A Project Report
on

VISUALIZING PREDICTION OF PSYCHOLOGICAL DISORDER OF A HUMAN BY CONSIDERING HEALTH CONDITIONS THROUGH ML

Submitted to

In partial fulfillment of requirement for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

By

NANDANAVANAM V NAGALAKSHMI SAI HIMAJA
20711F0035

Under the esteemed Guidance of

Y. RAJASEKHAR, M.Tech,
Dept of CSE

Department Of Master of Computer Applications

OCTOBER - 2022

NARAYANA ENGINEERING COLLEGE::NELLORE

AUTONOMOUS

DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS

BONAFIED CERTIFICATE

This is to certify that the project entitled **"VISUALIZING PREDICTION OF PSYCHOLOGICAL DISORDER OF A HUMAN BY CONSIDERING HEALTH CONDITIONS THROUGH ML"** that is being submitted by

NANDANAVANAM V NAGALAKSHMI SAI HIMAJA in partial fulfillment of the requirements for the award of degree

of **Master of Computer Applications** Department to the Narayana Engineering College

(Autonomous), Nellore is recorded to be the bonafied work carried out by him/her under my guidance and supervision.

PROJECT GUIDE

Y. RAJASEKHAR, M.Tech.
Department of CSE

HEAD OF THE DEPARTMENT

Dr. J. Suresh Babu, Ph.D.
Professor & HOD
Department of MCA

EXAMINER- 1

EXAMINER-2

ACKNOWLEDGEMENT

I am extremely thankful to **Dr.P.Narayana**, the **Founder Chairman** of **Narayana Group** for his good initiation starting technical institution in Nellore. I am also thankful to **Mr.K.Puneeth**, the **Chairman** of **Narayana Group** for providing the infrastructural facilities to work in, without this the work would not have been possible.

I would like to express our deep sense of gratitude to **Dr. G. Sreenivasulu Reddy, Principal, Narayana Engineering College(A), Nellore** for his continuous effort in creating a competitive environment in our college and encouraging throughout this course.

I would like to convey our heartfelt thanks to **Dr. J. Suresh Babu**, **Ph.D, Professor & HOD** of **Master of Computer Applications** for giving the opportunity to embark up on this topic and for her continues encouragement throughout the preparation of the project.

I would like to thank our guide **Y. Rajasekhar, M.Tech**, Department of **Computer Science Engineering** for his/her invaluable guidance, constant assistance, support, endurance and constructive suggestions for the betterment of the project.

I also thank all the **staff members** of the Department of **MCA** for helping us directly or indirectly in completing this project successfully.

NANDANAVANAM V NAGALAKSHMI SAI HIMAJA

20711F0035

DECLARATION

I hereby declare that the project entitled **"VISUALIZING PREDICTION OF PSYCHOLOGICAL DISORDER OF A HUMAN BY CONSIDERING HEALTH CONDITIONS THROUGH ML"** has been done by us under the guidance of **Y. RAJASEKHAR,** Department of **CSE.** This project work has been submitted to NARAYANA ENGINEERING COLLEGE(A), NELLORE as a part of partial fulfillment of the requirements for the award of degree of **Master of Computer Applications**.

I declare that this project report has not been submitted at any time to another institute or University for the award of any degree.

NANDANAVANAM V NAGALAKSHMI SAI HIMAJA

20711F0035

Place: NELLORE
Date:

ABSTRACT

On earth, there are a lot of people who suffer from depression. And it is a very weakening condition for all humans over the world. In some cases, it is difficult for people to live their own life normally, and in other cases, it may also lead to death by suicide. Initial stage doctors forcibly try to find there is a need to provide treatment for mental health conditions like depression without any proper training on how to handle such cases.

There is proof that there is an integrated approach where medical practitioners consistently screen patients for psychological disorders and work with psychologists and other psychological professionals a well to treat patients leads to a reduction in the cost and brings better patient outcomes as well. Though this kind of approach may require a lot of buy-ins, requires more training, and logistically it is not possible.

Using data from the Centers for Disease Control and Prevention National Health and Examination Survey, by applying machine learning it is predicted whether a person is depressed or not by considering the information in the patient's medical file. These predictions can be used to put psychologically disordered patients to be in touch with experienced mental health professionals shortly and trouble-free.

Keywords : psychological disorder, psychologist, psychological professionals, depression, prevention, training, prediction.

CONTENTS

LIST OF FIGURES, TABLES & PHOTOGRAPHS

1.INTRODUCTION

1.1 DATA VISUALIZATION

Data Visualization is the representation of data in a graphical format. This can be done with the help of visualization tools in the form of charts, graphs, and maps. Data Visualization provides an accessible way to understand outliers, trends, and maps. It is a format of translating visual context as a graph, to get the user i.e., the human brain easily understandable. Using Data visualization it is easier to identify patterns in large datasets.

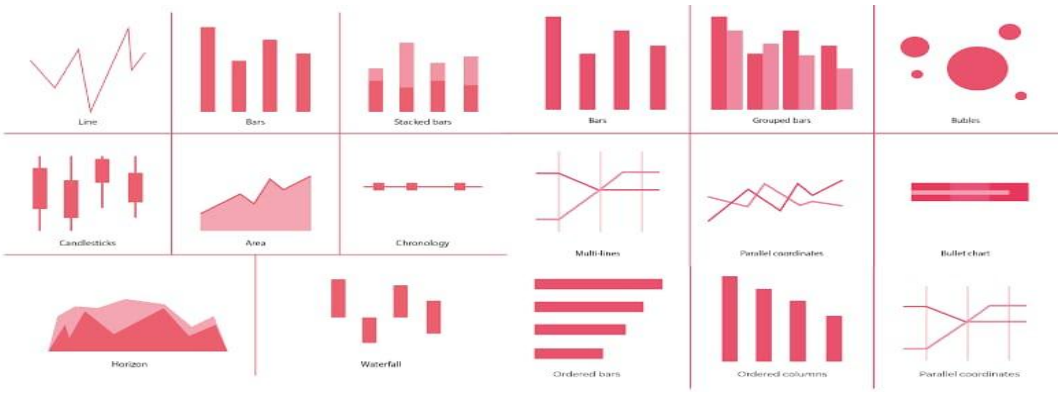


Fig 1.1 Data Visualization

1.2 DATA ANALYSIS

Any data which is needed to be visualized has to be analyzed. Analyzing data is basic for visualization. Data Analysis is the process of inspecting, cleansing, transforming, and data modeling which is used to discover useful information. Data Analysis. Data Analysis is the process that systematically applies statistical and logical techniques which describe condensing and evaluating data.

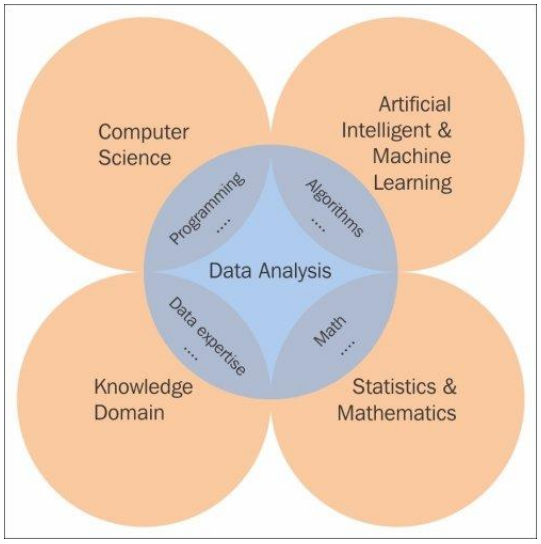
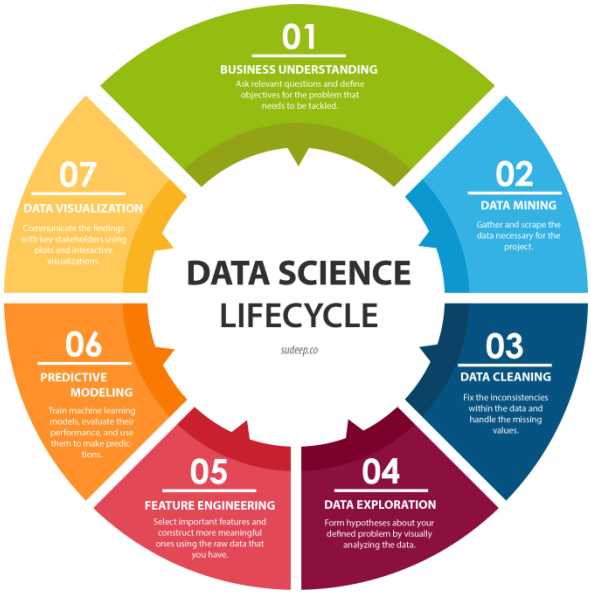


Fig 1.2 Data Analysis

1.3 DATA SCIENCE

Data Science deals with vast volumes of data by using modern techniques and tools which can be found using unseen patterns. Data Science creates predictive models with the help of complex machine learning algorithms. For data analysis data comes from different sources which can be represented in many different formats.

- Using Data Science we can capture the data, this can be done through data acquisition, data entry, and data extraction. This process gathers both raw structured and unstructured data.
- Maintenance of data can be done using data science, this process involves data cleansing, data staging, data processing, and data warehousing. With data science, data processing can happen this involves clustering, classification, and data modeling. determination of predictive analysis happens.
- Analyzing data is a key point in data science, this can be done along with predictive analysis, regression, and mining.
- Communication is the main theme of data science, here comes the main concept i.e., data visualization, and business intelligence. the analysis is in an easily understandable way such as graphs, and charts.



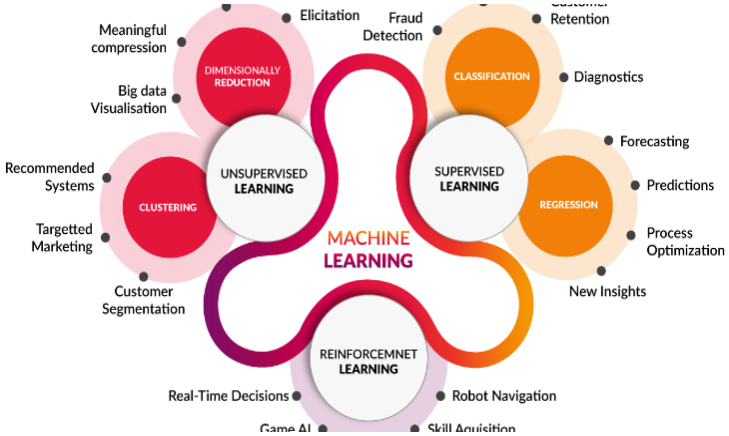
1.3 Data Science

1.4 MACHINE LEARNING

Machine Learning is a part of Artificial Intelligence, which is defined as the capability of a machine to act like an intelligent human behavior. Artificial Intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems. It is a modern innovation which has enhanced many industrial professional processes. Machine Learning focuses on using statistical techniques to build intelligent systems. Machine Learning has three kinds,

- Supervised learning
- Unsupervised Learning
- Reinforcement learning.

Machine learning allows software applications to become more accurate at predicting outcomes without being explicitly programmed.



1.4 Machine Learning Methods

1.5 ABOUT PROJECT

According to WHO, universally more than 200 million people have depression. Every year most suicides are caused by depression, which has been the main cause of young people's death. The National Inistitute of Mental Health has found that 7.1% of young adults are more depressive than others.

It is identified by Psychological Association that primary care doctors as physicians are requested to do medication for mental disorders. Such cases are considered depression without proper training about handling that kind of treatment. According to WHO, 70 percent of primary physician visits are caused by patients' psychological issues, other than 80 percent of patients who are having symptoms without a diagnosis will be receiving proper psychological treatment from a physician, and only 10 percent visit a psychological professional. 70 percent of patients are not getting diagnosed properly. Of the people who try suicide, there is 90 percent of patients have a psychological disorder and 40 percent of patients have already visited their physicians within the last 3 months.

From a Survey, physicians are being concentrated on different outcomes for patients, cost of care, and many different points between psychologically disordered persons which are provided more relevant between doctors and patients. From Survey we have found that mental health issues for patients who get intervention, low cost, better utilization of health care services, and outcomes of patients are improved.

1.6. PROBLEM STATEMENT

1.6.1 AIM

The objective of this project is to gather multiple kinds of data from patients who visit a psychologist and that which could be in a medical report of that particular patient. Through that data, prediction is done by considering different algorithmic performances.

1.6.2 SCOPE

The prediction can be visualized once it starts the performance of algorithms. Later on, the final prediction will be used and accessed by both psychiatrists and psychologically disordered people. This help to make the patient's file flagged to make the doctor & patient get in touch shortly & trouble-free.

2. LITERATURE SURVEY

Generally, there are a lot of people who suffer from depression. In normal terms, it is easy to find a person's emotions. This can be done by observing different movements of a person or can be expected from their expressions too. In this normal approach, a human sees another human with eyes, and then after processing data in the brain human will be delivered with proper results. But, hence this is a manual process it is hard to gain proper results because humans may do mistakes so the results would be affected.

So, Python Came into existence to detect a human face. For detecting a face firstly it is needed to be recognized. In python, there are different packages to perform face recognition. And also after detecting the face it will be processed o produce correct results through recognition.

The normal approach is fine but to do emotion detection a bit more accuracy is needed. Using Deep Learning Emotion Detection is identified using a model. Here model in the sense we train the model to make it learn. In Deep Learning two kinds of learning are Supervised Learning and Unsupervised Learning.

Supervised Learning means we have to train the model by representing different features and make the model learn the difference and classify among them. Unsupervised Learning means the model learns by itself, it learns by itself through previously trained data. Based on that trained data model will assume and calculate proper results.

In Deep Learning we have the concept of a Neural Network. This neural network scans the data from the user processes it in hidden layers and then produces results in the output layer. It captures the image through a webcam and recognizes it and then through the hidden layer it performs operations later on it detects the emotion of a person and produces the result.

Also after detecting the emotion of a person it shows the result in text format like happy, sad, and angry. But this approach is not helpful to reduce depression. It only recognizes and detects emotion it is not much useful.

So, Using Machine Learning Prediction of depression can be attained in the best format. It can be done by considering many psychological and medical conditions of humans. Through this, there is a scope of reduction of depressed people over the world.

3. SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

- In the existing system, there is a normal approach for detecting the emotions of a person.
- This approach has only the process of face recognition through this model will detect faces and can recognize what kind of expression does face shows.
- By this, it captures and processes expressions to analyze what kind of emotion it is and then produces output in the format of text.

3.2 CON'S OF EXISTING SYSTEM

- There's only one prediction in the existing system, no further approach is given.
- It is harmful if we leave mental disorders patients without any treatment.
- Traditional Solutions didn't work out in many cases.

3.3 PROPOSED SYSTEM

- In the proposed system prediction of the psychological disorder of a person is identified through healthcare-related data.
- The proposed system shows that 71% of those who have depression and 80% of those who don't have depression can be correctly identified.
- Better results could be yielded by adding more patient information to the data or testing more types of models.

3.4 PRO'S OF PROPOSED SYSTEM

- It will be helpful for both patients as well as doctors.
- Keen observation & personal guidance will be given to patients.
- Fast & accurate results will be produced.

3.5 FEASIBILITY STUDY

3.5.1 TECHNICAL FEASIBILITY

This project requires a technical setup for the development of a system. Implementation of this project requires python along with a lot of libraries offered with it. Since the

requirements are all open-sourced they are easily accessible. Also, the versions of the libraries are being updated we have to get updated to the latest version.

3.5.2 ECONOMIC FEASIBILITY

This project is developed within the budget because all the requirements which are used for the implementation of this project are freely available. This project is developed at no other additional cost. Developing this project is a bit cheaper when compared to other projects. So, developing this project is economically feasible.

3.5.3 TIME FEASIBILITY

The amount of time required to develop this project is dependent on the size of the dataset. Even if we have a large dataset it takes more time to do the implementation, Through this project using a lot of machine learning libraries we have decreased the execution time. Hence, the implementation of this project is highly feasible within a limited time.

3.6 SOFTWARE REQUIREMENTS

- Operating System – Windows 7 or above
- Programming Language – python
- Developing Environment – Visual Studio Code version 2022
- Domain – Backend

3.7 HARDWARE REQUIREMENTS

- Processor – intel CORE i3(64 bit)
- RAM – 6 GB
- SSD – 250 GB

3.8 USER REQUIREMENTS

3.8.1 SYSTEM

Operating System

An Operating System is system software that is used to manage both hardware and software resources. It manages memory and allows us to communicate with memory. There are different kinds of operating systems like Windows, Linux, and macOS we can use any of them.

Programming Language

Python is an object-oriented high-level programming language. Python is considered the easiest language with high productivity. It is used in many technologies Artificial Intelligence, Data Analytics, Data Visualization, web development, game development, etc. ,Python can be used in both front-end and back-end development.

Developing Environment

IDE is an Integrated Development Environment, a software that is used to build applications that combine both developer tools and Graphical User Interface(GUI).Different languages contain different IDEs, also some IDEs support multiple programming languages. Specifically, some IDEs that can work with python are Jupyter notebook and Visual Studio Code.

Domain

A domain is a targeted area of a specific task. Python and Machine Learning are always considered as the backend whereas data visualization acts as both frontend and backend.

3.8.2 PACKAGES

NumPy

NumPy stands for NumericalPython it is a python library that is used to work with multi-dimensional arrays and also using NumPy. We can access many functions and make use of pandas in this project to operate with algebra and matrices.

Pandas

Pandas is a python library it is built on top of NumPy. Pandas are used to perform machine learning tasks. We make use of pandas in this project to do data analysis through Machines. Pandas have more efficiency than NumPy.

Glob

- Glob is used to return all file paths that match a specific pattern. Using this to search for a specific file pattern.

OS

- OS provides functions for creating and removing and identifying current folders.

matplotlib

matplotlib is used for Visualization. Using matplotlib we can interactively visualize static data. plotting can be done through this itself.
Through matplotlib we can generate graphs like bar graphs line plots box plots, scatter lot plots, pie charts, etc.

Seaborn

Seaborn is used for data visualization. Though seaborn is dependent on matplotlib it gives a high-level interface to produce attractive statistical graphics full of information.

SkLearn

Sklearn is pronounced as scikit-learn. It is a machine learning library that is specifically designed for python language. Sklearn contains a bunch of algorithms that perform different kinds of machine learning tasks. Sklearn has various functionalities like classification regression clustering. It contains different modules to perform different tasks.

One Hot Encoder

OneHotEncoder is one of the package in sklearn. OneHotEncoder is used to preprocess categorical data. By doing this it creates a new binary feature for each possible category and assigns a value of 1 that corresponds to its original category.

Quantile Transformer

QuantileTransformer is a module from sklearn. QuantileTransformer is used to reprocess.

The method transforms the features to follow a uniform distribution or a normal distribution.

This is a technique for transforming numerical input or output variables to have a uniform probability distribution.

3.8.3 ML ALGORITHMS

KMeans

KMeans is a machine learning algorithm that is used to perform clustering. KMeans is an iterative algorithm, that tried to partition the dataset into Kpre-defined distinct non-overlapping clusters.

In this case, all data points belong to only one group. A cluster is nothing but a collection of data points that are together due to some specific similarities.

A target number K referred to the number of centroids need in the dataset.

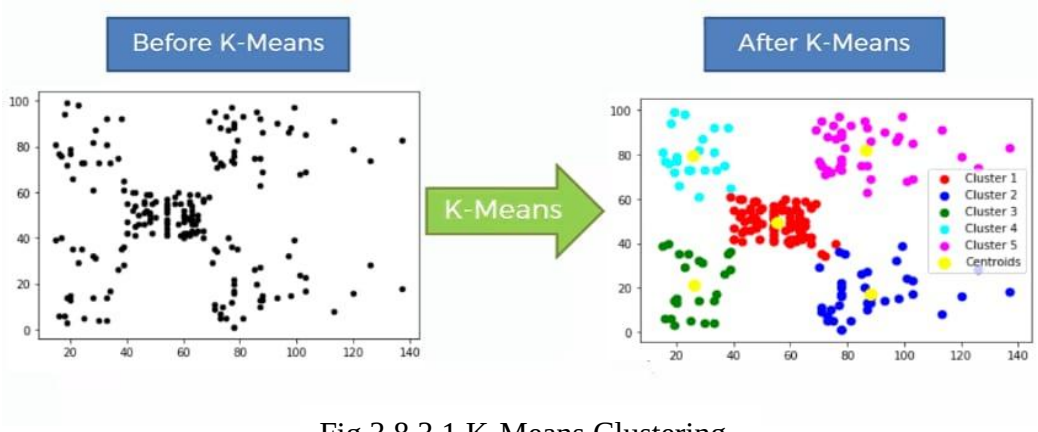


Fig 3.8.3.1 K-Means Clustering

Dummy Classifier

Dummy Classifier the name itself specifies its meaning. A Dummy classifier is a classifier model that makes predictions without trying to find patterns in the data. It also makes predictions that ignore the input features. This makes predictions using simple rules.

Regression

Regression is a Machine Learning algorithm it is used to predict continuous value. Regression means a mathematical approach to finding the relationship between single or multiple variables. The regression algorithm comes under supervised learning. supervised learning means the process of mapping between the input and output. Regression had split into two types Linear Regression, and Logistic Regression.

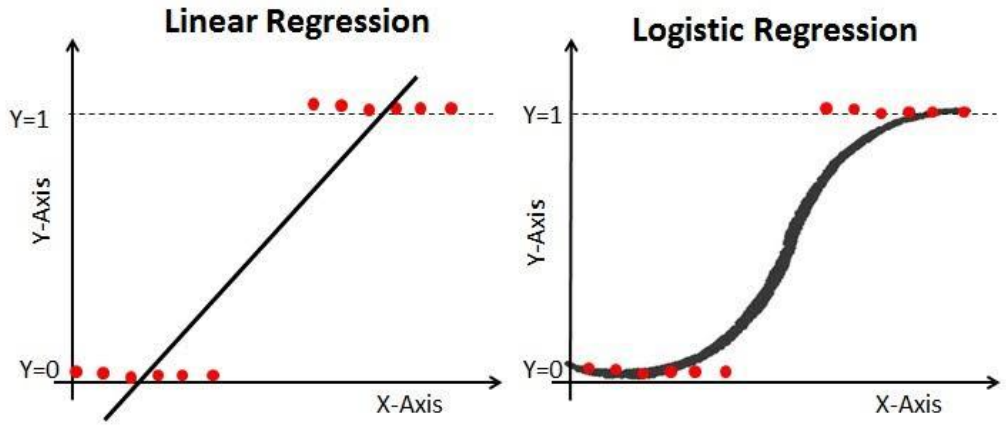


Fig 3.8.3.2 Regression

Linear Regression

Linear Regression is a machine learning algorithm. It is based on supervised learning. This is based on the value of another variable that means shows a kind of relationship between a dependent variable and multiple independent variables.It can be either simple linear regression or multiple linear regression. Here the relationship can be positively linear or negatively linear.

Logistic Regression

Logistic Regression is a kind of regression algorithm in machine learning. It also comes under supervised learning. Used to predict categorical data.Solves classification problems and gives probabilistic values that lie between 0 and 1.

In Logistic Regression, we use the sigmoid function to gain threshold value which helps to have an S-Shaped curve that provides accurate value prediction.

Grid Search CV

GridSearchCV, a library function from model_selection package, a package from sklearn. It is a technique that is used to search best parameter values.

Those Parameters are taken from a set of grid parameters. A cross-validation method that needs parameters to extract the best-parameterized predictions.

GridSearchCV can only be done on training data after the completion of training, testing, and splitting of data.

GridSearchCV can be applicable on non-differentiable and non-continuous functions. GridSearchCv is accurate when we apply it to multiple linear regression.

Random Forest

Random Forest is a machine learning algorithm. This algorithm can be used for regression and classification. Random Forest is based on ensemble learning that comes under supervised learning. By using random forest we can get efficient results with high accuracy for even large datasets.

Random Forest Classifier

- Randomforestclassifier is a classification algorithm that contains a lot of decision trees. It makes predictions based on many independent base models.

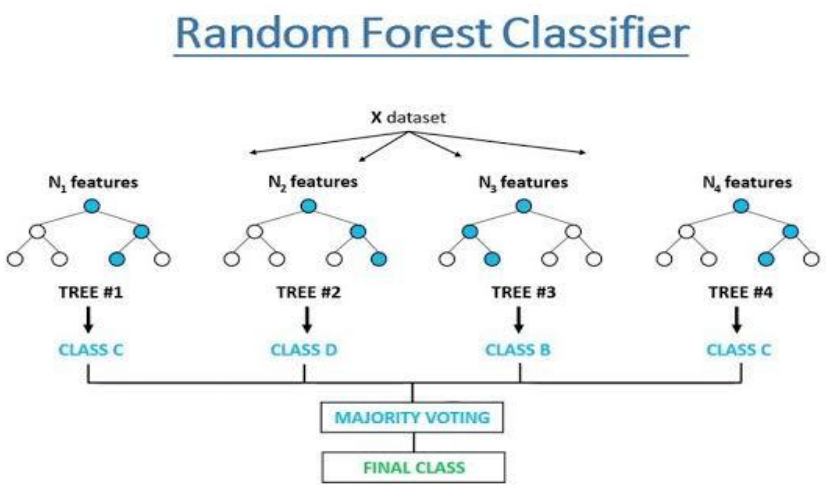


Fig 3.8.3.3 Random Forest Classifier

Extra Tree Classifier

Extratreeclassifier is a method of ensemble learning, that provides classification results. Combining trees is known as the ensemble method. Extremely Randomized Trees Classifier is called an extratreeclassifier. Used to fit on no.of randomized decision trees and it applies to the whole dataset. Feature on this is splitting the value from a random selection. a meta estimator fits for randomized trees, works on

dataset samples, and gains average which improves accuracy and control fitting for prediction.

XGBoost

XGBoost means Extreme Gradient Boosting, it is a form of classification from the path of supervised learning. Predicts a target variable after combining weaker models. Parallel tree boosting is provided by xgboost. It is an ensemble decision tree-based algorithm.

SVM

SVM stands for Support Vector Machine it is one of the machine learning algorithms. A supervised learning approach is used for classification. In many cases like detection, a recognition support vector machine is applied. It always works best on linearly separable data which means input should be flexible to split into two different classes. The separation can be done by using a hyperplane. To transform data there is a technique called kernel trick which finds the optimal boundary. This can be either a linear support vector machine or a polynomial support vector machine.

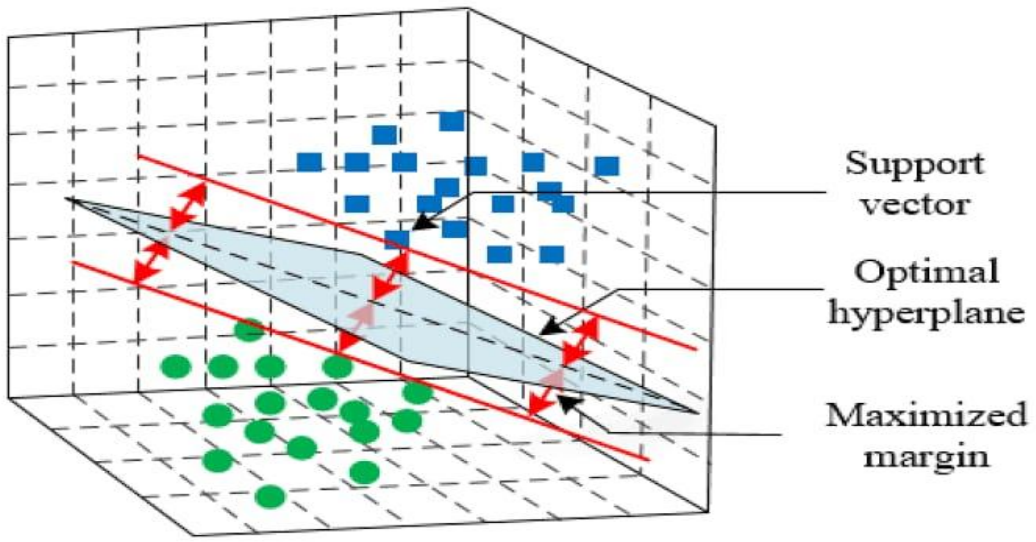


Fig 3.8.3.4 Support Vector Machine

SGD

Stochastic Gradient Descent is a machine learning optimization algorithm. This is used o calculate which is best between the predicted result and the actual data. Stochastic Gradient Descent performs classification as one vs all. Stochastic Gradient Descent doesn't belong to any model in machine learning, it's a path to train a model.

Voting Classifier

Voting Classifier is a model in machine learning, used for estimation training of base model also predicts based on the aggregation of every base estimator. Voting Classifier has two kinds Soft Voting Classifier, and Hard Voting Classifier.

Soft Voting

- Soft Voting Classifier provides a probability value of a specific target class and data points.

Hard Voting

- It also classifies according to the weights of each classifier whereas Hard vote classifier provides data depending on labels of class and associated weights to every classifier.

Confusion Matrix

A confusion matrix is a table that is used to describe the performance of a classification algorithm, this visualizes and summarizes about classification algorithm. In this confusion matrix, a machine learning method is used which allows the measurement of Recall, Precision, Accuracy, and AAUC-ROC curve. It makes it easy to see whether the system is confusing two classes that are commonly mislabeled one as another. In this, there are four terms called True Positives, False Positives, True Negatives, and False Negatives to make classification more accurate.

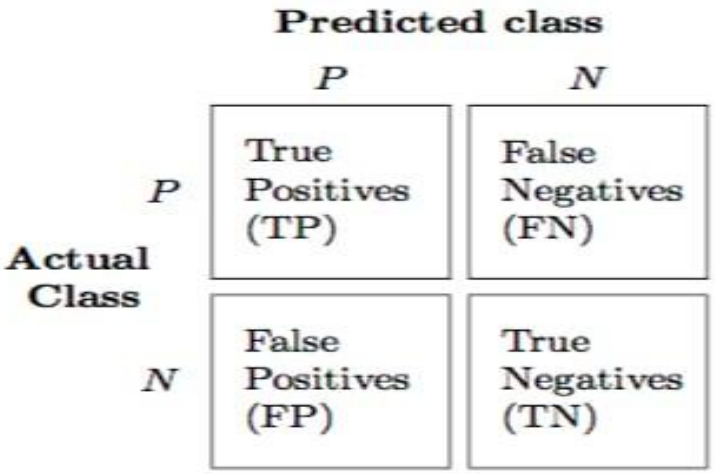


Fig 3.8.3.5 Confusion Matrix

4. SYSTEM DESIGN

4.1 ARCHITECTURE

- Architecture is the main step in a project, it specifies the whole path of the project.
- In architecture there will be short and clear diagrammatic approach is represented.
- The design of the project can be shown along with the different layers involved in it.

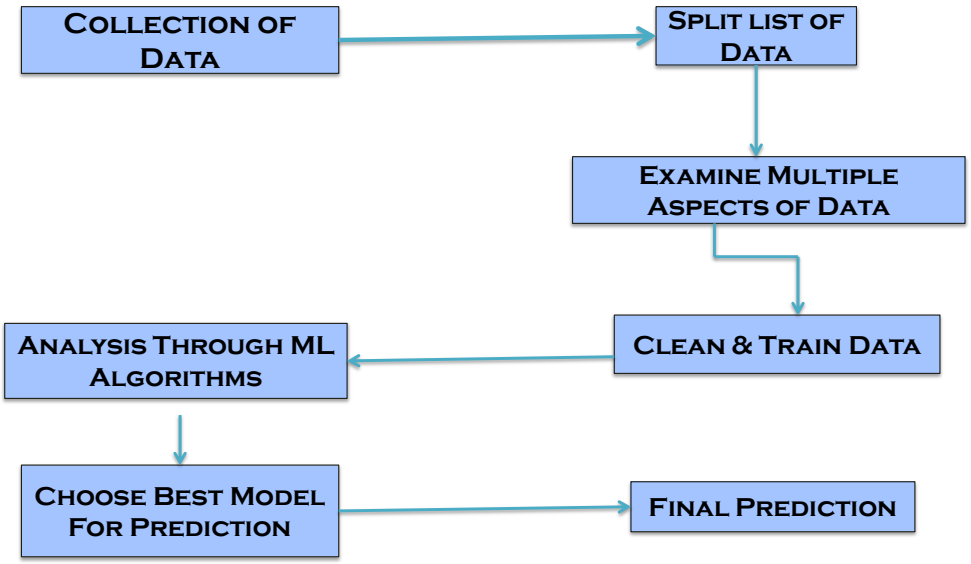


Fig 4.1 Project Architecture

4.2 UML DIAGRAMS

UML means Unified Modeling Language. It is a general-purpose modeling language for development. In the IT industry, UML diagrams are used to visualize system design in a standard way. UML Diagrams are used for a better understanding of the system. Through UML Diagrams one can maintain and document the system.

UML Diagrams are similar to blueprints but not the same. Unified Modeling Language is not a Programming language but a graphical language.UML Diagrams are split into many different categories, every single type of UML Diagram has its features.

- In different types of UML Diagrams, there are some primary diagrams, that are
  - Use Case Diagram
  - Sequence Diagram
  - Class Diagram
  - Activity Diagram

4.2.1 USE CASE DIAGRAM

Use Case Diagram is one of the Diagrams in Unified Modeling Language Diagrams.By using the Use case Diagram we can capture requirements and can model the behavior of the system Through the use case diagram, we can specify events of the system along with its workflow. Use case comes under behavioral diagrams Use case diagrams containing four main parts.

- Actor
- System
- Use cases
- Connecting Lines

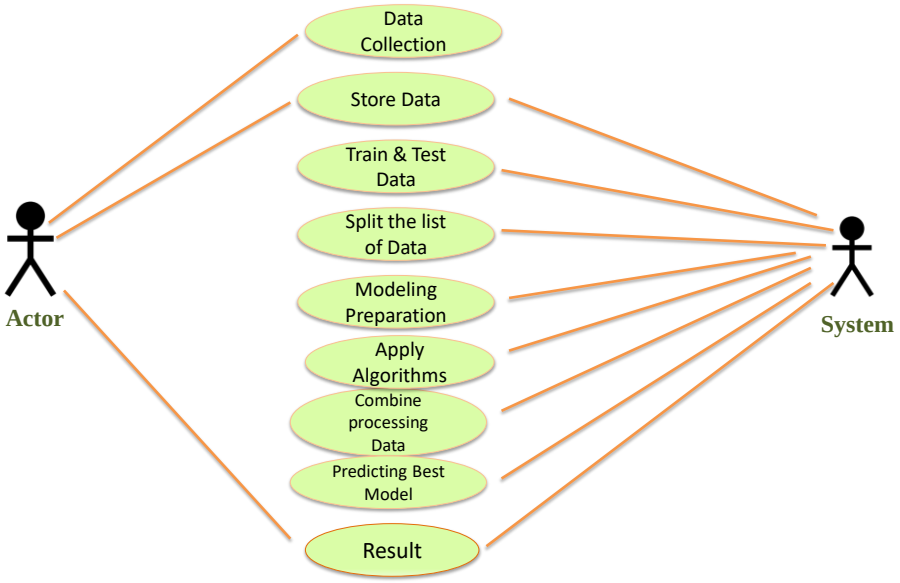


Fig 4.2.1 Use Case Diagram

- In this we have actor , System, Use cases.
- Use cases are,

- Firstly, collection of data should be done as a basic process because without data we can't do any kind of further process.
- Then, after collecting data we have to store that data in a dataset.
- After completion of storing data into the dataset we have to do training and testing to that data.
- Then, after completion of training and testing of stored data we can split data into different categories.
- Now we have to do Modeling Preparation for that categorized data.
- After that we have to Apply some machine learning Algorithms on that data
- Now we have to Combine Processing Data into a single form.
- Then prediction of best model can be done
- At last final prediction is considered as result.
- Here, collection of data , storing of data and results are related to user.
- Other use cases are related to the system.

4.2.2 SEQUENCE DIAGRAM

In Unified Modeling Language, one of the best diagrams is the sequence diagram. A Sequence diagram shows an interaction between objects through a sequence of messages. In Sequence Diagram, there will be objects associated with lifelines through messages that pass over time. Through Sequence Diagram we can describe how it is happening, in which order the process will happen, working of objects. Also, can get a better understanding of requirements. The sequence Diagram shows details of generated events through actors which are from outside of the system. The sequence Diagram shows the complete flow of the project. We can get interactive individuals from the top passing messages in temporal order to the bottom of the page. A lifeline in a sequence diagram represents a classifier it can also be "Self" which represents workflow from one another object.

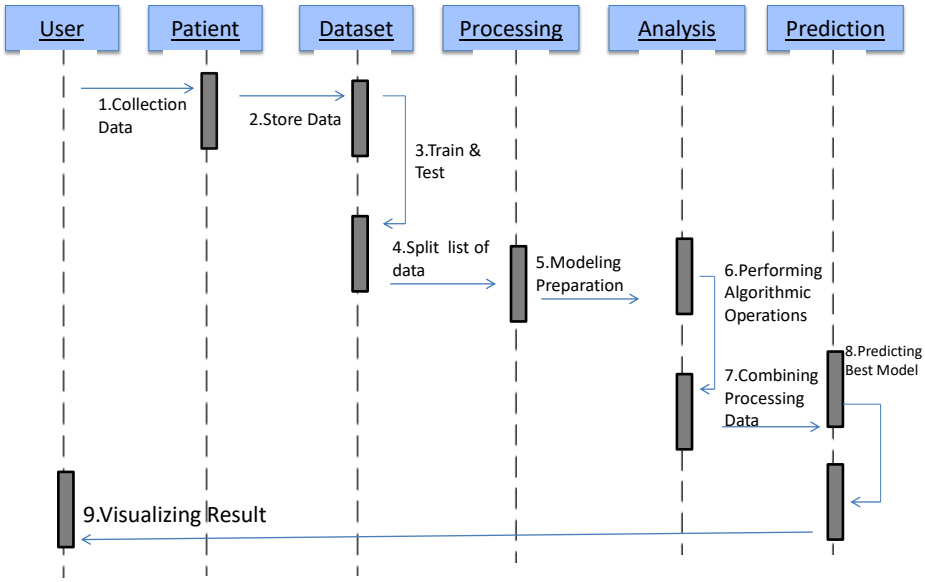


Fig 4.2.2 Sequence Diagram

Objects

- User
- Patient
- Dataset
- Processing
- Analysis
- Prediction

Work Flow

- Firstly, collection of data should be done as a basic process because without data we can't do any kind of further process.
- Then, after collecting data we have to store that data in a dataset.
- After completion of storing data into the dataset we have to do training and testing to that data.
- Then, after completion of training and testing of stored data we can split data into different categories.
- Now we have to do Modeling Preparation for that categorized data.
- After that we have to Apply some machine learning Algorithms on that data
- Now we have to Combine Processing Data into a single form.
- Then prediction of best model can be done

- At last final prediction will be visualized.
- In this every single message is passed from one another in a sequence fomat.

4.2.3 CLASS DIAGRAM

Class Diagram is one of the important diagrams in Unified modeling language. A class Diagram is considered a blueprint of the project. By using a class diagram one can specify the relation between two or more objects.

In the class diagram, there will be,

- Classes
- Attributes
- Operations

In Class Diagram mapping the structure of a specific system through modeling its classes and objects can be done. Through the class diagram, Every object-oriented method is the main building block. Specifications of base classes are done in a class diagram of the unified modeling language. In Class Diagram Attributes are represented with a “+” Symbol.

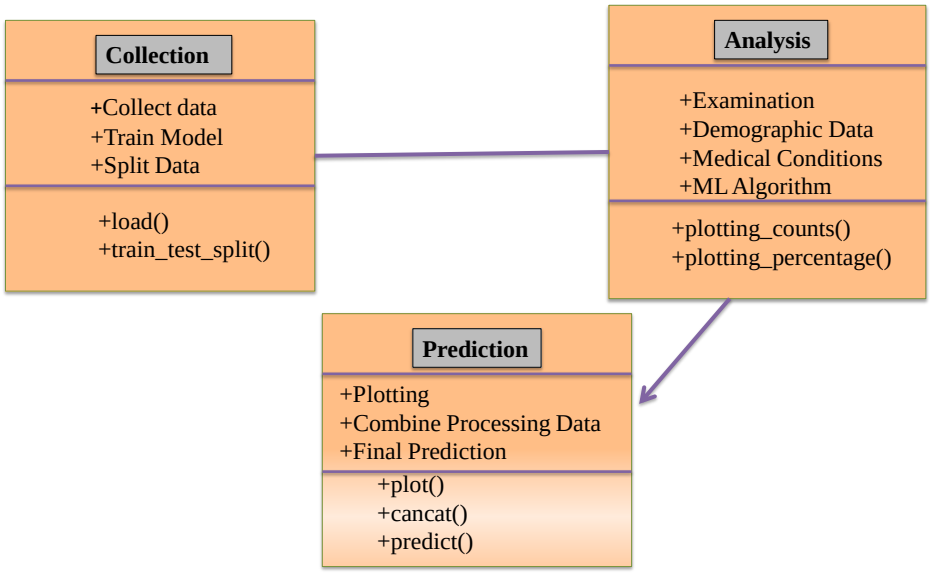


Fig 4.2.3 Class Diagram

Classes

- Collection
- Analysis
- Prediction

Attributes of class Collection

- Collect Data
- Train Model
- Split Data

Attributes of Class Analysis

- Examination
- Demographic Data
- Medical Conditions
- ML Algorithm

Attributes of Class Prediction

- Plotting
- Combine Processing data
- Final Prediction

Operations of Class Collections

- load()
- train_test_split()

Operations of Class Analysis

- plotting_counts()
- plotting_percentage()

Operations of Class Prediction

- plot()
- cancat()
- predict()

4.2.4 ACTIVITY DIAGRAM

The Activity Diagram is one of the best diagrams in the unified modeling language. By using an activity diagram we can describe a system's dynamic activities. In simple terms, an activity diagram is a form of activity that flows from one another. An activity Diagram is a kind of flow chart that captures a system's behavior dynamically. Activity Diagrams are used for constructing executable systems, this can be done with the help of forward and reverse engineering. The activity Diagram will have a set of activities from the initial state to the final state.

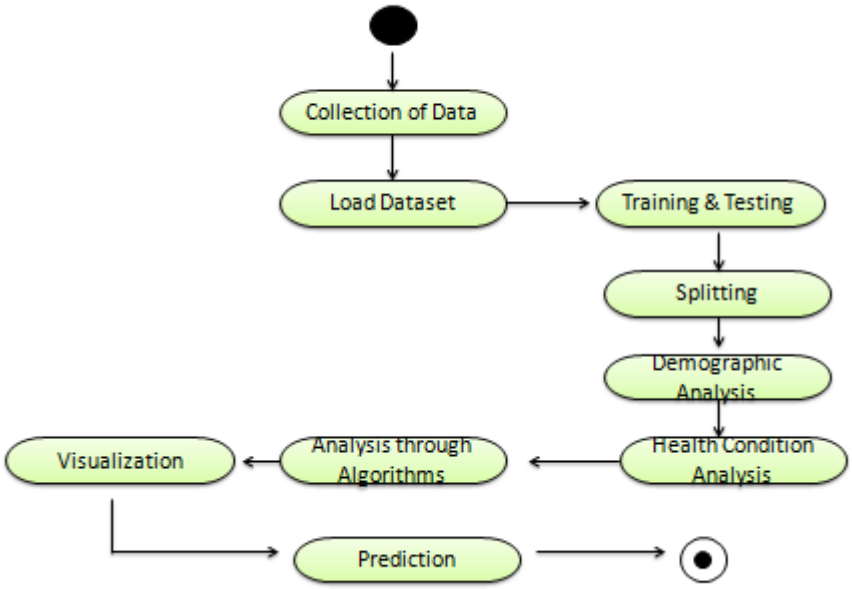


Fig 4.2.4 Activity Diagram

Activities

- The First Activity is Collection of Data
- The Second Activity is Loading of the Data Set
- Training and testing the model
- Splitting the list of data
- Demographic data analysis
- Analysis on health conditions
- Performing Machine learning algorithms to analyze the best model.
- Visualizing the analysis.
- Predicting final results.

4.3 MODULE DESCRIPTION

A Module is considered as a collection of source files along with build settings that allow units of functionality division. A Module can represent a detailed explanation of what we do in the project. A project can contain several modules, every single module represents a specific functionality.

- This Project is divided into four different modules
    - Collect data from CDC
    - Examine Data
    - Combine Data
    - Final Prediction

4.3.1 COLLECT DATA FROM CDC

- There will be different datasets in CDC, we have to collect the perfect dataset with which we have to do operations.
- In this module, we collect information about several health conditions of humans from CDC Dataset.
- After getting the perfect data we have to do data preprocessing training testing and splitting of data.
- After this, the data will be split into several categories.

4.3.2 EXAMINE DATA

- Examining Data is the primary module in this project.
- In this module, we try to inspect different aspects, considering a person who comes to a psychiatrist consulting due to his/her mental illness.
- We gather information from that particular patient, by surveying their daily lifestyle, present health issues, and past health issues.
- That means firstly we gather Demographic Data from patients, here demographic data has different categories in it. Demographic data includes,
    - Gender
    - Race
    - Region
    - Income
    - Birthplace
    - Education level
    - Financial Status
    - Marital Status

- Household Size
- Household income

- And then after that, we start a survey about Medical Conditions,
  - Asthma
  - Anemia
  - Stroke
  - Cancer
  - Thyroid
  - heart attack
  - diabetes
  - liver conditions
  - BMI
  - Blood Pressure
  - irregular pulse
  - Cholesterol
  - Standard Biochemistry profile
  - Blood Count
  - Sleep Disorders
  - Physical Activity
  - Alcohol Consumption
  - Smoking
  - Drug Addiction

- Every single category of the data is visualized, data visualization is done in the format of plotting bar charts.

- Now we have to Perform Algorithmic Operations on every Single Aspect.
- We do perform different Machine Learning algorithms for analyzing train and test data.
- The Machine learning Algorithms which are used to analysis are,
    - Logistic Regression
    - Random Forest
    - K-mean Clustering
    - Support Vector Machine
        - Linear SVM
        - Polynomial SVM
- In Every single Algorithm we do analysis on
    - Base Model
    - Elastic Net
    - Grid Search.
- All the algorithms are visualized according to their functionalities.

4.3.3 COMBINE DATA

- After performing different machine learning algorithms a data model is created.
- After that, we do combine all the data models. There are some algorithms to combine data.
- Those algorithms through which data models are combined are,
    - XGBoost
    - Confusion Matrix
    - Dummy Classifier
    - Extra Tree

- Prediction of data models are visualized for further process of best prediction.
- From those predicted data models we will choose the best model among them.
- The best Model is Chosen by considering several variations which are happened in trained data during algorithm performance.

4.3.4 FINAL PREDICTION

- Finally, after choosing the best model the prediction is categorized into two types that are either depressed or not depressed.
- This Final Prediction is also done with the help of some machine learning libraries which contains a set of machine learning libraries.
- Those algorithms which are used for final prediction are,
    - Stochastic Gradient Descent
    - Confusion Matrix
    - Standard Scaling
    - Stacking Classifier
    - Voting Classifier
        - Soft Voting Classifier
        - Hard Voting Classifier
- And then top featured categories are plotted again as non-tree models.
- At last, the final prediction of the best model is visualized.

5. IMPLEMENTATION

5.1 IMPLEMENTATION

Installation

- Firstly we have to install python 3.10 on our system
- And then we have to install visual studio code v2022 as Developing Environment.
- Now we have to Set up visual studio code with the required settings.
- Then we should have to install all the required packages of python using pip command.
- Syntax to install packages is

  "pip install package name".

Importing Packages

- After installation, we have to import every single package which we wanted to use. This can be done using import package name as alias name.

Loading Data Set

- We have to load the dataset by specifying its path using

  read_csv() method

Training, Testing & Splitting

- Then Training testing and splitting of data should be performed this can be done using train_test_split()
- Now we have to create separate lists to categorize the number of columns in the data set

Visualizing Categorical columns

- Then we have to plot those columns. this can be done through

  pf.plotting_counts()

  pf.ploting_percentages()

  pf.lot_num_of_cols()

- The above methods are performed multiple times to visualize different categories of data.

Modeling Preparation

- Now Modeling Preparation should be executed this can be done from sklearn preprocessing.
- To do this we have One Hot Encoder in sklearn.
  encoder = OneHotEncoder()
- Now scaling should be done on numerical columns through
- scale_cols = X_train.select_dtypes('number').columns

5.2 APPLYING ALGORITHMS

Performing Machine Learning Algorithms

Clustering

- Now KMeans Clustering should be performed to classify categories of data using cluster format.
- A Dummy Model should be taken to classify among models. that dummy classifier is from sklearn

Regression

- Now we have to apply logistic regression, logistic regression is from the linear model of sklearn.
- Fitting logistic regression to training data
- logreg_clf.fit(X_train_final, y_train)
- Evaluating logistic regression model pf.evaluate_model()
- then we do the ElasticNet model and GridSearch model.

Classification

- Now Classification should be done through Random Forest.
- RandomForestClassifier is accessible from the ensemble of sklearn
- Now classification should be done by ExtraTreeClassifier, it is from ensemble sklearn.
- And then XGBoostClassifier is used to get a better classification of the model.

- And then classification of the support vector machine should be performed as linear SVM and Polynomial SVM.
- Now to get the best performance and clear visualization we use the Stochastic Gradient Descent Linear Classifier.
- This can be done on the original preprocessing data through SGDClassifier of sklearn.linear_model.

Choosing Best Model

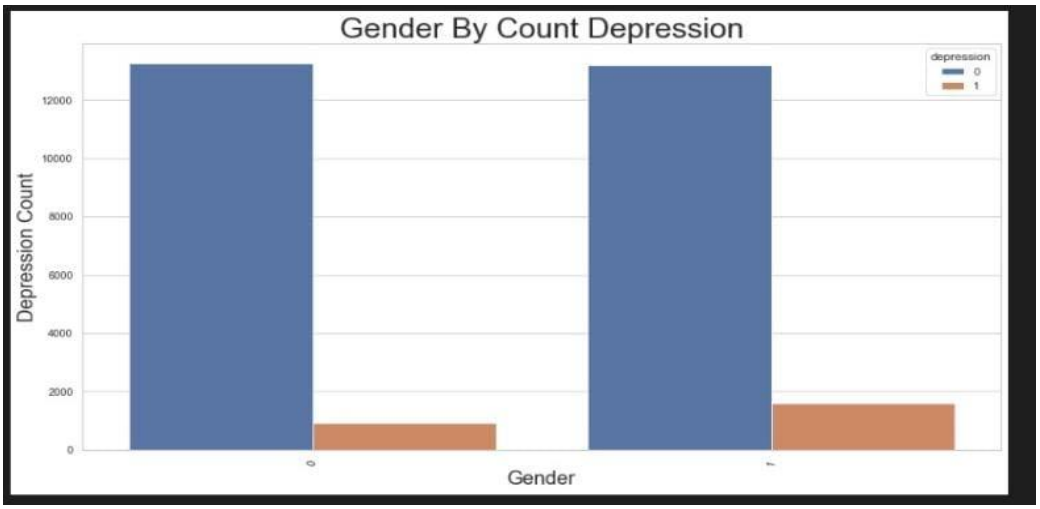
- At last, we classify all models with Voting Classifiers. Voting classifiers are two kinds one is a soft voting classifier and another is a Hard voting classifier.
- It is in VotingClasifier of sklearn.ensemble we do use estimators to classify among models.
- Finally using Voting Classifier, Stochastic Gradient Descent we got the visualization of the Top featured results who have depression.

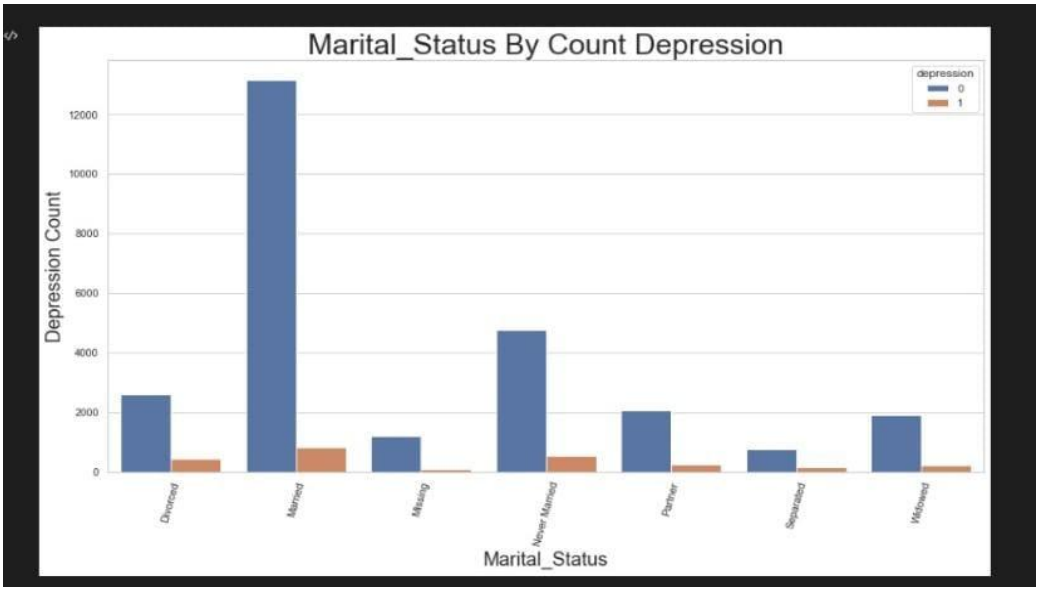
Final Prediction

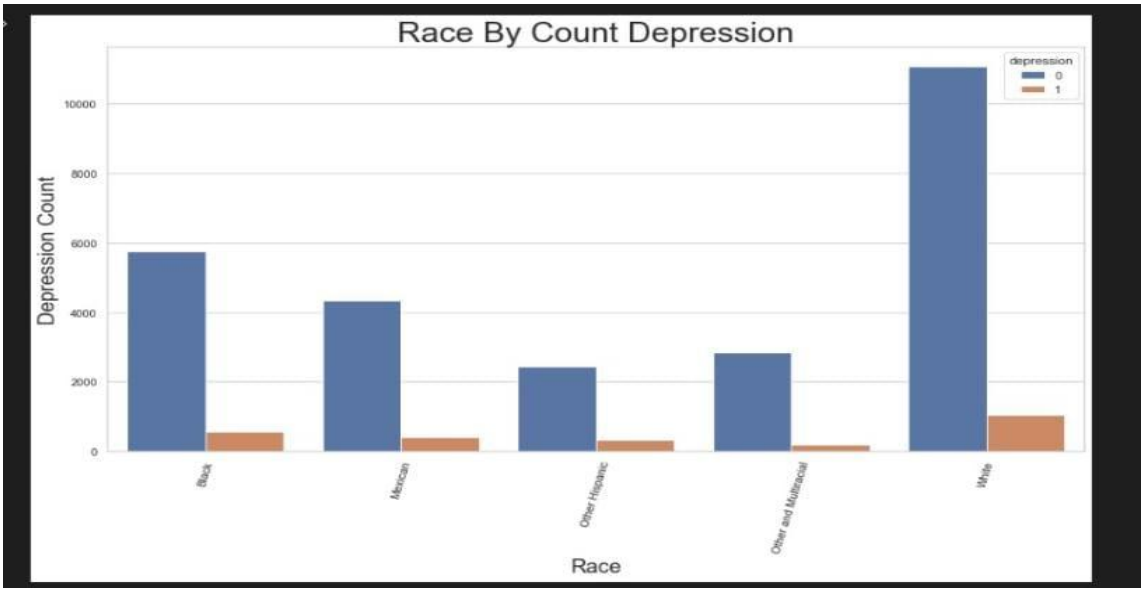
- Those Top featured results will be plotted again to visualize changes between the base model and rained the best model.
- Using Different classifiers finally we have found the best model to predict depressed / not depressed.

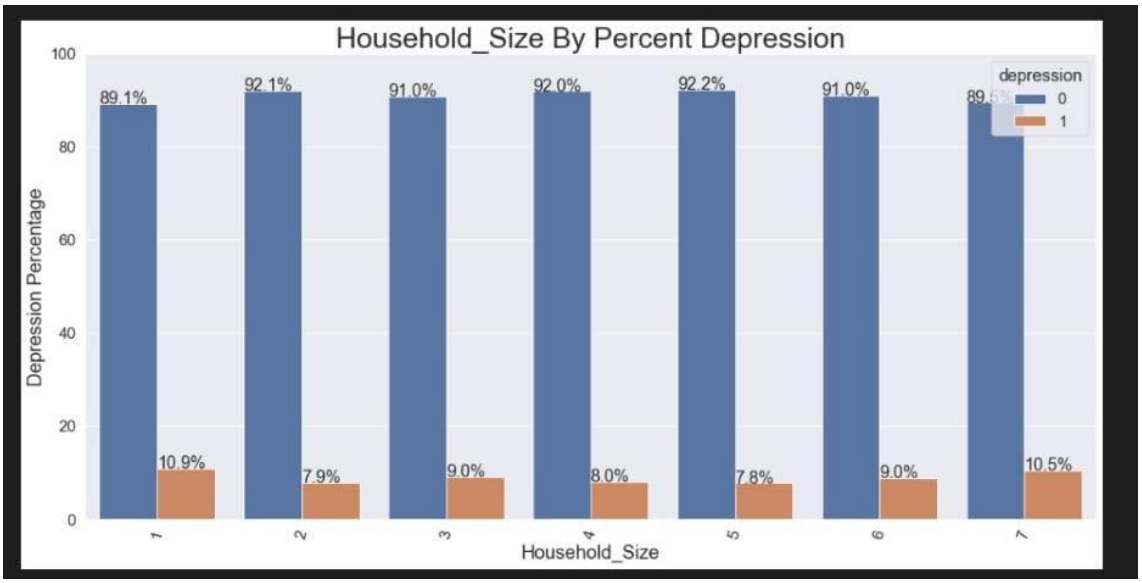
6. VISUALIZING RESULTS

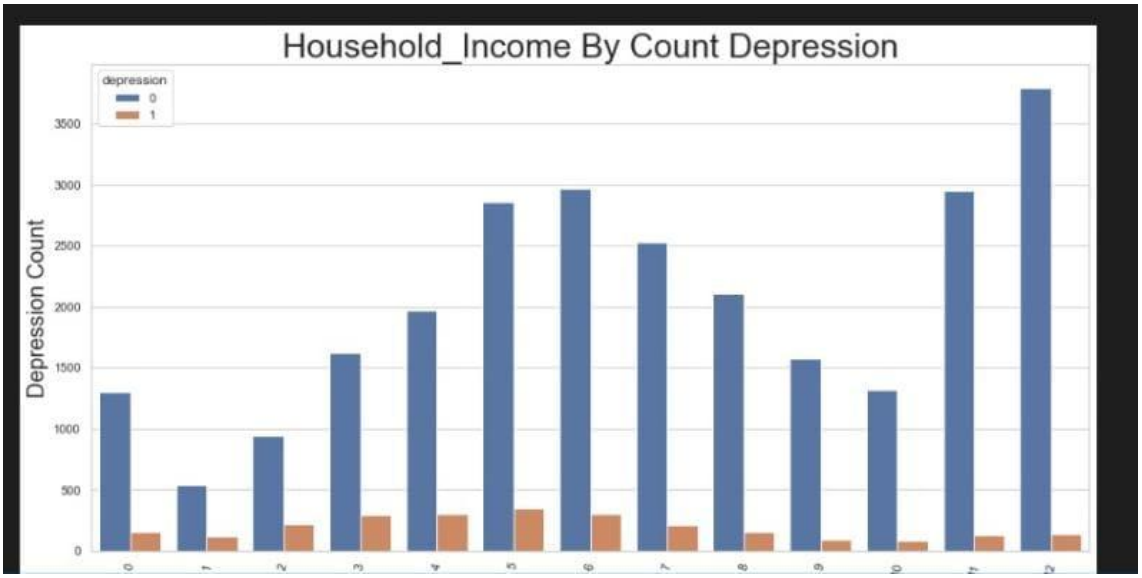
6.1 VISUALIZING DEMOGRAPHIC DATA

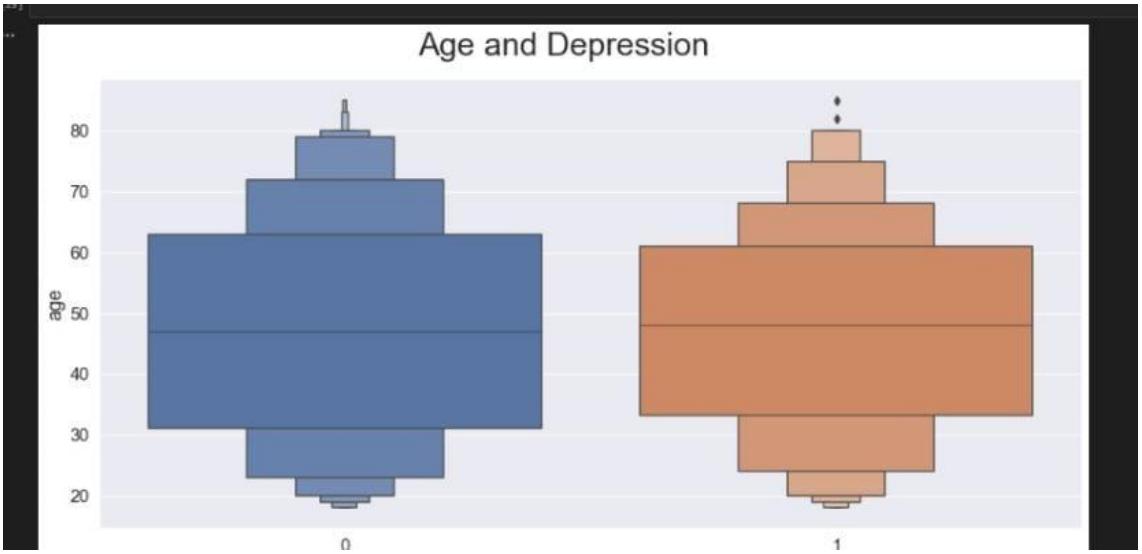




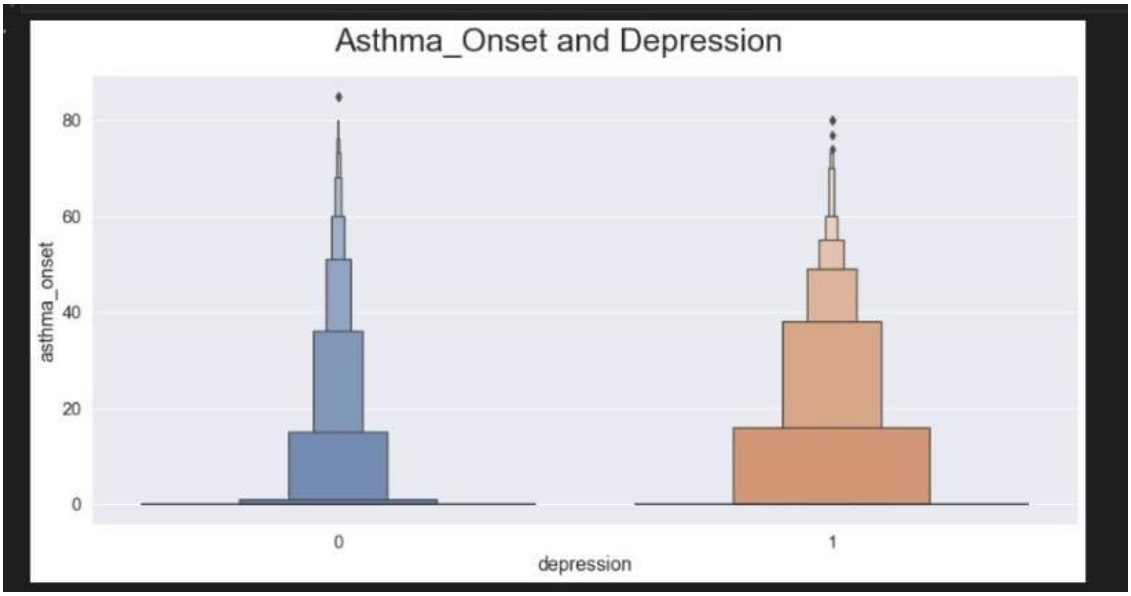

Race By Count Depression
depression
0
1
Depression Count
10000
8000
6000
4000
2000
0
Black
Mexican
Other Hispanic
Other and Multiracial
White
Race

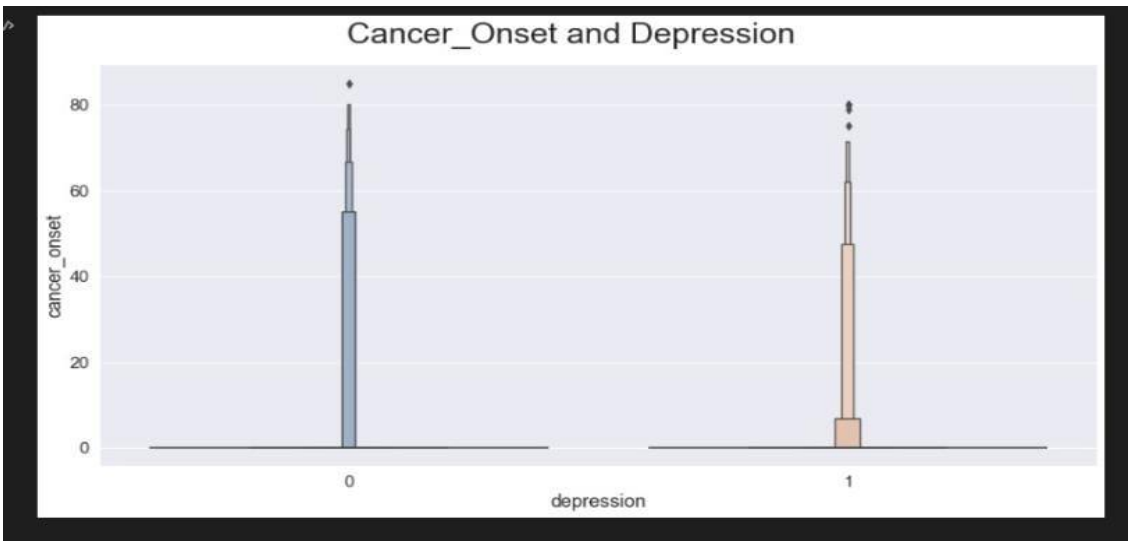

Household_Size By Percent Depression
depression
0
1
Depression Percentage
100
80
60
40
20
0
89.1%
92.1%
91.0%
92.0%
92.2%
91.0%
10.9%
7.9%
9.0%
8.0%
7.8%
9.0%
10.5%
1
2
3
4
5
6
7
Household_Size

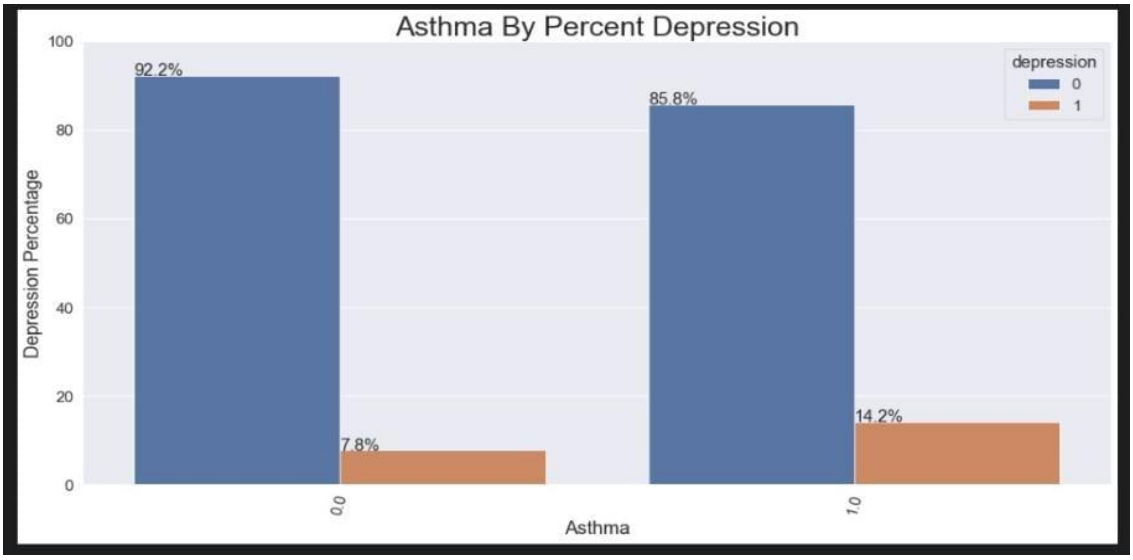

Household_Income By Count Depression
depression
0
1
Depression Count
3500
3000
2500
2000
1500
1000
500
0

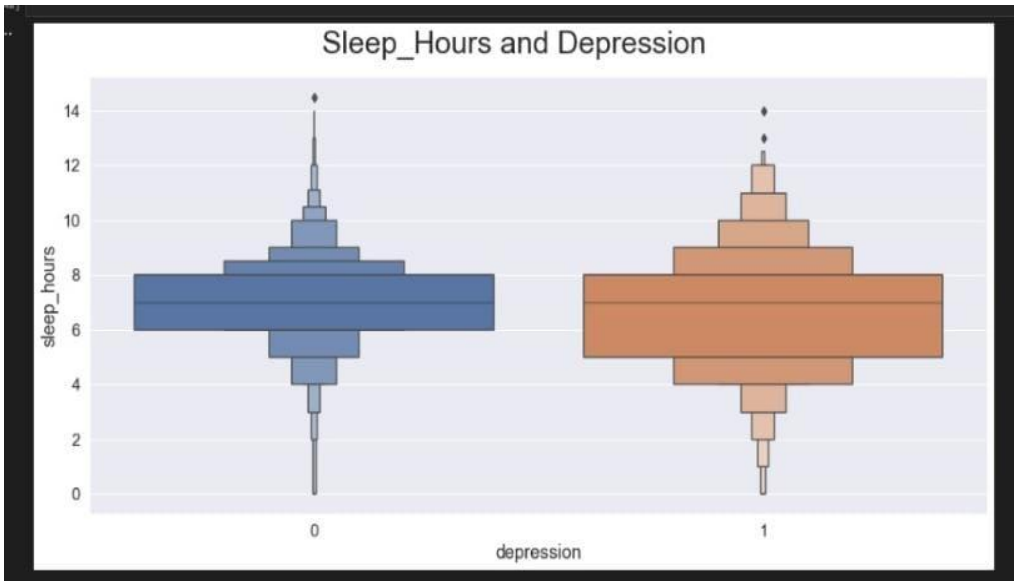

Age and Depression
age
80
70
60
50
40
30
20
0
1

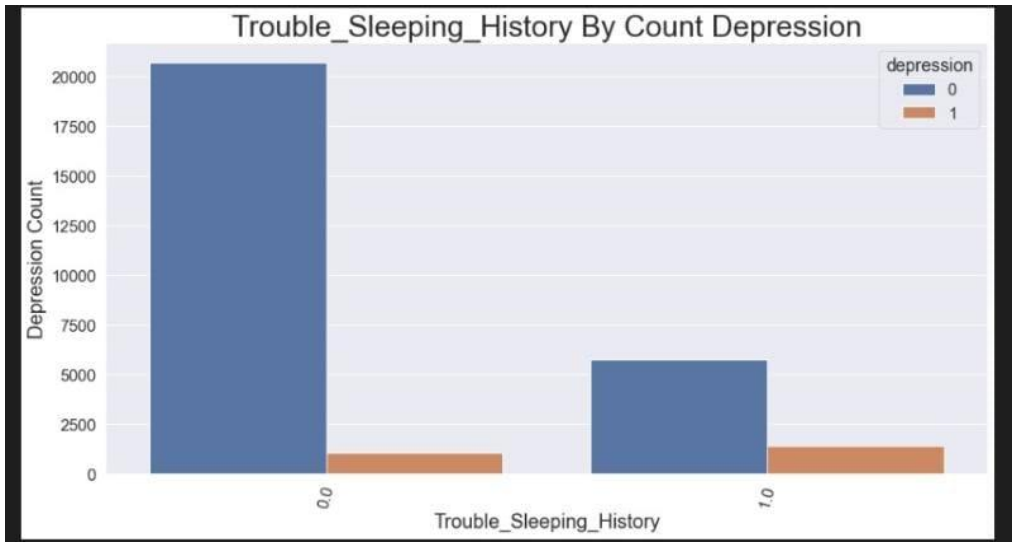
6.2 VISUALIZING MEDICAL CONDITIONS

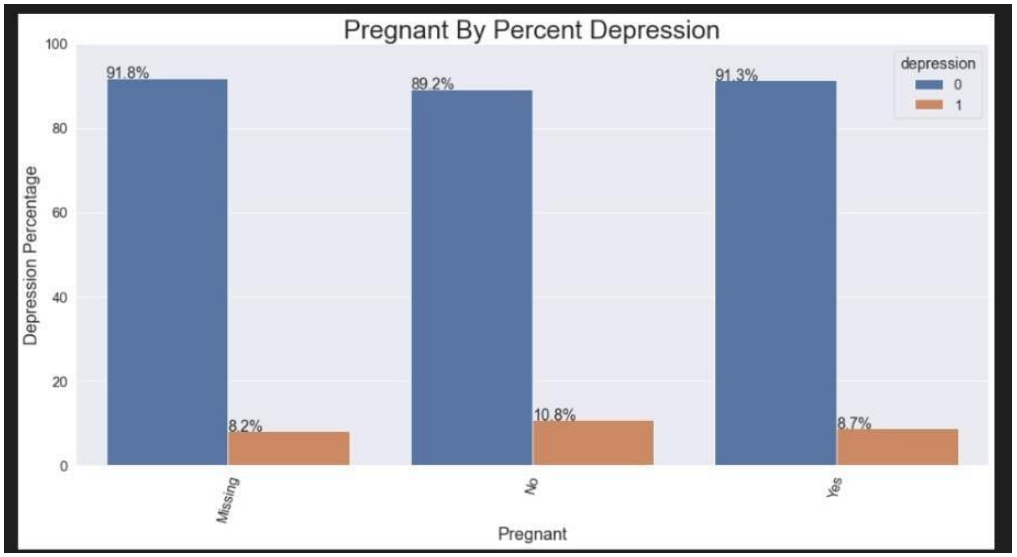




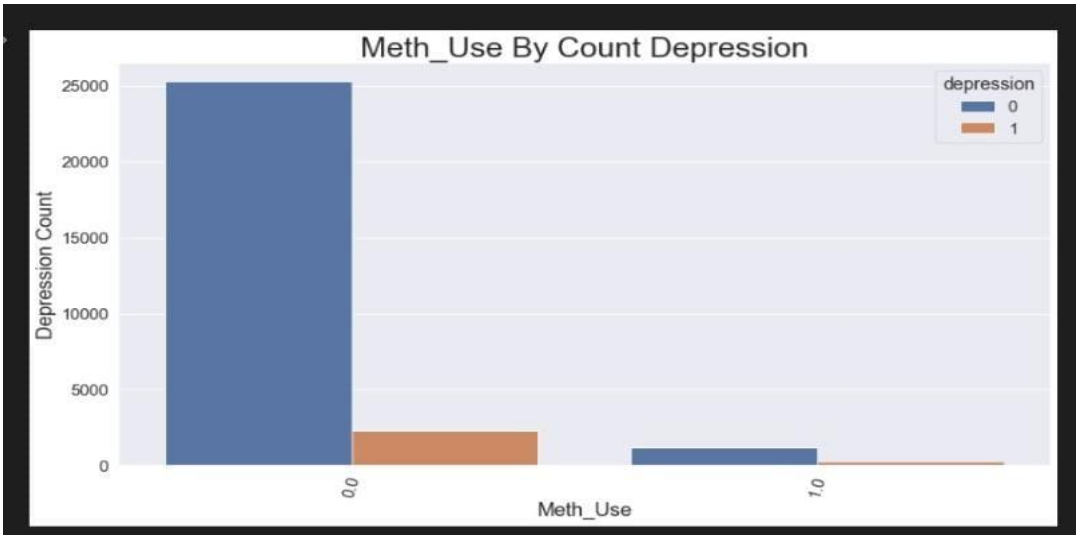


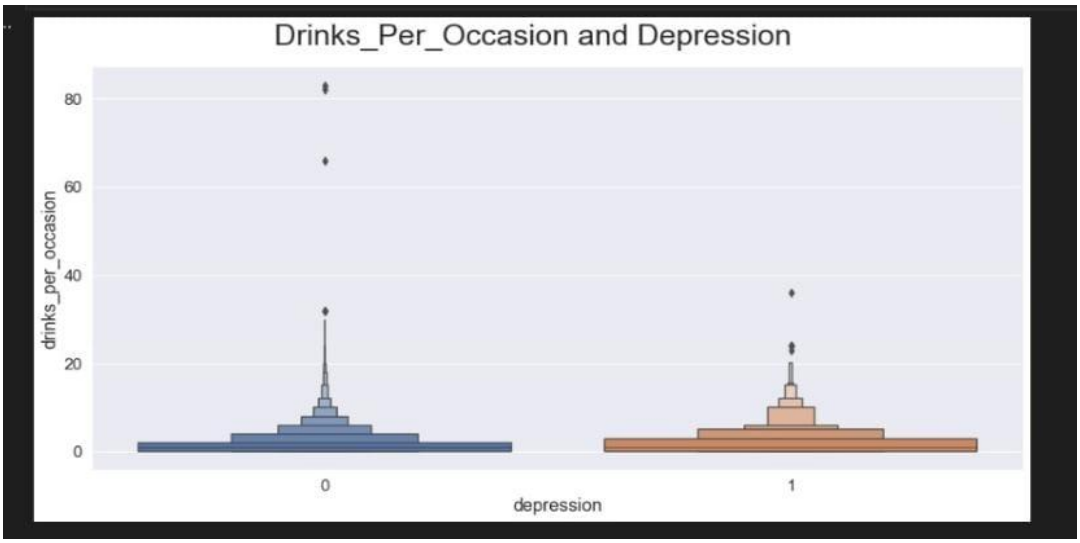

Sleep_Hours and Depression
sleep_hours
depression
0
1
0
2
4
6
8
10
12
14

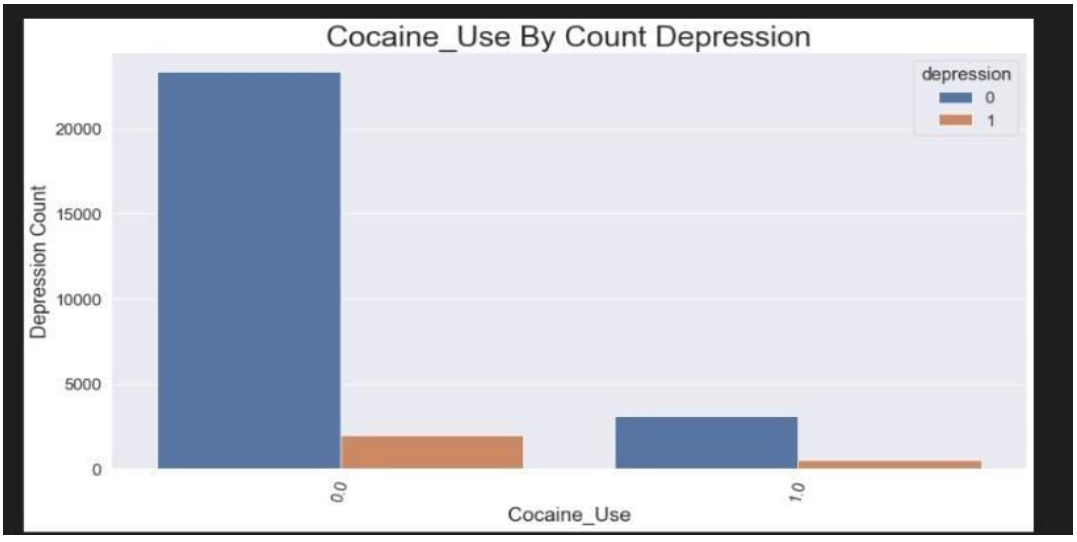

Trouble_Sleeping_History By Count Depression
depression
0
1
Depression Count
Trouble_Sleeping_History
0.0
1.0
0
2500
5000
7500
10000
12500
15000
17500
20000


Pregnant By Percent Depression
depression
0
1
Depression Percentage
Pregnant
Missing
No
Yes
91.8%
8.2%
89.2%
10.8%
91.3%
8.7%
0
20
40
60
80
100

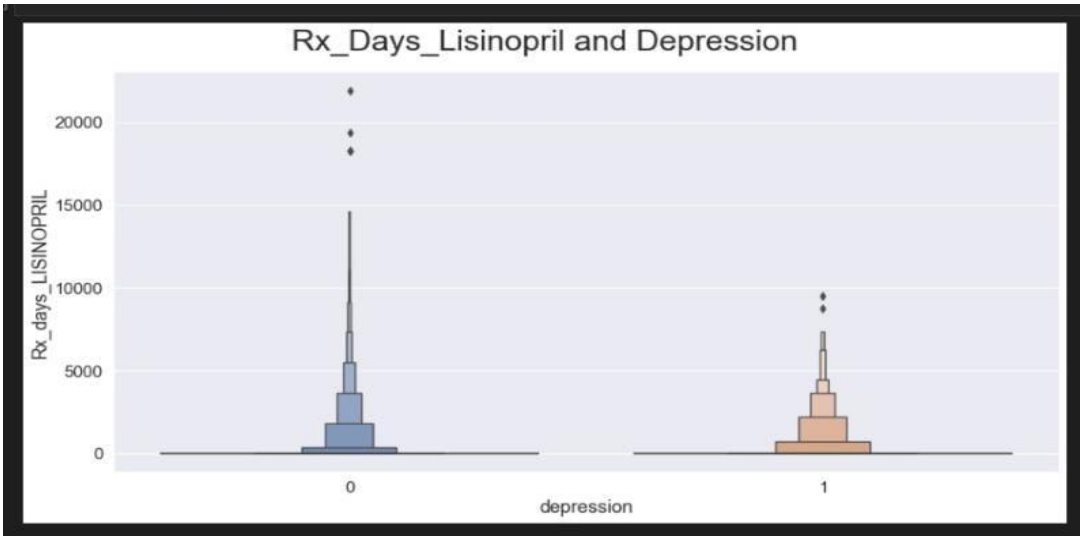
6.3 **VISUALIZING DRUG USAGE**

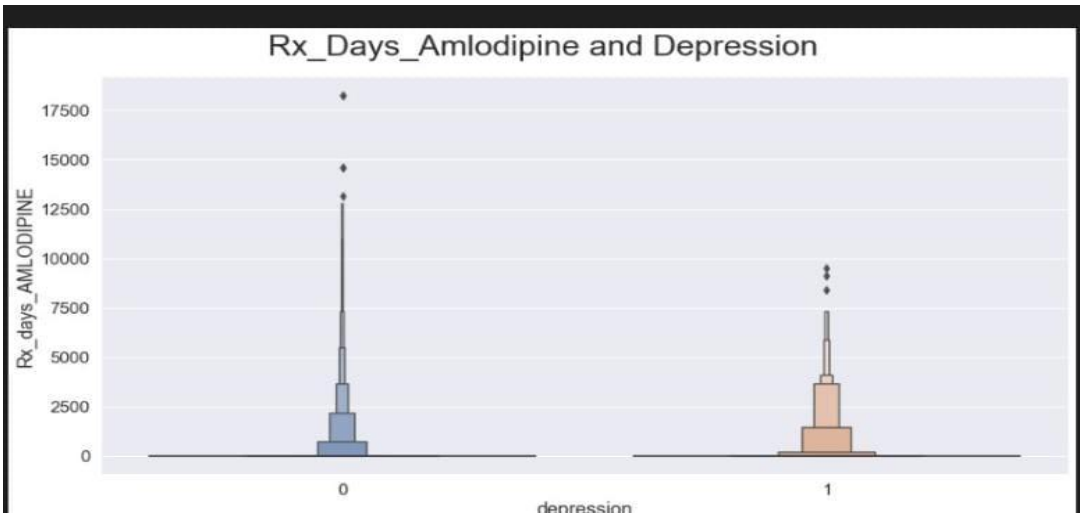


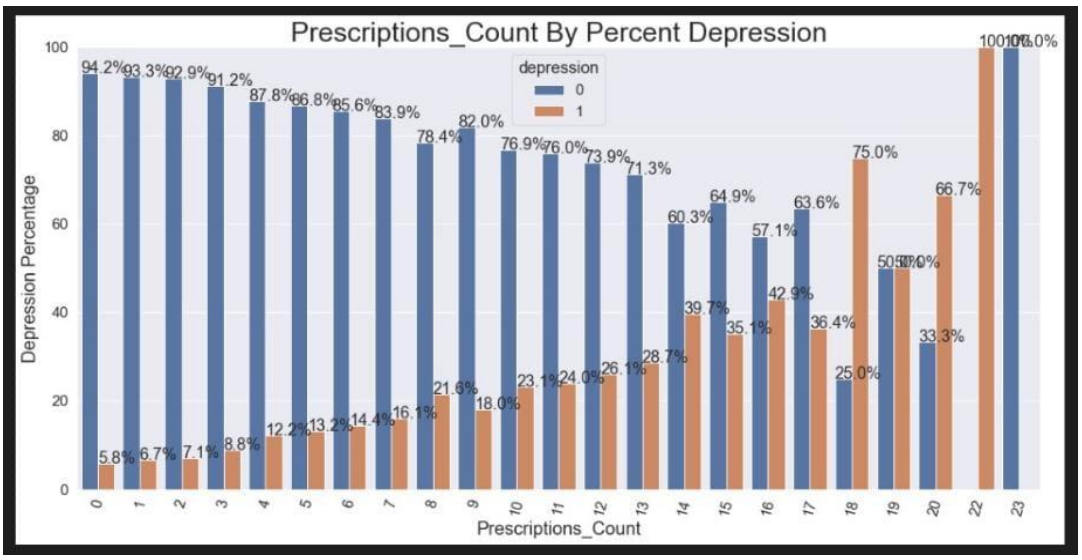


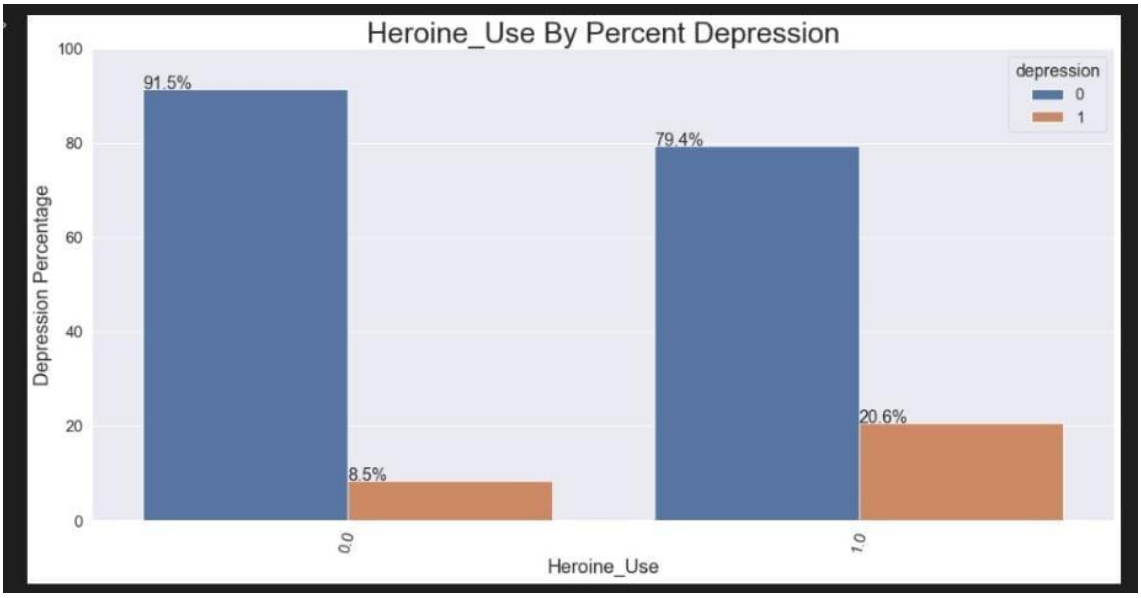


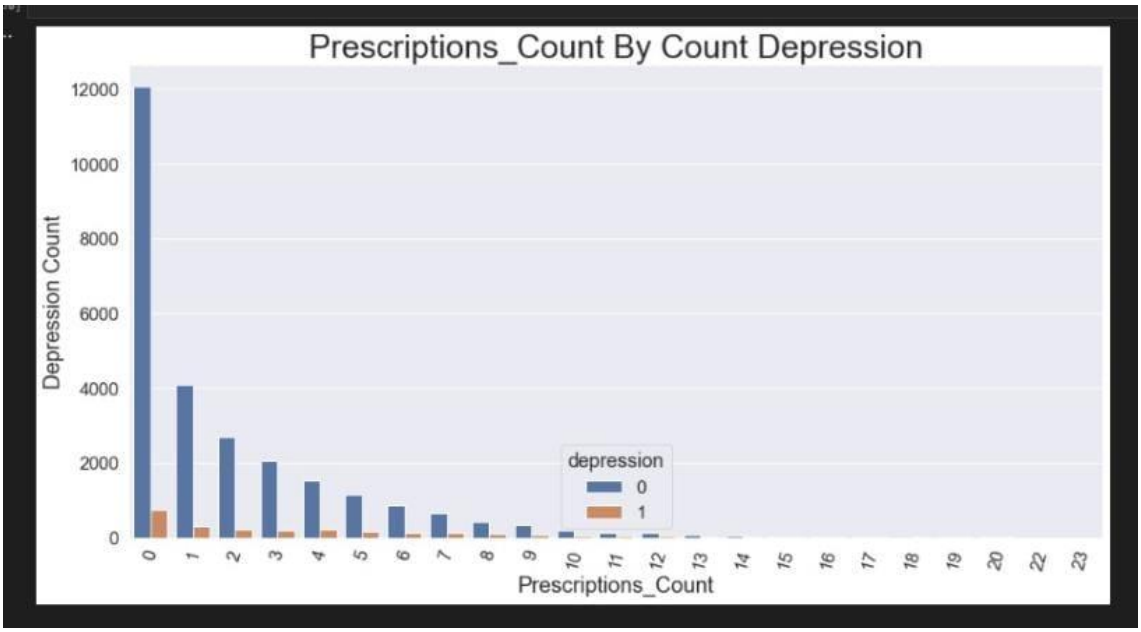
6.4 VISUALIZING PRESCRIPTION MEDICATION



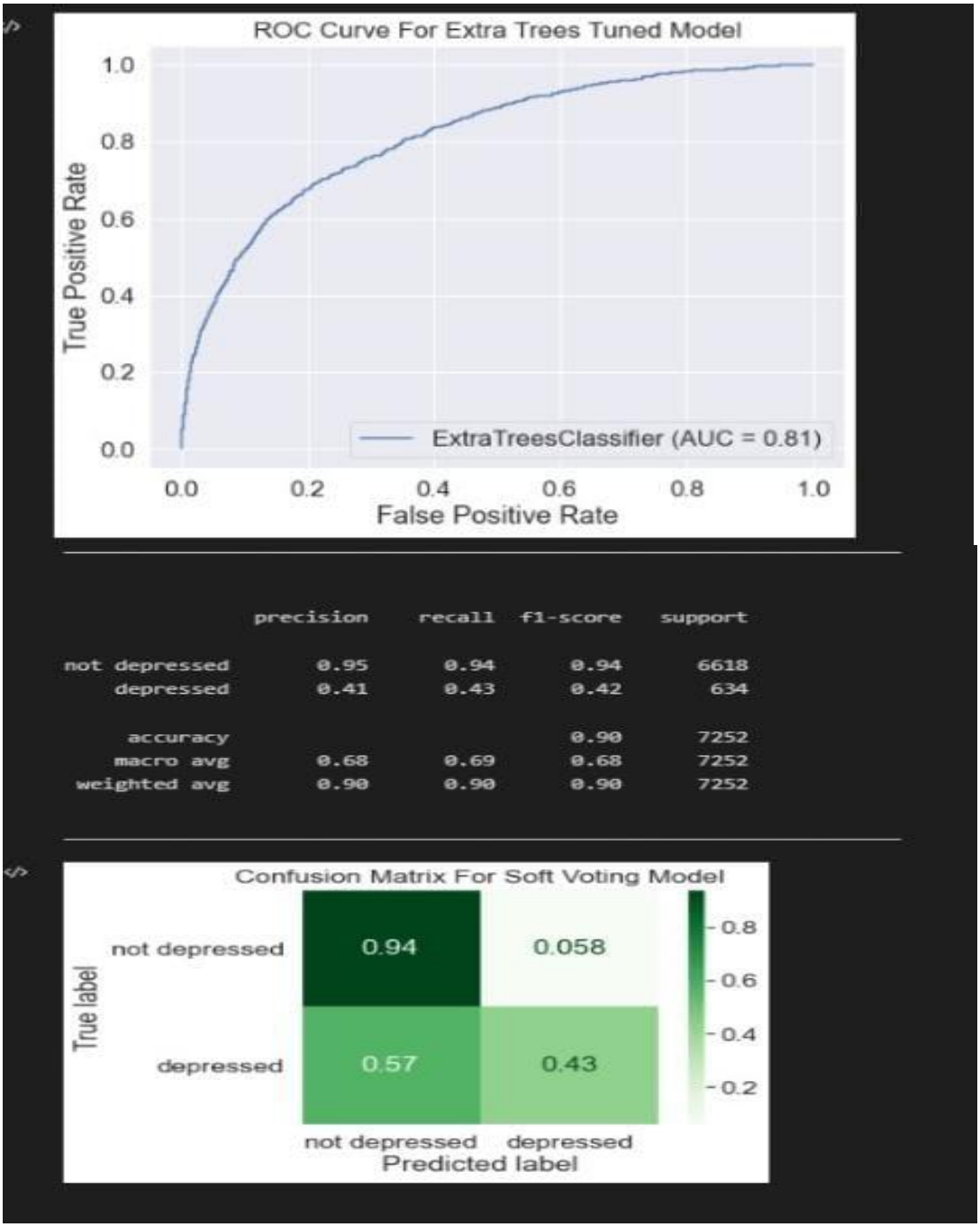




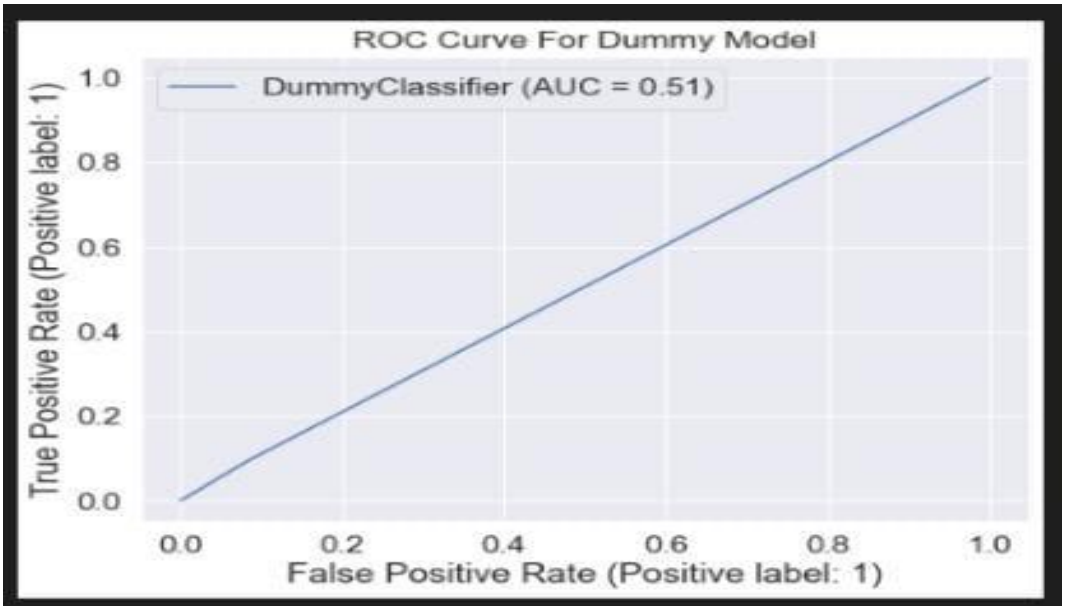

Heroine_Use By Percent Depression
depression
0
1
91.5%
8.5%
79.4%
20.6%
Depression Percentage
Heroine_Use
0.0
1.0


Prescriptions_Count By Count Depression
depression
0
1
Depression Count
Prescriptions_Count

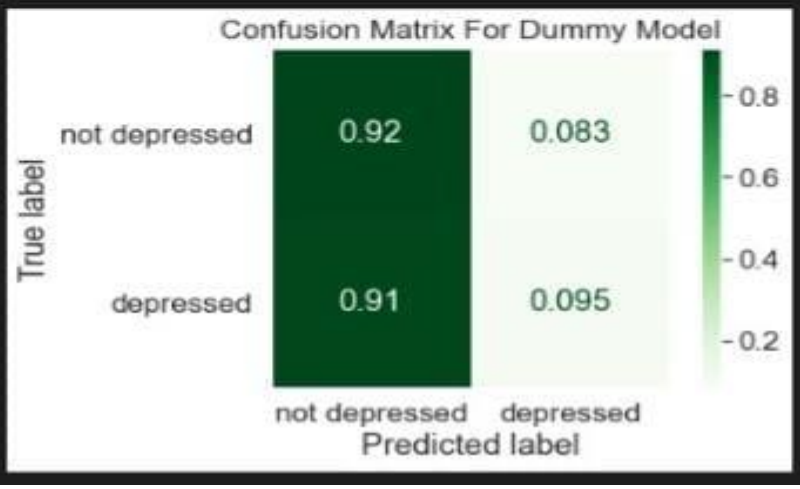
6.5 PERFORMING ML ALGORITHMS

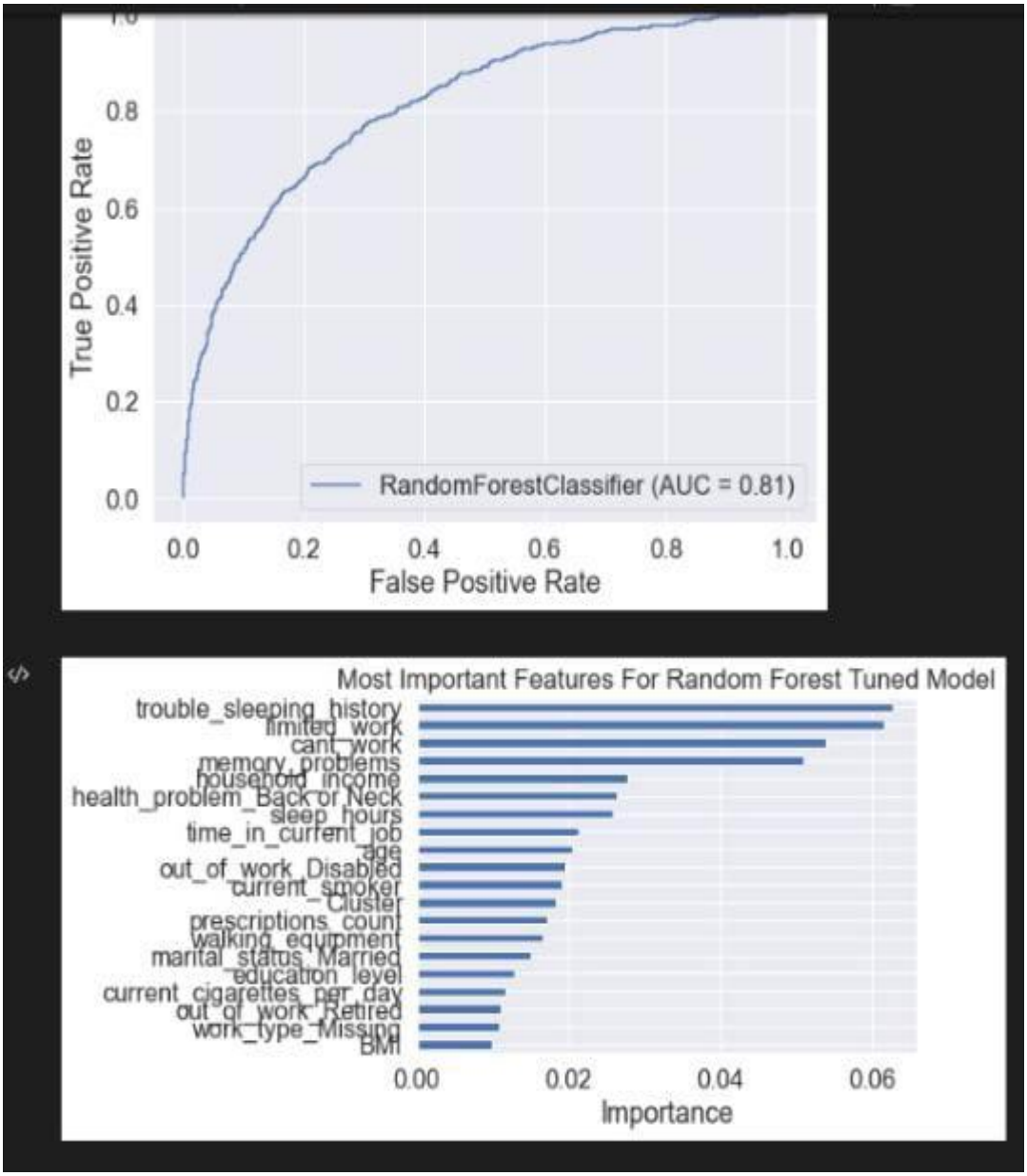


	precision	recall	f1-score	support
not depressed	0.95	0.94	0.94	6618
depressed	0.41	0.43	0.42	634
accuracy			0.90	7252
macro avg	0.68	0.69	0.68	7252
weighted avg	0.90	0.90	0.90	7252

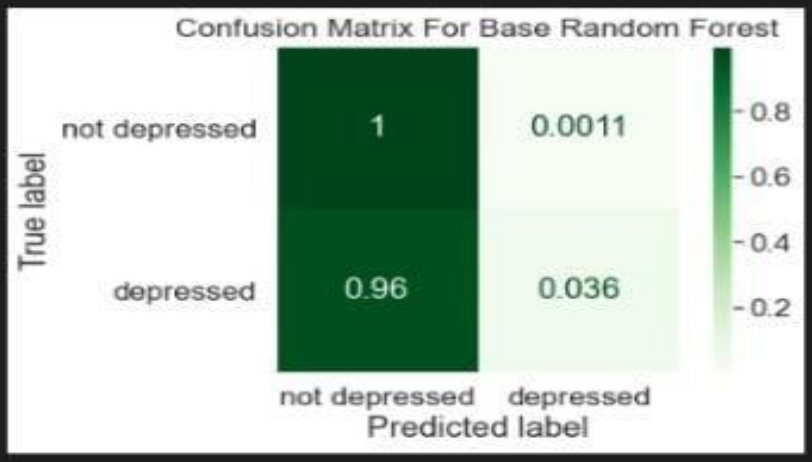


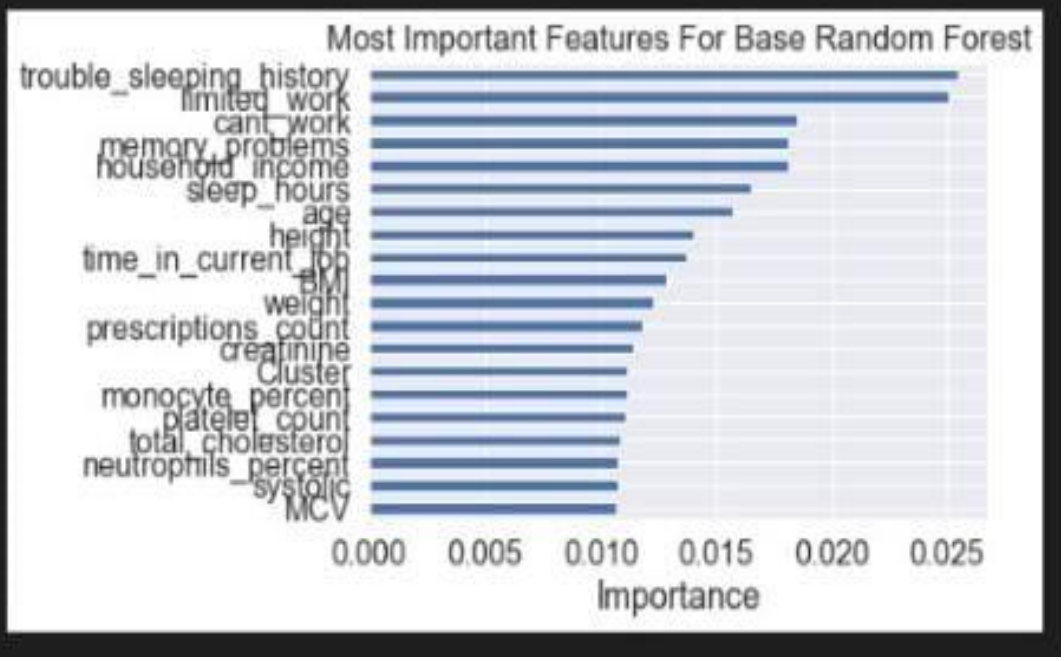
	precision	recall	f1-score	support
not depressed	0.91	0.92	0.92	6618
depressed	0.10	0.09	0.10	634
accuracy			0.84	7252
macro avg	0.51	0.51	0.51	7252
weighted avg	0.84	0.84	0.84	7252

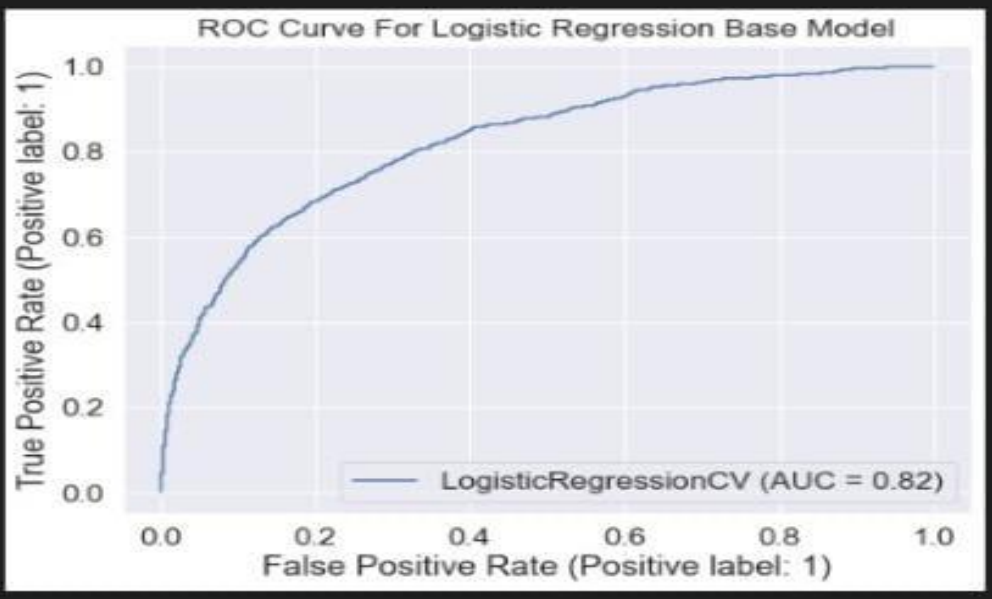



True Positive Rate
0.0
0.2
0.4
0.6
0.8
RandomForestClassifier (AUC = 0.81)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Most Important Features For Random Forest Tuned Model
trouble_sleeping_history
limited_work
cant_work
memory_problems
household_income
health_problem_Back or Neck
sleep_hours
time_in_current_job
age
out_of_work_Disabled
current_smoker
Cluster
prescriptions_count
walking_equipment
marital_status_Married
education_level
current_cigarettes_per_day
out_of_work_Retired
work_type_Missing
BMI
0.00
0.02
0.04
0.06
Importance

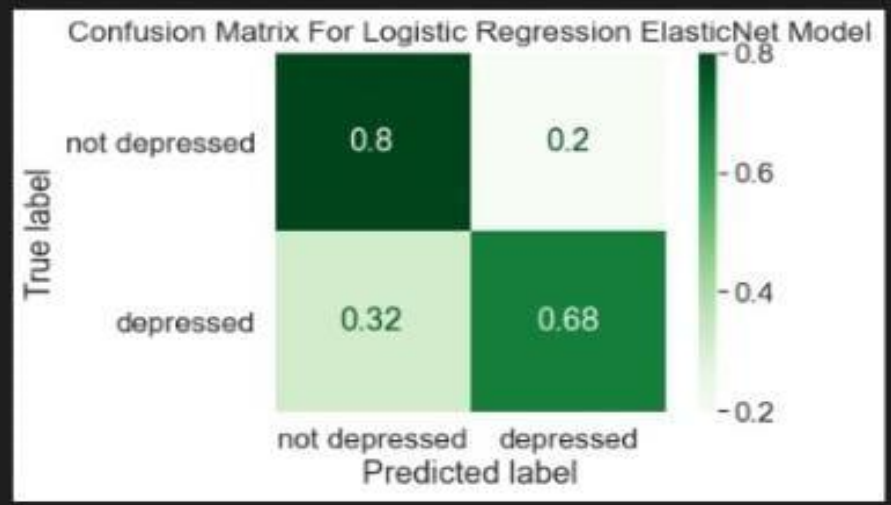
	precision	recall	f1-score	support
not depressed	0.92	1.00	0.96	6618
depressed	0.77	0.04	0.07	634
accuracy			0.91	7252
macro avg	0.84	0.52	0.51	7252
weighted avg	0.90	0.91	0.88	7252



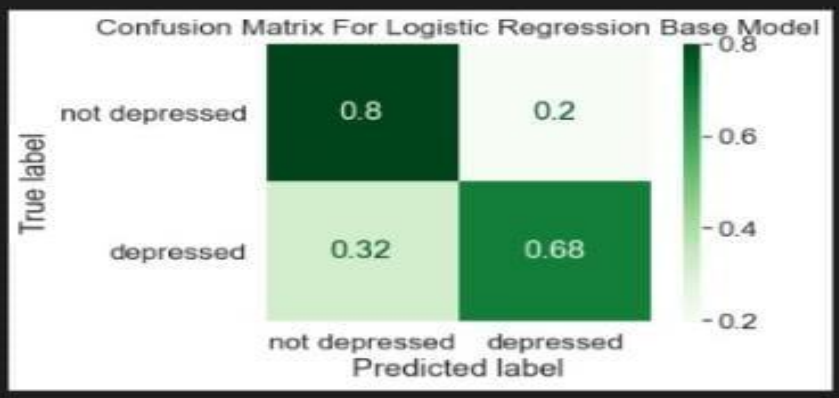




	precision	recall	f1-score	support
not depressed	0.96	0.80	0.88	6618
depressed	0.25	0.68	0.36	634
accuracy			0.79	7252
macro avg	0.61	0.74	0.62	7252
weighted avg	0.90	0.79	0.83	7252

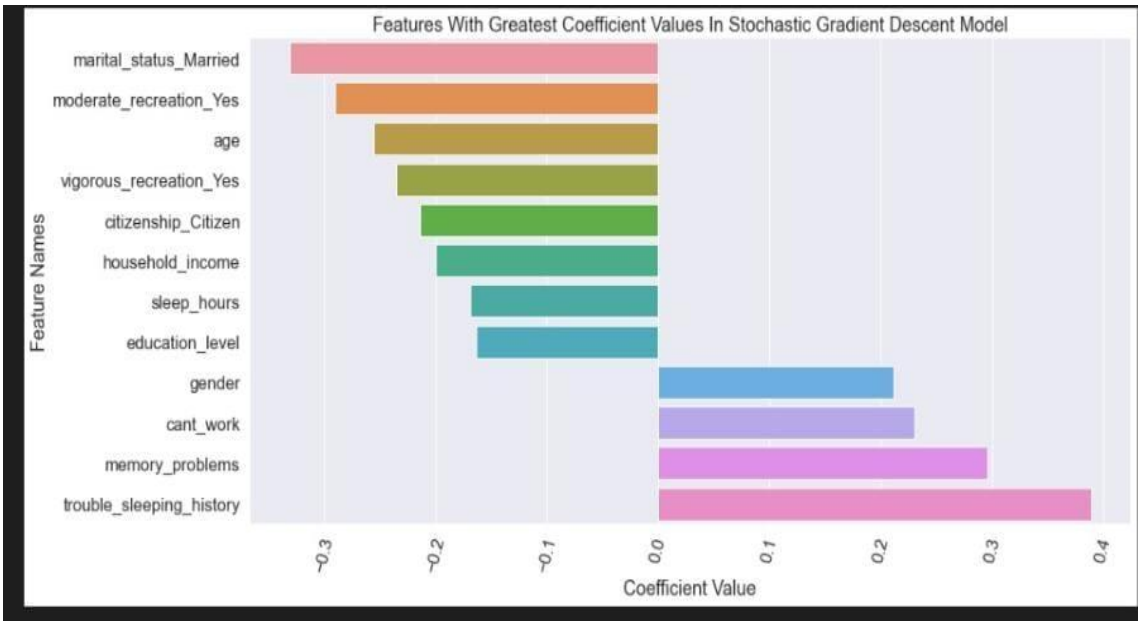


	precision	recall	f1-score	support
not depressed	0.96	0.80	0.88	6618
depressed	0.25	0.68	0.37	634
accuracy			0.79	7252
macro avg	0.61	0.74	0.62	7252
weighted avg	0.90	0.79	0.83	7252



6.6 FINAL MODEL PREDICTION

579	0.035671	0.055073	health_problem_Diabetes
580	0.035863	0.010713	bilirubin
581	0.037038	0.000000	heart_disease
582	0.037938	0.000000	limitations
583	0.039616	0.000000	total_cholesterol
584	0.044446	0.000000	Rx_DIAZEPAM
585	0.046904	0.000000	start_smoking_age
586	0.049162	0.000000	Rx_days_PANTOPRAZOLE
587	0.049349	0.109273	diabetes_relative
588	0.050376	0.000000	Rx_LORAZEPAM
589	0.054371	0.139414	walking_equipment
590	0.056602	0.065773	sedentary_time
591	0.061509	0.151842	prescriptions_count
592	0.064958	0.000000	health_problem_Birth Defect
593	0.068899	0.180388	Rx_GABAPENTIN
594	0.069940	0.145285	household_size
595	0.079865	0.000000	Rx_days_NITROFURANTOIN
596	0.081347	0.214710	marijuana_per_month
597	0.090045	0.333152	current_cigarettes_per_day
598	0.091420	0.000000	Rx_CLONAZEPAM
599	0.098980	0.226381	marijuana_use
600	0.099806	0.154187	Rx_ALPRAZOLAM
601	0.116064	0.334445	health_problem_Back or Neck
602	0.143569	0.306399	ever_overweight
603	0.211841	0.340252	gender
604	0.230622	0.489722	cant_work
605	0.295808	0.911684	memory_problems



7. VALIDATION & TESTING

VALIDATION

Validation is considered a process of evaluating testing datasets with the trained model. It is a set of activities and processes which are designed to make sure that the model is performing at its best. Validation is mostly used to improve the accuracy of the model. This improves the data quality and quantity. We can do validations in different formats depending on the objectives.

TRAINING DATA

Training data is a large dataset used to teach machine learning models. With this training, data is used for the prediction of models. For feature extraction, machine learning algorithms are used. A dataset or data that is used to train a model is considered training data, predictions are done with this model. This kind of approach includes supervised learning. That means we make the model learn how to perform some specific task.

TESTING DATA

Testing data is considered unseen data, using this we can test our models. Testing data is needed once the machine learning model is built. The test data is a part of the training data which is used for the evaluation of predicting the final model. Test data will provide an evaluation in an unbiased way which will fit on training data. Sometimes the data which is in the test dataset has never been used in the training dataset and are considered a holdout dataset.

VALIDATION AND TESTING

- Validation is done on predicted models to increase the level of accuracy.
- The validation dataset is used to improve the machine learning model.
- Machine Learning algorithms are applied to models so that for every model there will be a different accuracy in prediction.
- For this, to avoid inconsistent predictions we use a test dataset. The test dataset is used for testing the accuracy of the model.
- With this prediction of unseen data, the best model can be identified with accurate prediction results.

8. CONCLUSION

Prediction of psychological disorder of a person is implemented by considering several health conditions i.e., by the data which is given by Centers of Disease Control. Categorical data is analyzed in different formats. The original data is trained and then applied to machine learning algorithms like classification, regression, and clustering. Linear models gave better results when compared to other models. Using Stochastic Gradient Descent coefficient values there is a scope for capturing less impactful and most impactful features. Those Top features are visualized again to compare the difference between the base model and the predicted model. The most affected features that cause depression are marital status married, age, household income, trouble sleeping history, can't work, and memory problems. Hence, visualized data of these features are helpful for both psychological practitioners and psychologically disordered persons can get benefitted. Visualization on predicting the psychological disorder of a person using health care data results with the help of the confusion matrix model was able to identify 80% of true negatives and 71% of false positives accurately.

9. FUTURE SCOPE

Furthermore, The prediction of depression is a multidimensional problem also considered a complex problem. we could also try different models, there might be a possibility that other models may provide more features for better prediction. Also, we can add more entries to the dataset so there may be differences in prediction. Evaluation of more markers may provide valuable results. A continuing evaluation may improve model building with much-benefited features.

10. BIBLIOGRAPHY

[1] World health Organization Depression Overview: https://www.who.int/news-room/fact-sheets/detail/depression

[2] National Institute of Mental Health on Major Depression: https://www.nimh.nih.gov/health/statistics/major-depression.shtml

[3] American Psychological Association Briefing Series on the Role of Psychology in Health Care: https://www.apa.org/health/briefs/primary-care.pdf

[4] Association of Integrated Team-Based Care With Health Care Quality, Utilization, and Cost: https://jamanetwork.com/journals/jama/fullarticle/2545685

[5] The PHQ-9: validity of a brief depression severity measure https://pubmed.ncbi.nlm.nih.gov/11556941/

[6] Take the PHQ-9 depression screener online: https://www.mdcalc.com/phq-9-patient-health-questionnaire-9

[7] World Health Organization. The World Health Report 2001: Mental Health: New

Understanding, New Hope (World Health Organization, Switzerland, 2001)

[8] Durstewitz, D., Koppe, G. & Meyer-Lindenberg, A. Deep neural networks in

psychiatry. Mol. Psychiatry 24, 1583–1598 (2019).

[9] Li, X. et al. EEG-based mild depression recognition using convolutional neural

network. Med. Biol. Eng. Comput. 47, 1341–1352 (2019).

[10] MEPS HC-135E: 2010 emergency room visits. Agency for Healthcare Research and

Quality, Center for Financing, Access, and Cost Trends, Rockville, MD; July 2012. http://

meps.ahrq.gov/mepsweb/data_stats/download_data/pufs/h135e/h135edoc.pdf, Accessed

2012.

[11] Birnbaum HG, Kessler RC, Kelley D, Ben-Hammadi R, Joish VN, Greenberg PE.

Employer burden of mild, moderate, and severe major depressive disorder: mental

health services utilization and costs, and work performance. Depress Anxiety 2010; 27

[14] L. Pan, G. Liu, X. Mao, H. Li, J. Zhang, H. Liang, and X.

Li, Development of Prediction Models Using Machine Learning Algorithms for Girls with Suspected Central Precocious Puberty: Retrospective Study, JMIR medical informatics 7 (2019), e11728-e11728.

[15] WHO. Preventing suicide: A global imperative. http://www.who.int/ mental_health/suicide-prevention/world_report_2014/en/ 2014. Accessed January 5, 2018.

[16] WHO. National Aged Standardised Suicide Rates 2015. http://www. who.int/gho/mental_health/suicide_rates/en/. Accessed January 5, 2018.

[17] Liu KY. Suicide rates in the world: 1950-2004. Suicide Life Threat Behav. 2009;39(2):204-213.

[18] Ajdacic-Gross V, Weiss MG, Ring M, et al. Methods of suicide: international suicide patterns derived from the WHO mortality database. Bull World Health Organ. 2008;86(9):726-732.

[19] Kreitman N. The coal gas story. United Kingdom suicide rates, 1960- 71. Br J Prev Soc Med. 1976;30(2):86-93. 6. Large MM, Nielssen OB. Suicide in Australia: meta-

analysis of rates and methods of suicide between 1988 and 2007. Med J Aust. 2010;192(8):432- 437.

[20] Gunnell D, Knipe D, Chang SS, et al. Prevention of suicide with regulations aimed at restricting access to highly hazardous pesticides: a systematic review of the international evidence.

[21] Kerr WC, Subbaraman M, Ye Y. Per capita alcohol consumption and suicide mortality in a panel of US states from 1950 to 2002. Drug Alcohol Rev. 2011;30(5):473-480.

[22] Niederkrotenthaler T, Fu KW, Yip PS, et al. Changes in suicide rates following media reports on celebrity suicide: a meta-analysis. J Epidemiol Community Health. 2012;66(11):1037-1042.

[23] Tondo L, Albert MJ, Baldessarini RJ. Suicide rates in relation to health care access in the United States: an ecological study. J Clin Psychiatry. 2006;67(4):517-523.

[24] Fountoulakis KN, Kawohl W, Theodorakis PN, et al. Relationship of suicide rates to economic variables in Europe: 2000-2011. Br J Psychiatry. 2014;205(6):486-496.

[25] King M, Semlyen J, Tai SS, et al. A systematic review of mental disorder, suicide, and deliberate self harm in lesbian, gay and bisexual people.