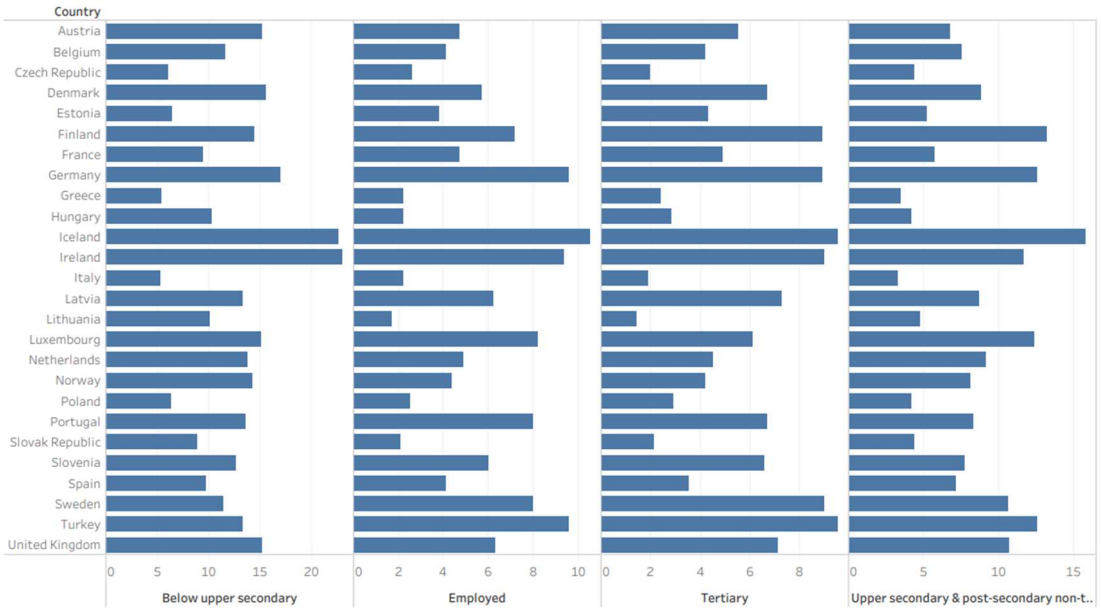


Depression by Education Level in Different Countries

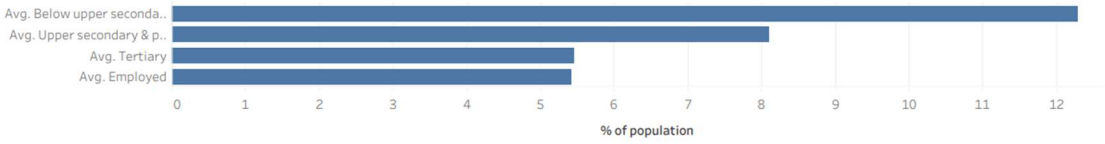


Depression by Education Level in Different Countries

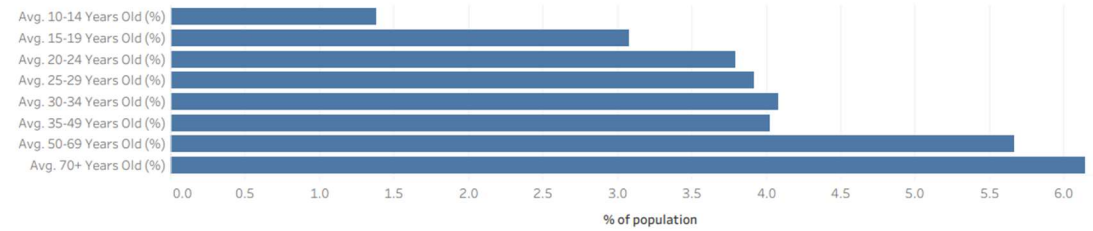
Depression by Education Level in Different Countries



Depression by Education Level (Average of Different Countries)



Depression by Age (Average of Different Countries)



CHAPTER 4

METHODOLOGY

4.1 PROCEDURE TO SOLVE THE GIVEN PROBLEM:

Data collection: The first step is to collect a dataset of an organisation with employees information init. The dataset can be found from organisations website or any dataset websites which are publicly available datasets, or collecting data from a company.

Data cleaning and pre-processing: The collected data needs to be cleaned and pre-processed to remove any inconsistencies or irrelevant information it may be null values, duplicates. This involves techniques such as removing special characters, numbers, and punctuation, converting the text to lowercase, and removing stop words.

Feature extraction: The pre-processed text data is then transformed into a numerical representation that can be used as input to machine learning algorithms. This can be done using techniques such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), or word embeddings.

Model selection: A suitable machine learning algorithm is chosen as the predictive model to classify the employee leaves the organisation or not. This is a binary classification problem. Popular algorithms for classification include logistic regression, Support Vector Machines (SVM), Naive Bayes, and Neural Networks.

Model training and testing: The model is trained on a subset of the data and tested on the remaining data to evaluate its performance. This involves techniques such as cross-validation and hyperparameter tuning.

Model evaluation and analysis: The performance of the model is evaluated using various metrics such as accuracy, precision, recall, and F1-score. The analysis involves investigating the factors that contribute to the accuracy of the model, such as the quality of the training data, the choice of features, and the selection of the machine learning algorithm.

Deployment: Once the model is trained and tested, it can be deployed in a production environment where it can be used to automatically to predict employee's attrition. This can be done using web applications, APIs, or mobile applications.

4.2 MODEL ARCHITECTURE:

The model architecture of our model is as follows:

Data collection: Obtain data from organisation/companies from there official websites or open sources.

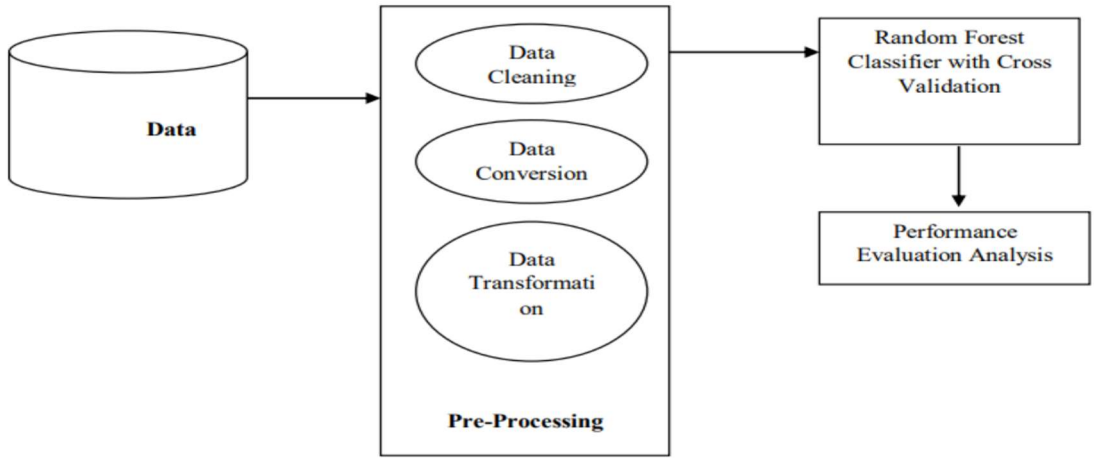
Data preprocessing: Clean the data by dropping null values, selecting relevant independent variables (population density, stringency index, new tests, new vaccinations, and positive rate), and choosing adependent variable (new cases). Train-test split: Split the data into training and testing sets using the train_test_split function from the sklearn library.

Model selection and training: Choose a classification model, such as logistic Regression, and fit the model to the training data using the fit method. Model evaluation: Evaluate the model's performance on the testing set by calculating various metrics, such as F1 score, precision, recall. Use the sklearn.metrics library to compute these metrics.

Prediction on new data: Test the model on new data by making predictions using the predict method. Create a new dataset of predictor variables and use the model to predict new cases.

Visualization: Plot the results of the model on different variables, such as positive rate and stringency index, to gain insights into the relationship between these variables and the number of new cases. Use the matplotlib library to create visualizations. Iteration and improvement: Analyze the results and refine the model as needed, such as adding or removing variables, changing the model type, or adjusting parameters.

Performance evalution: By using different metrice like f1 score, precision, recall the performance of the model can be found.



4.3 SOFTWARE DESCRIPTION

Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that you create can be simultaneously edited by your team members - just the way you edit documents in Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

As a programmer, you can perform the following using Google Colab.

- Write and execute code in Python
- Document your code that supports mathematical equations
- Create/Upload/Share notebooks
- Import/Save notebooks from/to Google Drive
- Import/Publish notebooks from GitHub
- Import external datasets e.g. from Kaggle
- Integrate PyTorch, TensorFlow, Keras, OpenCV
- Free Cloud service with free GPU

CHAPTER 5

RESULTS AND DISCUSSION

We got less error values, which means the model is good. The predictions are also quite close to the actual values. These are the output screens:

```
x = disorders[['Schizophrenia', 'BipolarDisorder', 'EatingDisorders', 'AnxietyDisorders', 'DrugUseDisorders', 'AlcoholUseDis
x
```

	Schizophrenia	BipolarDisorder	EatingDisorders	AnxietyDisorders	DrugUseDisorders	AlcoholUseDisorders
0	0.160560	0.697779	0.101855	4.828830	1.677082	0.672404
1	0.160312	0.697961	0.099313	4.829740	1.684746	0.671768
2	0.160135	0.698107	0.096692	4.831108	1.694334	0.670644
3	0.160037	0.698257	0.094336	4.830864	1.705320	0.669738
4	0.160022	0.698469	0.092439	4.829423	1.716069	0.669260
...	...	...	...	...	...	...
6463	0.155670	0.607993	0.117248	3.090168	0.766280	1.515641
6464	0.155993	0.608610	0.118073	3.093964	0.768914	1.515470
6465	0.156465	0.609363	0.119470	3.098687	0.771802	1.514751
6466	0.157111	0.610234	0.121456	3.104294	0.772275	1.513269
6467	0.157963	0.611242	0.124443	3.110926	0.772648	1.510943

6468 rows × 6 columns

```
y = disorders['Depression']
y
```

```
0       4.071831
1       4.079531
2       4.088358
3       4.096190
4       4.099582
          ...
6463    3.128192
6464    3.140290
6465    3.155710
6466    3.174134
6467    3.192789
Name: Depression, Length: 6468, dtype: float64
```

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)

x_train.shape
```

```
(4851, 6)
```

```
from sklearn.linear_model import LinearRegression

regressor = LinearRegression(fit_intercept = True)
regressor = regressor.fit(x_train, y_train)

print('Linear Model Coefficient (m): ', regressor.coef_)
print('Linear Model Intercept (b): ', regressor.intercept_)
```

```
Linear Model Coefficient (m):  [-2.78611872 -1.23113194  0.70760191  0.25303977  0.32120983  0.11842917]
Linear Model Intercept (b):  3.332056000234563
```

```
y_predict = regressor.predict(x_test)
y_predict
```

```
array([3.23089458, 3.4301233 , 3.41117655, ..., 3.39117297, 3.69611957,
       3.54102037])
```

```
y_test
```

```
2022    5.489475
1902    4.164680
1666    3.074819
2066    4.340423
5212    2.821500
          ...
2999    3.271182
4040    2.859093
5047    2.538367
764     3.902352

1635    2.588871
Name: Depression, Length: 1617, dtype: float64
```

```
from sklearn import metrics
acc = metrics.mean_squared_error(y_test, y_predict)
print("Mean Square Error: ", acc)
acc = metrics.mean_absolute_error(y_test, y_predict)
print("Mean Absolute Error: ", acc)
wts = regressor.coef_
incpt = regressor.intercept_
print("Slope: ", wts, "\nIntercept ", incpt)
```

```
Mean Square Error:  0.3485639688411424
Mean Absolute Error:  0.46563167394972427
Slope:  [-2.78611872 -1.23113194  0.70760191  0.25303977  0.32120983  0.11842917]
Intercept  3.332056000234563
```

```
regressor1 = pd.DataFrame({'Actual value':y_test, 'predicted value':y_predict})
regressor1.head()
```

	Actual value	predicted value
2022	5.489475	3.230895
1902	4.164680	3.430123
1666	3.074819	3.411177
2066	4.340423	4.081722
5212	2.821500	3.625155

```
# Predictions
inpredict = [
    [0.1, 0.5, 0.1,  5, 1.5, 0.5],
    [0.12, 0.43, 0.12, 4.5, 1.34, 0.45],
    [0.11, 0.41, 0.13, 4.99, 1.49, 5.1]
]
regressor.predict(inpredict)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but Linea
  warnings.warn(
array([4.31486651, 4.17564049, 4.95806693])
```

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

In conclusion, the project "Mental Health Disorders Analysis and Prediction" is an important step towards improving mental health care by leveraging machine learning techniques to analyze and predict mental health disorders. The project aimed to address the issue of under-diagnosis and misdiagnosis of mental health disorders by providing a more accurate and reliable diagnosis, as well as predicting potential risks and taking preventive measures. The use of machine learning algorithms and data analytics has enabled us to analyze large volumes of data, identify patterns, and predict outcomes with greater accuracy. The project has the potential to improve the quality of mental health care by providing early detection and intervention for those at risk, thereby reducing the overall burden of mental health disorders on individuals and society as a whole.

Our future plan is to expand the scope of the project by including a wider range of mental health disorders and developing more accurate and reliable prediction models. This could involve collecting more data from a broader range of sources, including electronic health records, social media, and wearable devices, and to integrate the prediction models developed in the project into mental health care systems and clinical workflows. This would require further development and validation of the models, as well as integration with electronic health records and clinical decision support systems.

CHAPTER 7

REFERENCES

Mental Health Disorders Dataset

https://data.world/vizzup/mental-health-depression-disorder-data

Pandas documentation (2021)

https://pandas.pydata.org/docs/

NumPy documentation. (2021)

https://numpy.org/doc/stable/

Scikit-learn documentation. (2021)

https://scikitlearn.org/stable/documentation.html

Matplotlib documentation. (2021)

https://matplotlib.org/stable/contents.html

Seaborn documentation. (2021)

https://seaborn.pydata.org/

Literature Survey Paper - I

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3645145/

Literature Survey Paper - II

https://uwaterloo.ca/mental-health-wellness/recommendation-22-literature-review-summary