Project Report
on
Depressed Patient Classification

IE7275 – Data Mining in Engineering

Submitted by: Group 39

Team Members

Ratna Manjeera Grandhi

Praneeth Korukonda

Term and Year: Spring 2023

Submitted to: Prof. Sagar Kamarthi

Submitted Date: April 21st, 2023

ACKNOWLEDGEMENT

Apart from the efforts of team, the completion of this project wouldn‟t have been possible without the inputs and encouragement of the Teaching Assistants. We also take this opportunity to express our gratitude to Professor Sagar Kamarthi. This wouldn‟t have been materialized without his support and the knowledge he shared with us throughout the semester.

Our heartfelt appreciation goes to Northeastern University – College of Engineering for givingus this opportunity to test our knowledge and use it for practical purposes.

Contents

Depression Disorder

Project Setting:

Depression is a common but serious mood disorder. It causes severe symptoms that affect how you feel, think, and handle daily activities, such as sleeping, eating, or working. Depression can happen at any age, but often begins in adulthood. Depression is now recognized as occurring in children and adolescents, although it sometimes presents with more prominent irritability than low mood. Many chronic mood and anxiety disorders in adults begin as high levels of anxiety in children. The severity of a depression is determined by the quantity of symptoms, their seriousness and duration, as well as the consequences on social and occupational function. Depressions are also common in bipolar disorder, another severe psychiatric disorder. The main difference between uni-polar depression and bipolar disorder is the periodic occurrence of mania in the latter, a state associated with inflated self-esteem, impulsivity, increased activity, reduced sleep and goal-directed actions. Both diseases are genetic disorders and might be understood as a genetic vulnerability to the environment disturbing the internal biological state and potentially trigger mood episodes.

Problem Definition:

We present the analysis of a unique dataset containing sensor data collected from patients suffering from depression. The dataset contains motor activity recordings of 23 unipolar and bipolar depressed patients and 32 healthy controls. Our project classifies if a patient is depressed or not depressed.

Data Source & Data Description:

We acquired data from Kaggle, and datasets simulation

https://www.kaggle.com/datasets/arashnic/the-depression-dataset
https://datasets.simula.no/depresjon/#dataset-details

Our data set contains 56 dataset points consisting of 23 conditions and 32 controls. Within those each condition contains thousands of data evaluation points and same with controls containing thousands of data evaluation points. To summarize data and club the datasets together we will categorize the data by scores. The various columns/factors that give us the score are number (patient identifier), days (number of days of measurements), gender (1 or 2 for female or male), age (age in age groups), afftype (1: bipolar II, 2: unipolar depressive, 3: bipolar I), melanch (1: melancholia, 2: no melancholia), inpatient (1: inpatient, 2: outpatient), edu (education grouped in years), marriage (1: married or cohabiting, 2: single), work (1: working or studying, 2: unemployed/sick leave/pension), madrs1 (MADRS score when measurement started), madrs2 (MADRS when measurement stopped). Predictors used are numerical values which are mainly from madrs 1 and madrs 2 scores, actigraph (sensor data). Actigraph dataset are the main objective type used for observing depression. Study will also include data from participants sleep pattern analysis of depressed and non-depressed candidates. The dataset will also be consumed in analyzing machine learning methods and cost sensitive classification. Upon feasibility, the same dataset will be suitable for comparing different machine learning classification models.

we have two types of datasets for depression classification - one is called the "score" dataset and the other is called the "control and condition" dataset. In the score dataset, we have numbers that represent how severe a person's depression symptoms are. In the control and condition dataset, we have two groups of people - one group has been diagnosed with depression (called the "condition" group) and the other group does not have depression (called the "control" group), I am showing the data description of Both the data sets below.

As we have described above this is one of the data descriptions for one control subject and condition subject

Data Processing:

For better data visualization we have altered specific columns as shown below

```
# Altering the categories for specific columns based on the data
scores.gender = scores['gender'].cat.rename_categories({1.0 : 'female',2.0:'male'})
scores.afftype = scores['afftype'].cat.rename_categories({1.0 : 'bipolar II',2.0:'unipolar', 3: 'bipolar I'})
scores.melanch = scores['melanch'].cat.rename_categories({1.0 : 'melancholia',2.0:'No melancholia'})
scores.inpatient = scores['inpatient'].cat.rename_categories({1.0 : 'inpatient',2.0:'outpatient'})
scores.marriage = scores['marriage'].cat.rename_categories({1.0 : 'married/cohabitating',2.0:'single'})
scores.work = scores['work'].cat.rename_categories({1.0 : 'Work/Study',2.0:'Unemployed/Sick leave/Pension'})
scores.head()
```

	number	days	gender	age	afftype	melanch	inpatient	edu	marriage	work	madrs1	madrs2	MADRS_dELTA
0	condition_1	11	male	35-39	unipolar	No melancholia	outpatient	6-10	married/cohabitating	Unemployed/Sick leave/Pension	19.0	19.0	0.0
1	condition_2	18	male	40-44	bipolar II	No melancholia	outpatient	6-10	single	Unemployed/Sick leave/Pension	24.0	11.0	-13.0
2	condition_3	13	female	45-49	unipolar	No melancholia	outpatient	6-10	single	Unemployed/Sick leave/Pension	24.0	25.0	1.0
3	condition_4	13	male	25-29	unipolar	No melancholia	outpatient	11-15	married/cohabitating	Work/Study	20.0	16.0	-4.0
4	condition_5	13	male	50-54	unipolar	No melancholia	outpatient	11-15	single	Unemployed/Sick leave/Pension	26.0	26.0	0.0

Figure 1: Data Processing

Correlation between the variables:

Here we are depicting the correlation between the variables in the scores data set which describes the strength and direction of the variables relationship

```
[32] df.corr()
```

	days	gender	afftype	melanch	inpatient	marriage	work	madrs1	madrs2
days	1.000000	-0.073955	-0.184464	0.010070	-0.223032	0.390647	0.045526	0.122678	0.027726
gender	-0.073955	1.000000	0.064884	0.253629	0.175655	0.038169	-0.079259	-0.104832	-0.303410
afftype	-0.184464	0.064884	1.000000	-0.089984	-0.259938	-0.143091	0.053060	-0.132523	0.088863
melanch	0.010070	0.253629	-0.089984	1.000000	0.397360	0.253629	-0.076472	-0.202332	-0.270243
inpatient	-0.223032	0.175655	-0.259938	0.397360	1.000000	-0.293590	-0.204124	-0.523501	-0.410235
marriage	0.390647	0.038169	-0.143091	0.253629	-0.293590	1.000000	0.146077	0.447590	0.150551
work	0.045526	-0.079259	0.053060	-0.076472	-0.204124	0.146077	1.000000	0.281104	0.027913
madrs1	0.122678	-0.104832	-0.132523	-0.202332	-0.523501	0.447590	0.281104	1.000000	0.653090
madrs2	0.027726	-0.303410	0.088863	-0.270243	-0.410235	0.150551	0.027913	0.653090	1.000000

Figure 2: Correlation

Heatmap to identify the strength of the relationships between variables

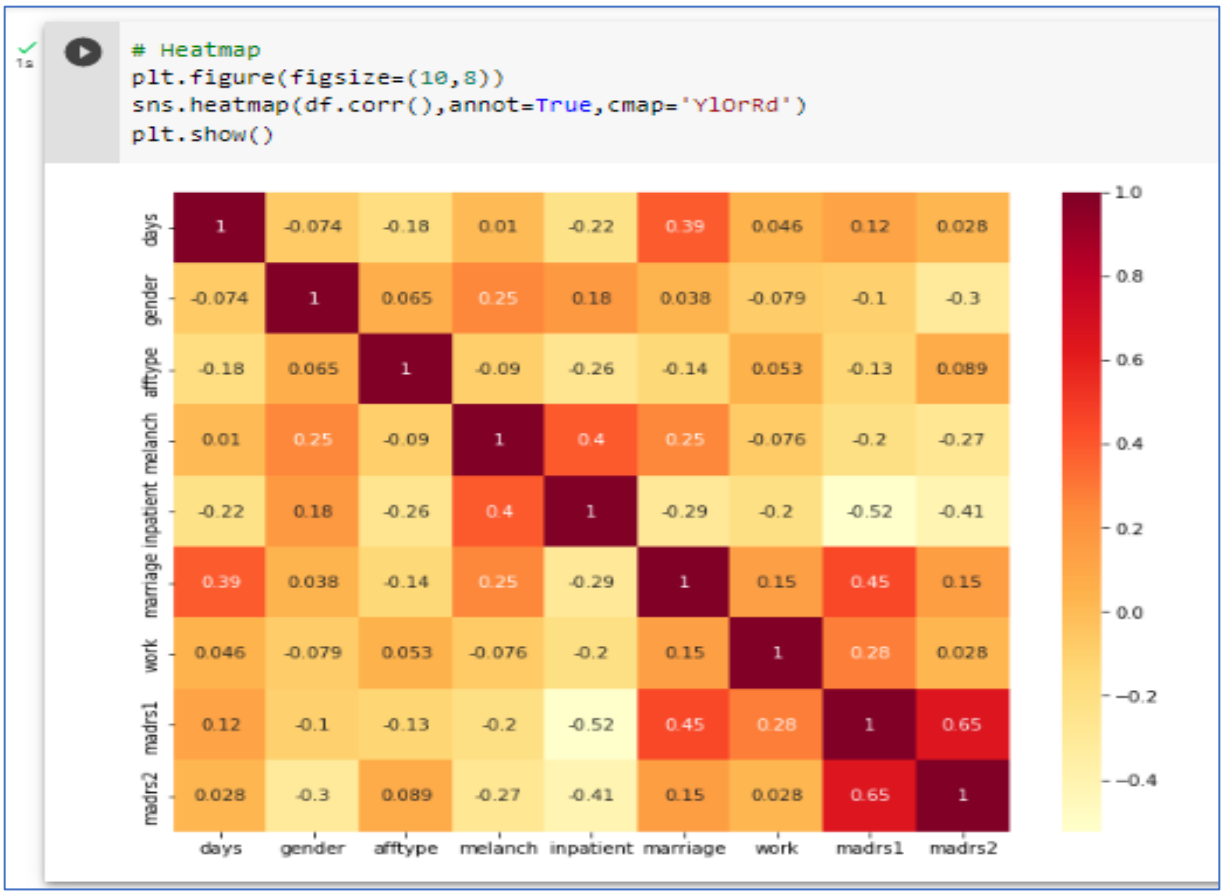

Figure 3: Heat Mapping

To understand and derive insights from the data we have plotted these plots

This is the box plot between MADRS 1 and all the variables

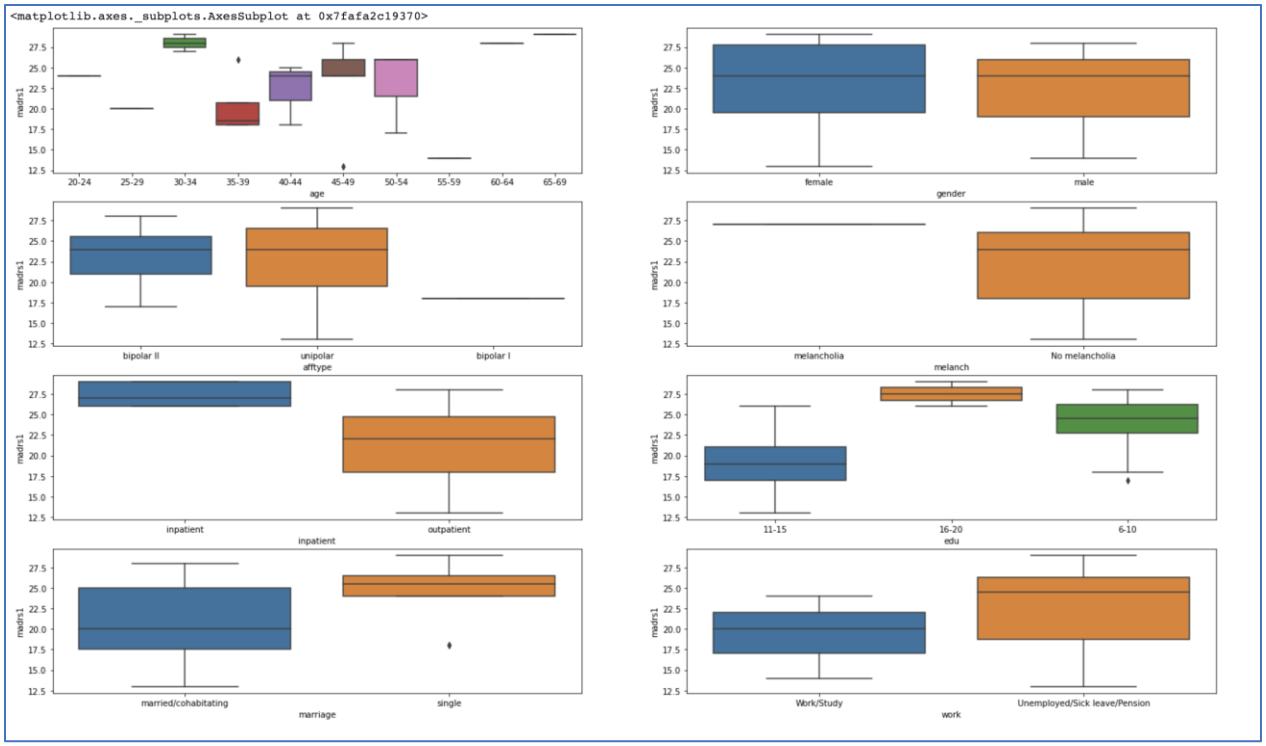

Figure 4: Box plot MADRS1

This is the Violon plot between MADRS 1 and all the variables

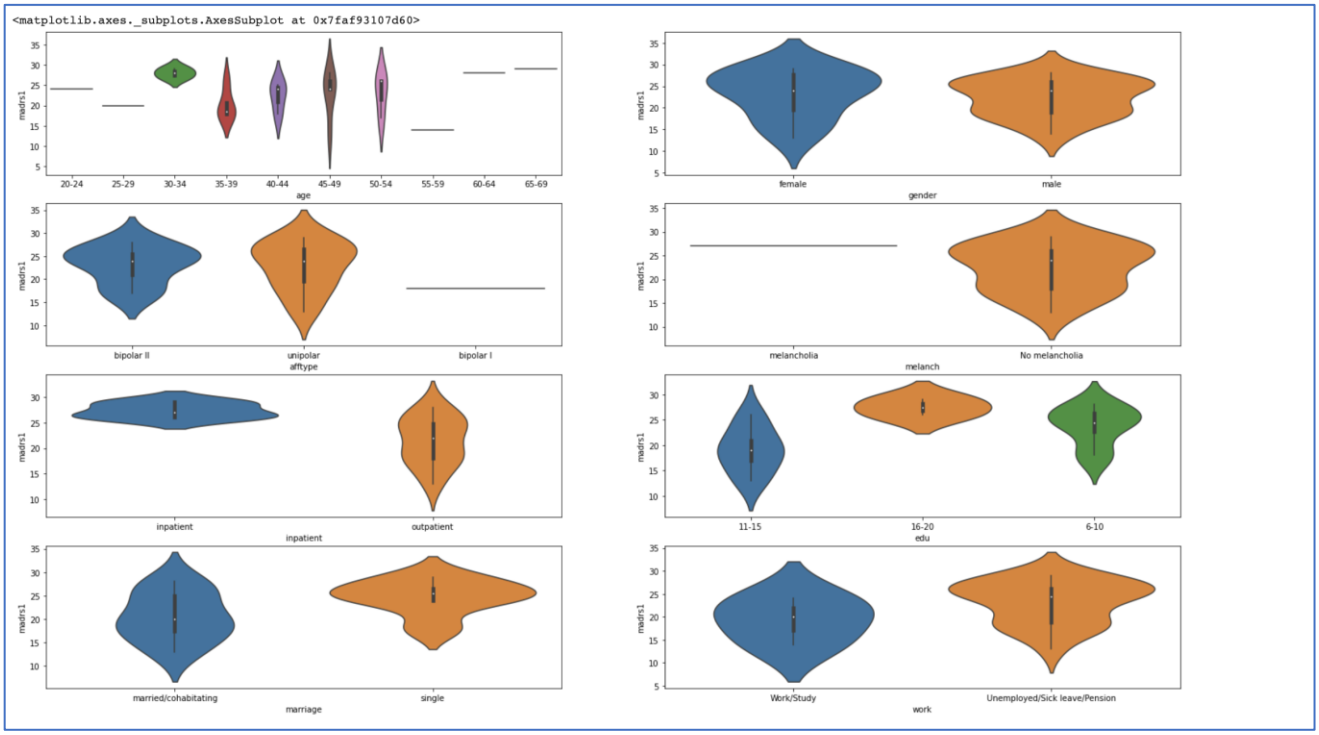

Figure 5: Violon Plot MADRS1

This is the box plot between MADRS 2 and all the variables

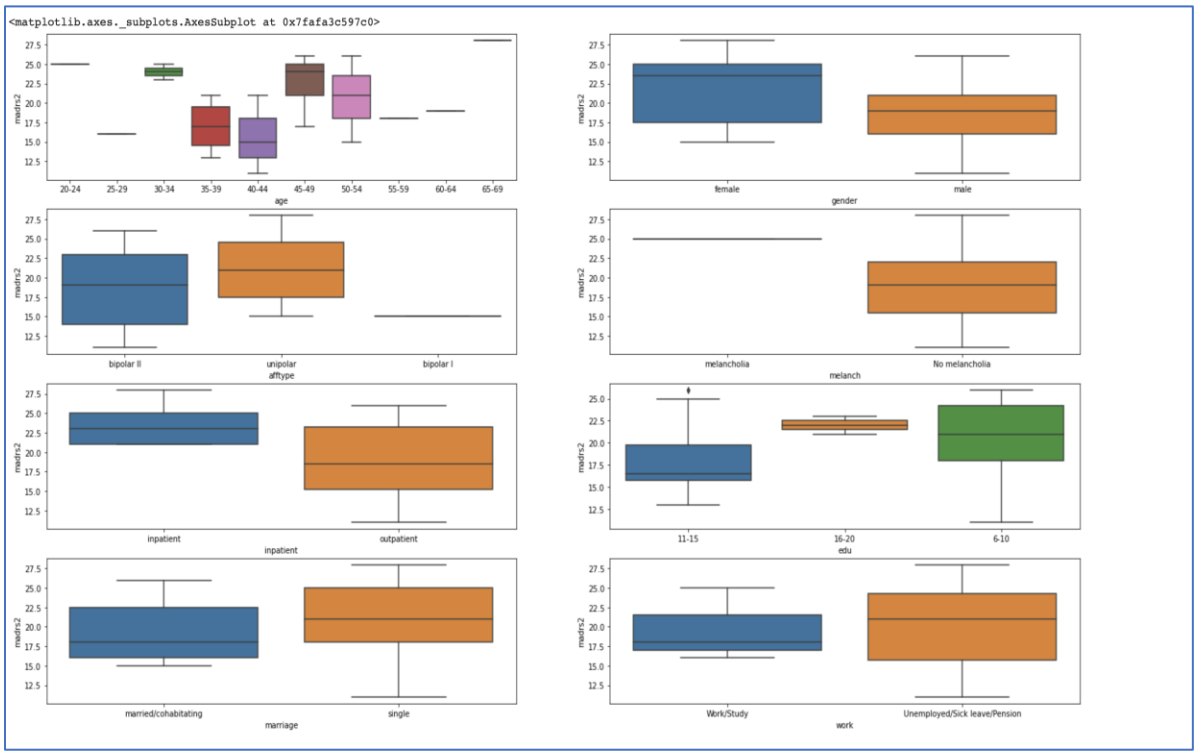

Figure 6: Box plot MADRS2

This is the Violon plot between MADRS 2 and all the variables

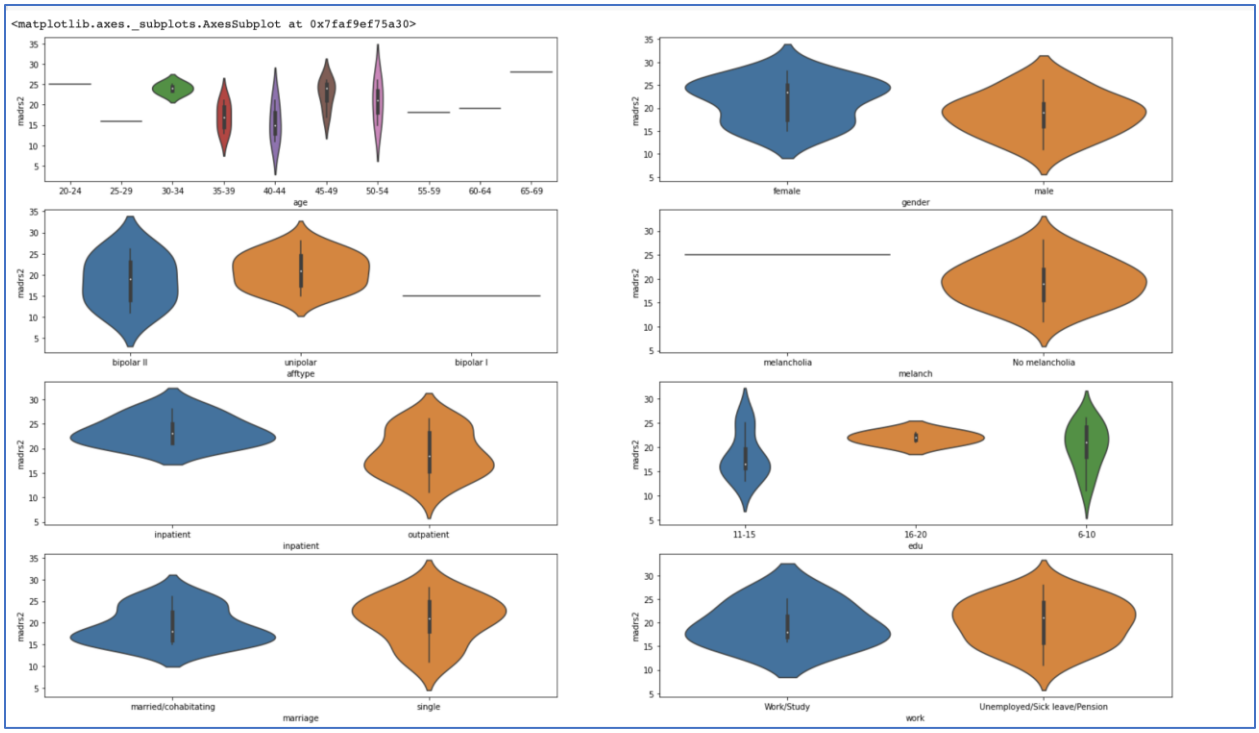

Figure 7: Violon plot MADRS2

Pair Plots:

Pair plots are a way to visualize relationships between different variables in a dataset. They show scatterplots of each possible combination of variables, as well as histograms of each variable on its own. This helps us see patterns and trends in the data, identify outliers or problems with the data, and explore relationships between variables that may not be obvious otherwise.

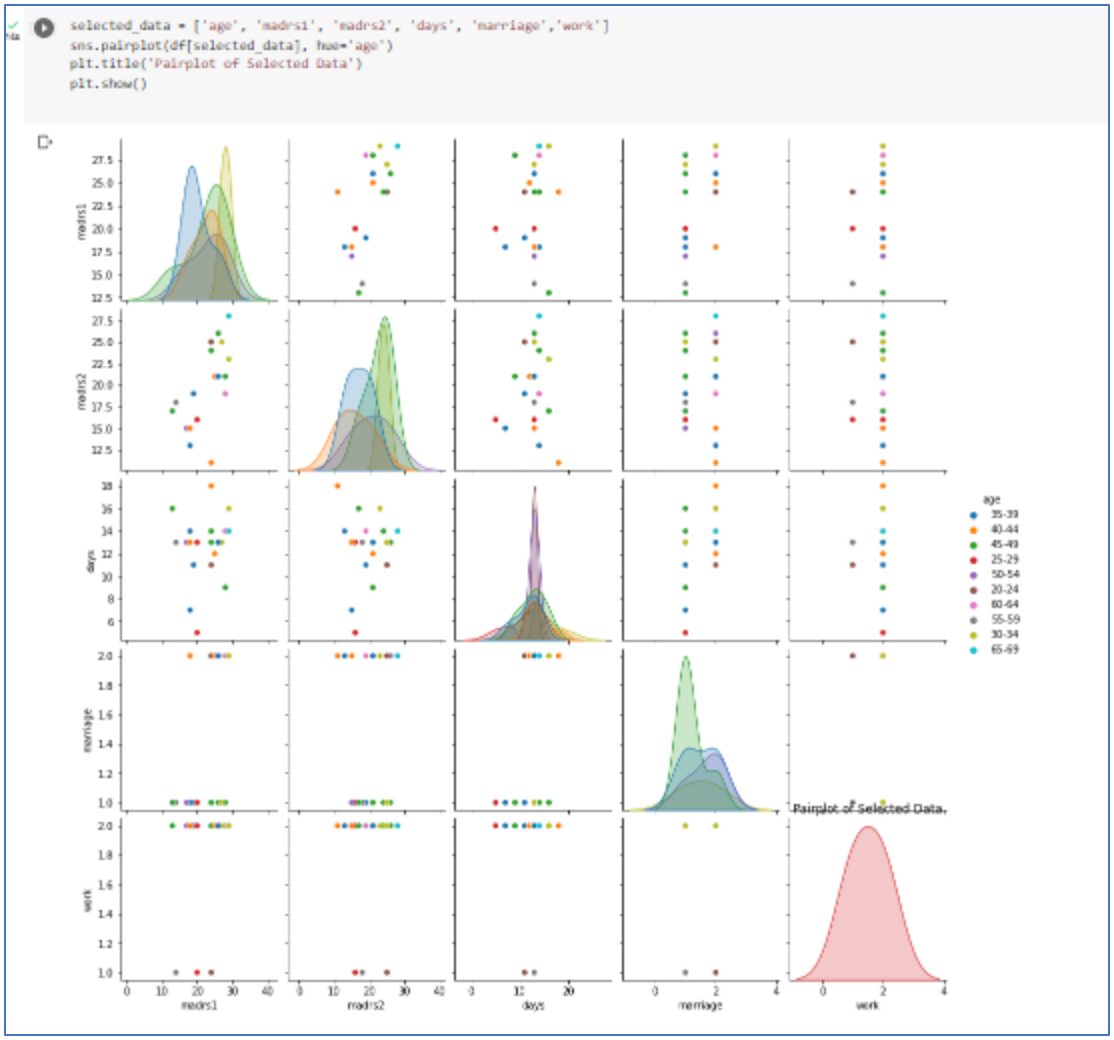

Figure 8: Pair plot MADRS1

To further understand the features we plotted a pair plot for the variables 'days', 'gender', 'afftype', 'melanch', 'inpatient','marriage','work','madrs1','madrs2'

Here for this pair plot we took the hue=. As madrs 1

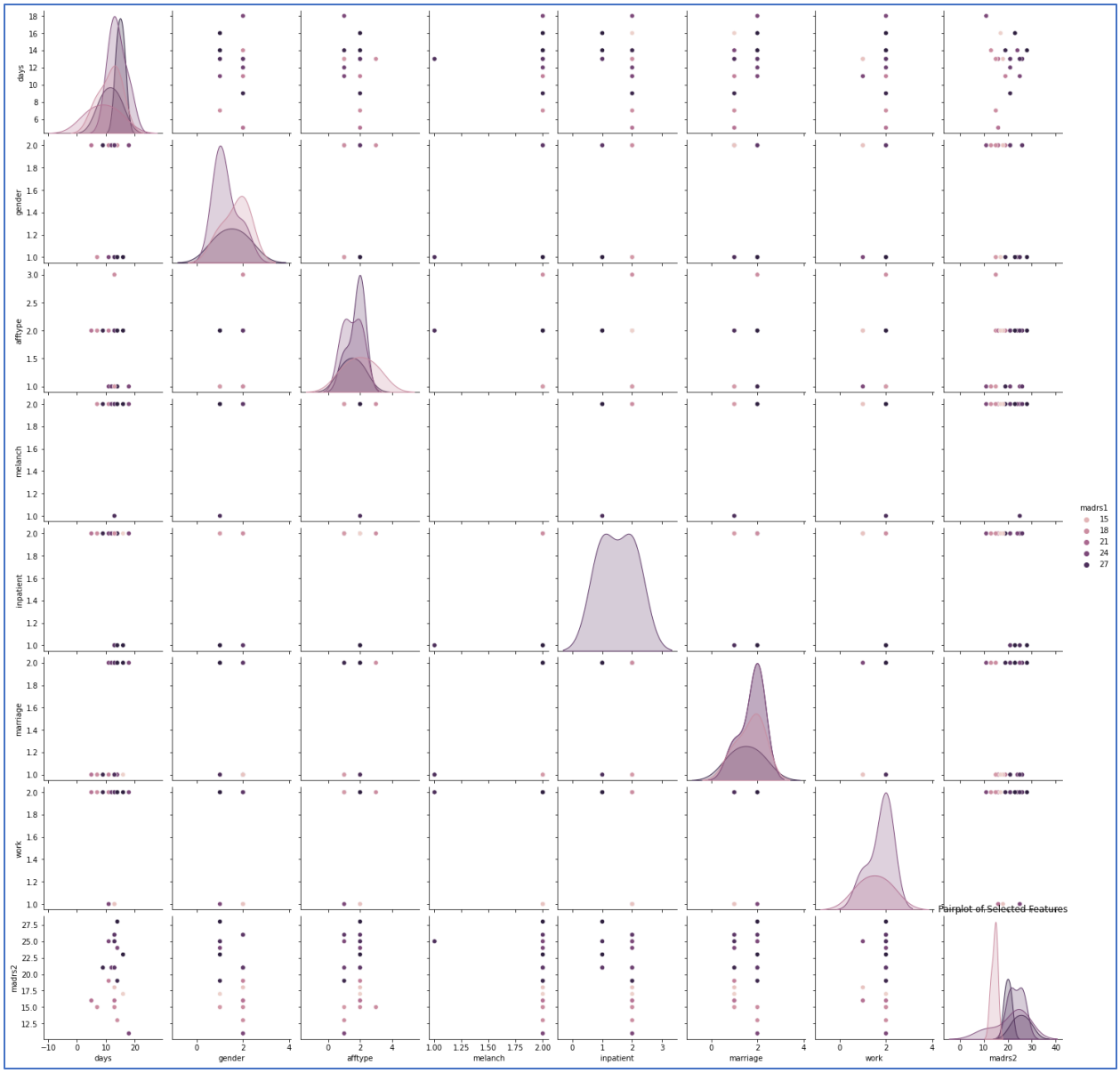

Figure 9: Pair plot MADRS2

Here for this pair plot, we took the hue as Madrs 2

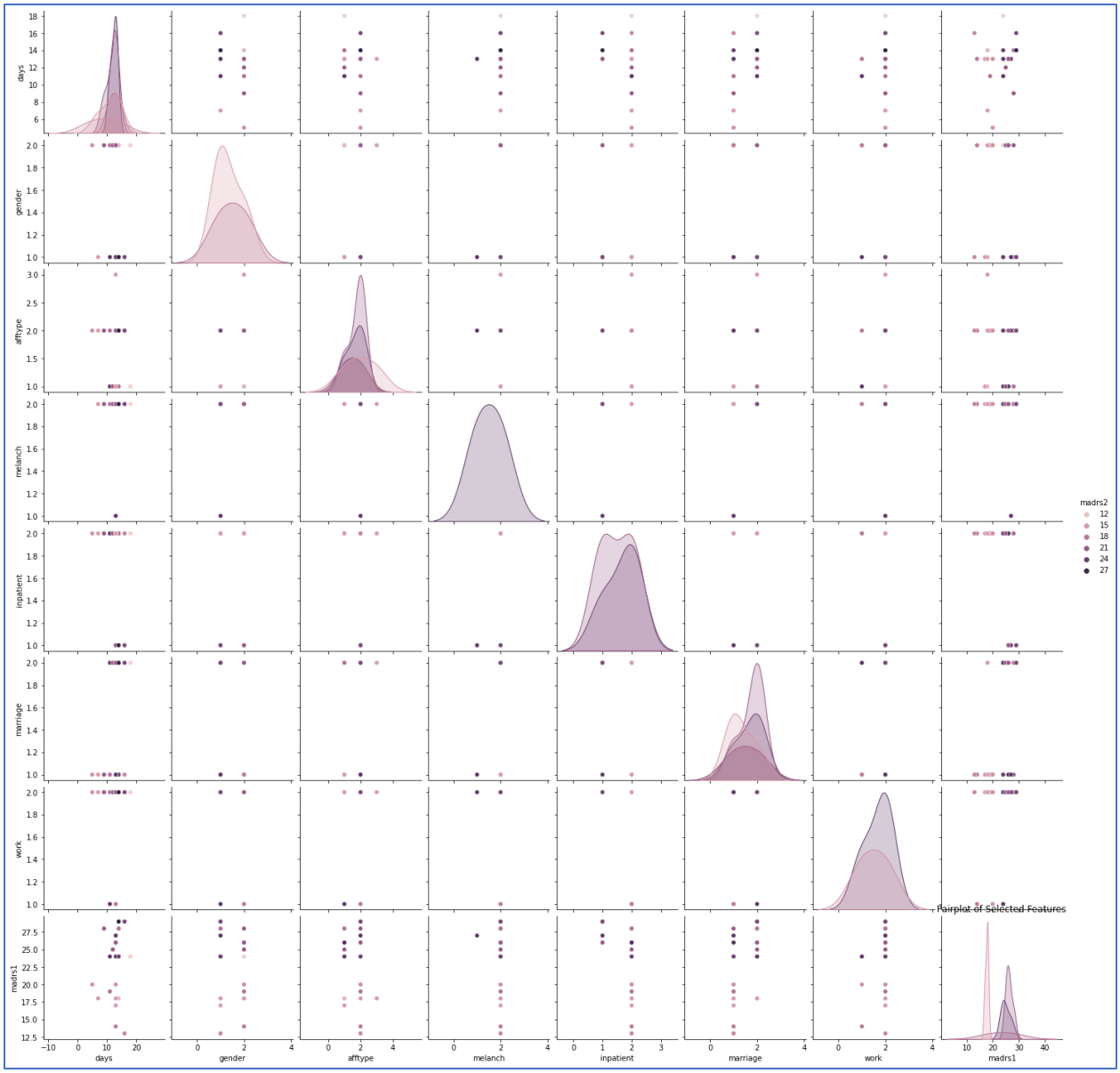

Figure 10: Pair plot

Distribution of Depression Data:

Figure 11: Distribution of Depression Data

Dimensional Reduction:

Dimension reduction is a technique that is commonly used to reduce the number of variables in a dataset while retaining the relevant information. It is helpful when you have a large number of variables and want to simplify the analysis. However, in our case, we need to use all the variables in the dataset, including demographic variables, MADRS scores, and activity data, to predict depression. Therefore, we don't need to use dimension reduction techniques.

In other words, since we are using all the variables in our analysis, there is no need to reduce the number of variables. Dimension reduction is only necessary when you have a large number of variables and want to identify the most important ones to include in your analysis.

Data Mining Methods:

1) Random Forest: This model is popular ensemble learning algorithms that generates a forest of decision trees and selects the best one by calculating the average of results of all trees.

   Advantages:

   1) Can be utilized in classification and regression problems.
   2) Performs good with null data and missing values.
   3) It's not prone to dimensionality since each tree does not consider all variables and spaces are reduced.

   Disadvantages:

   1) It's highly complex compared to decision trees where it follows a path of the tree.
   2) Training time is more due to its complexity. Whenever a decision is to be given it has to generate an output for all the tree lines.

2) KNN: This model is denoted as K-Nearest Neighbors which is a non-parametric classification algorithm that classifies samples based on similarity to other samples in the evaluation sets.

   Advantages:

   1) No training period is needed for this model as the data itself is a model definition within itself.
   2) Easy application due to the fact only the distance between different points need to be calculated and formulas like Euclidian and Manhattan can be used.
   3) Since there is no training period, data can be added and removed at any point in time.

Disadvantages:

1) Does not suit large data sets as the formula usage and calculations will become costly.
2) Not a good fit for high dimensionality as it will complicate the distance calculation from one point to another point.
3) Sensitivity to noise and missing data.

3) Naïve Bayes: This model is a probability classification algorithm that calculates the probability of a sample belonging to each class and filters out all low probabilities.

Advantages:

1) Simple and easy to execute.
2) Does not need much training to use the model.
3) Can handle both continuous and discrete data for model usage.

Disadvantages:

1) This model assumes all values are independent which limits the usage in real world scenarios.
2) Model uses zero frequency problem which assigns zero probability to categorical variables in the training dataset.
3) The probability calculations are misleading most of the time due to its nature of model.

4) XGBoost: This model is a popular gradient boosting algorithm creating a decision tree iteratively and combine results.

Advantages:

1) XGBoost is very flexible in terms of calculations and data.
2) It takes advantage of parallel processing and regularization.
3) Faster than Gradient Boosting.

Disadvantages:

1) Does not do with unstructured and sparse datasets.
2) The overall model is very unstable and hardly scalable due to estimations from previous calculations.
3) This model is highly sensitive to data that are categorized as outliers.

5) Decision Trees: finally this model is a popular one which creates tree-like model of decisions and their consequences.

Advantages:

1) Very easy to understand.
2) Minimal effort for data preparation.
3) Non-linear parameters do not result in less performance.

Disadvantages:

1) Overfitting: noise in the data
2) Instability: it can be unstable due to various variations of data

	model	accuracy	precision_0	precision_1	recall_0	recall_1	f1-score_0	f1-score_1
0	Random Forest	0.65	0.88	0.35	0.65	0.67	0.75	0.46
1	KNN	0.69	0.94	0.35	0.67	0.80	0.78	0.48
2	Naive Bayes	0.49	0.19	0.91	0.75	0.45	0.30	0.60
3	XGBoost	0.58	1.00	0.00	0.58	0.00	0.74	0.00
4	Decision Trees	0.73	0.91	0.48	0.71	0.79	0.79	0.59

Figure 12: Model Accuracy

The models that were selected for depression disorder are Random Forest, KNN, Naïve Bayes, XGBoost and Decision Trees. Based on the performance metrics, the most accurate and keen models that are suggested for depression study of data are Random Forest and Decision Trees.

Model Performance and Evaluation:

Random Forest:

This model technique is also a machine learning methodology used for classification and regression tasks. It resembles learning method building collection of decision trees, where each tree is trained on random sampling data set involving data and random dataset of the features. The Random Forest sampling calculates the individual trees and combine with other data set to reach its destination of data sets and results. This ensemble learning technique creates a group of decision trees, each of which is trained using a random portion of the training data and a random subset of the features. The final prediction is then created by combining the predictions of the various trees.

	precision	recall	f1-score	support
0	0.88	0.65	0.75	43
1	0.35	0.67	0.46	12
accuracy			0.65	55
macro avg	0.61	0.66	0.60	55
weighted avg	0.76	0.65	0.68	55

Figure 13: Prediction

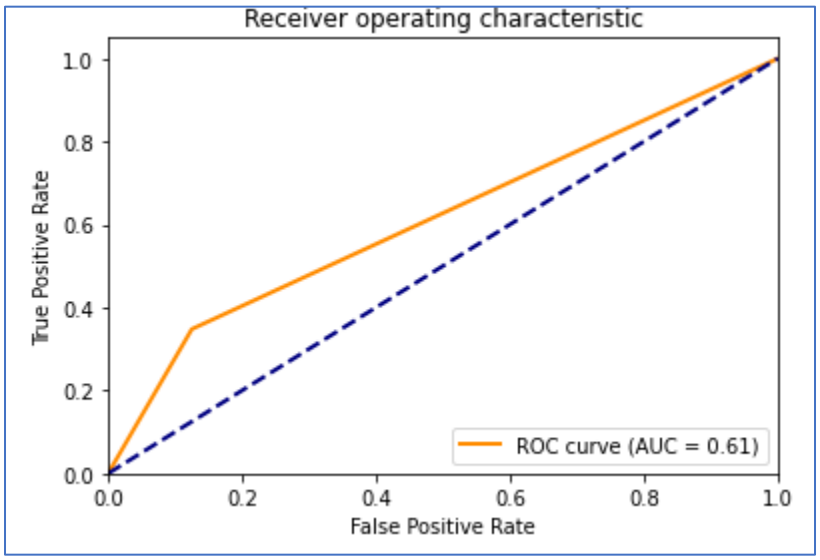

Figure 14: ROC Curve

In this model, the accuracy achieved is 65% with a precision score for class 0 of 88% indicating the model has a low false-positive rate. Though the recall score for class 1 is higher with a value of 67% referring to as the model having high true-positive rate. The F1-score for class 1 is very low with a value of 46%.

In conclusion, Random Forest performed moderately well, but unfortunately had obstacles with class 1 samples.

KNN:

K-Nearest Neighbors model is a non-parametric classification algorithm which categories samples based on similarity to other samples in training set.

	precision	recall	f1-score	support
0	0.94	0.67	0.78	45
1	0.35	0.80	0.48	10
accuracy			0.69	55
macro avg	0.64	0.73	0.63	55
weighted avg	0.83	0.69	0.73	55

Figure 15: Precision Data

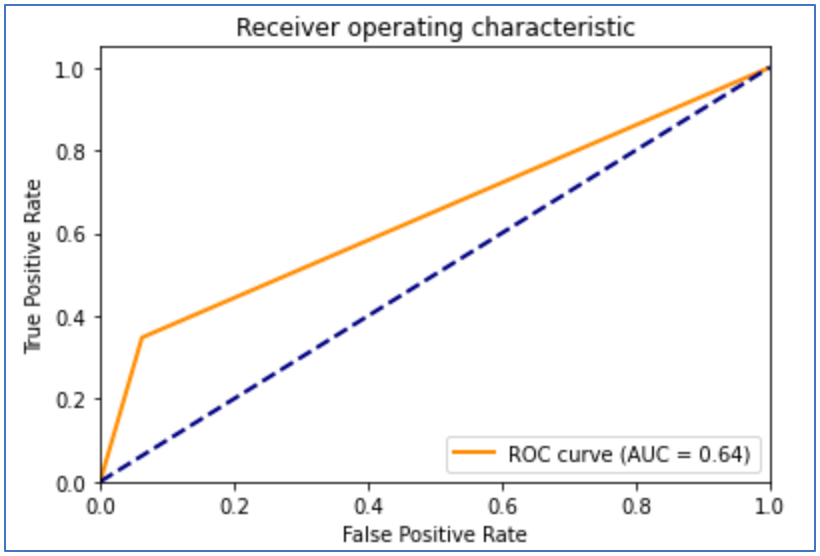

Figure 16: ROC Curve

In this model, the accuracy was at 69% with a precision score of class 0 is high 94% which resulting to have a low false-positive rate. With the recall score for class 1 being on the higher side 80%, indicating the model with a very high-true positive rate. The F1-score for class 1 is comparatively very low which is at 48%.

In conclusion, KNN performed moderately decent but seems to have a struggle in identifying class 1 sample data.

Naïve Bayes:

This model is a statistical classification algorithm which evaluates the probability of a sample belonging to each class and selects class with highest probability.

	precision	recall	f1-score	support
0	0.19	0.75	0.30	8
1	0.91	0.45	0.60	47
accuracy			0.49	55
macro avg	0.55	0.60	0.45	55
weighted avg	0.81	0.49	0.56	55

Figure 17: Precision Data

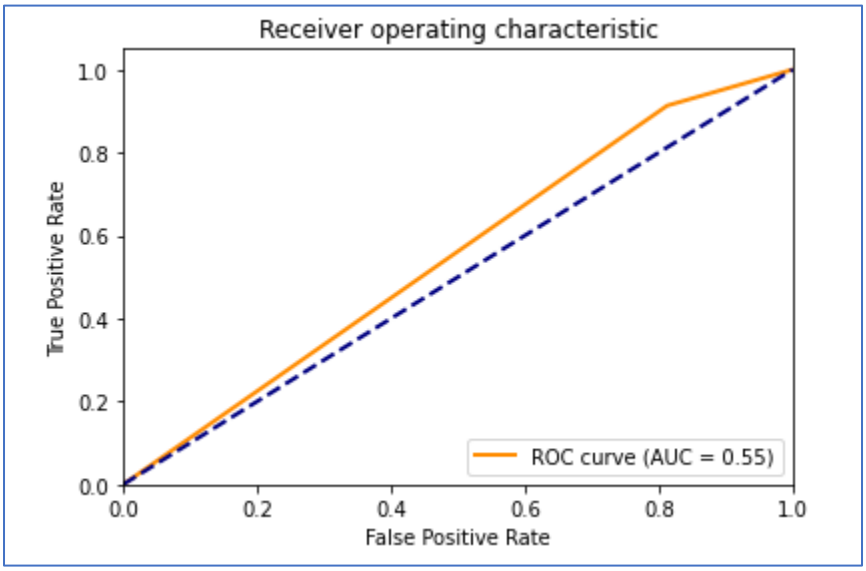

Figure 18: ROC Curve

The model evaluated with an accuracy of 49% and with a precision score for class 1 with a value of 91% which resulting the model to have a low false-positive rate. The recall score for class 0 is higher with a percentage of 75% proving to have a high true-positive rate. The F1-score for class 0 is very low with a percentage of 30%.

In conclusion, Naïve Bayes performance is very poor with the accuracy being very low, and the model seems to have difficulties identifying class 0 sample records.

XG Boost:

This model is known as Extreme Gradient Boosting is solutioned for regression, classification, and ranking issues. It is a unique tool technique which is based on trees that turn each weak model into a decision tree into a significant strong model. With XGBoost, the decision trees are incrementally added to the model in order to optimize the factual function. The differences between expected values and real values are known as residuals which are in-turn used to exercise each of the consequent tree. XGBoost is known for its management of interactions and nonlinear correlations between features.

	precision	recall	f1-score	support
0	1.00	0.58	0.74	55
1	0.00	0.00	0.00	0
accuracy			0.58	55
macro avg	0.50	0.29	0.37	55
weighted avg	1.00	0.58	0.74	55

Figure 19: Precision Data

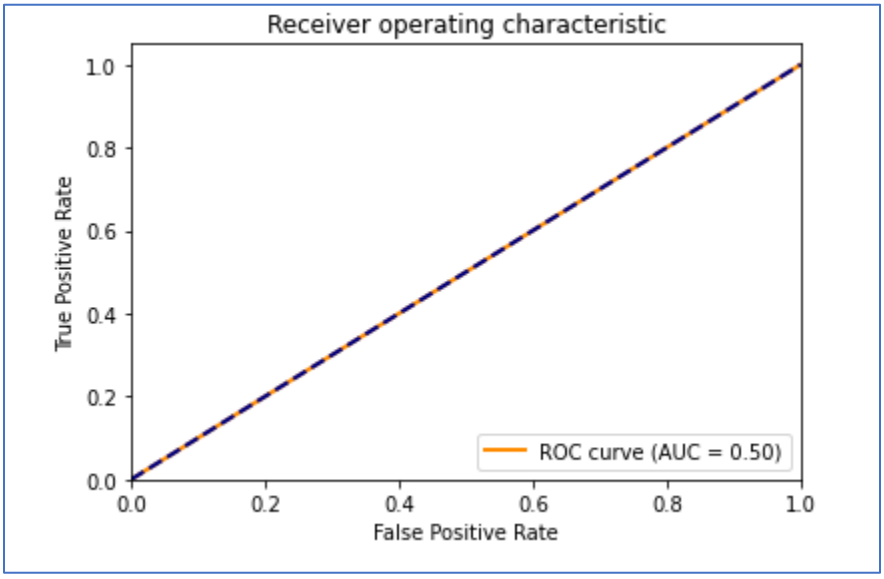

Figure 20: ROC Curve

This model was able to get an accuracy of 58% but unfortunately failed to predict any class 1 samples, resulting in 0 F1-score for class 1.

In conclusion, XGBoost is not a keeper for depression data set evaluation as it was not able to classify and differentiate the class 1 sample data sets.

Decision Trees:

This model performance is a clean and popular algorithm which creates tree-like model of decisions and their penalties.

	precision	recall	f1-score	support
0	0.91	0.71	0.79	41
1	0.48	0.79	0.59	14
accuracy			0.73	55
macro avg	0.69	0.75	0.69	55
weighted avg	0.80	0.73	0.74	55

Figure 21: Precision Data

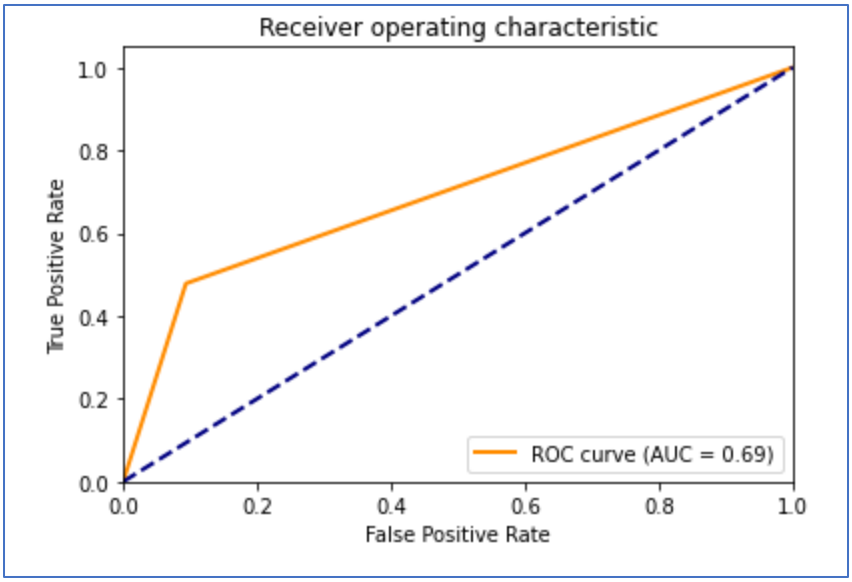

Figure 22: ROC Curve

The performance resulted with an accuracy of 73% and with a precision score of class 0 having a percentage of 91% which indicates low false-positive rate. The recall score of 79% for class 1 which results in high true-positive rate. The F1-score for class 1 is moderate with a percentage of 59%.

In conclusion, the Decision Tree model has a decent rate.

Conclusion:

Both Random Forest and Decision Trees models provided accurate and good performance for their classification tasks with Random Forest model having slightly low accuracy but better equilibrium between precision and recall for both class 0 and class 1. The advantages of these models are they are significantly easy to understand, decode, handle non-linear relationships between variables and less likely to overfitting than other models. Other models like KNN provide good performance for identifying positive cases but tend to misclassify negative cases as positive. The Naïve Bayes model was a very poor model for depression data set due to misclassification of positive cases as negative. Finally, XGBoost model is also not quite suitable for depression due to its vague performance in classifying positive cases. By evaluating all models and tracking their accuracies and performances we can conclude that Random Forest and Decision Tree models are the best models for Depression dataset evaluations.

In real world scenarios, Random Forest is widely used in E-commerce, banking, medicine, and stock markets. For example, in the banking sector, this model is used to find which customer will default on loans. The model has many advantages which are it can be used in classification and regression problems; solves problems of overfitting based on majority voting or averaging; performs well without any values in few columns; model has parallelism; it is immune to curse of dimensionality. A few disadvantages to this model are its highly complex, and training time is more than other models due to its complexity. It is one of the best models with high performance, used in many industries. The model can handle binary, continuous, and categorical data and overall, it is fast, simple, flexible, and a very robust model. Decision Trees on the other hand have their own advantages in the real world. Decision Trees models are faster in computation. When a data set with features is taken as input by a decision tree, it will formulate some rules to make predictions. The decision tree model is used in many various areas such as engineering, civil planning, law and business. A frequent and most common example we can relate to decision model is buying an item from online shopping where based on our search we get suggestions and recommendations to buy other products. The advantages of decision model are its simple to understand; very little effort for data prep. The disadvantages are noise in data will cause overfitting; model in general is very unstable due to variations in data.