# Using Machine Learning Models to Predict Depression

October 12, 2022

**Varun Pininty, California High School, San Ramon CA**

1. Abstract:

Depression is a mood disorder that negatively affects performance in work and school and also ruins your social life and leads to risk taking behaviors that endangers peoples' lives. Timely intervention and diagnosis can help identify and mitigate the effects of depression. In this study, I used a dataset formed by a survey to create multiple machine learning models in order to create a model that would accurately predict if someone had depression. Out of 4 different machine learning models I used, I was able to create a machine learning model using the Random Forest Classifier to achieve an accuracy of 90.48%.

Keywords: depression, predict, machine learning

## 1 Introduction:

Depression is a mental disorder that negatively impacts your life in various ways and left untreated, can ruin relationships and cause problems at school and work. It's estimated that nearly  of all cases of depression are left undiagnosed, leading to the depression getting worse and causing that person's lifestyle and work performance to deteriorate [1].I wanted to develop a machine learning model that can accurately predict if someone has depression, as timely intervention can give undiagnosed patients the chance to receive treatment and therefore improve their life [2].

# 2 Methodology:

## 2.1 Data Acquisition:

A survey was given to students at International Islamic University Malaysia, IIUM by a fellow student named Sharifal Islam. In this survey, students were asked various questions as can be seen in Figure 1 and their responses were compiled into a dataset that was then posted on Kaggle.

```
['Choose your gender',
 'Age',
 'What is your course?',
 'Your current year of Study',
 'What is your CGPA?',
 'Marital status',
 'Do you have Depression?',
 'Do you have Anxiety?',
 'Do you have Panic attack?',
 'Did you seek any specialist for a treatment?']
```

Figure 1: List of questions asked in survey and names of the columns in the dataset used.

## 2.2 Data Cleaning

To ensure that the model properly works, we have to make sure there aren't any null values or repeating values. When checking for null values, we find that there is one null value for Age. To correct this, we simply take the average of all the other participants' age to replace the null value.

```
Timestamp                                      0
Choose your gender                             0
Age                                            1
What is your course?                           0
Your current year of Study                     0
What is your CGPA?                             0
Marital status                                 0
Do you have Depression?                        0
Do you have Anxiety?                           0
Do you have Panic attack?                      0
Did you seek any specialist for a treatment?   0
```

We also want to get rid of Timestamp due to the time of turning in the survey having no real effect on if someone has depression or not. The next step is to check to make sure the course names are alright. There are many redundancies in names of course with spelling and capitalization errors.

We will manually change the course names to ensure that there won't be any unnecessary course names.

### 2.3 Label Encoding

If we were to directly try to train and test our machine learning models right now, it wouldn't work. This is due to many of the values being objects, not numerical values. For our machine learning models to run, it's necessary to be able to convert or treat our object values as numerical values. To do this, we do label encoding, a technique that allows us to assign a unique integer to each data point.

```
Choose your gender                              object
Age                                            float64
What is your course?                            object
Your current year of Study                      object
What is your CGPA?                              object
Marital status                                  object
Do you have Depression?                         object
Do you have Anxiety?                            object
Do you have Panic attack?                       object
Did you seek any specialist for a treatment?    object
```

### 2.4 Creating Machine Learning Models

The first thing to do to create a machine learning model is to split the dataset into two parts, a training set and a testing set. We will use the train-test-split from Scikit-Learn to train the model on 80% of the data and to test the models on 20% of the data. After doing this, we will create 4 different machine learning models to choose the best algorithm for finding the accuracy. We will use Logistic Regression, SVM, Decision Tree Classifier, and Random Forest Classifier.

## 2.5 Results:

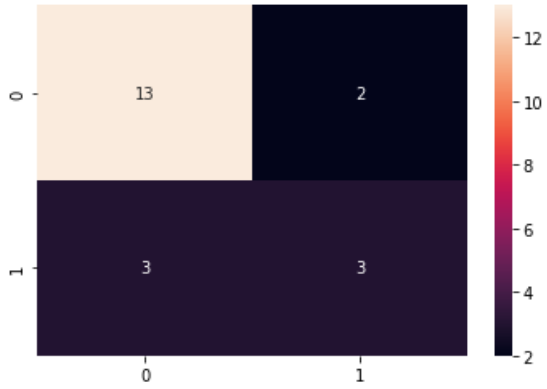
```
[[13  2]
 [ 3  3]]
              precision    recall  f1-score   support

           0       0.81      0.87      0.84        15
           1       0.60      0.50      0.55         6

    accuracy                           0.76        21
   macro avg       0.71      0.68      0.69        21
weighted avg       0.75      0.76      0.75        21

Accuracy of logistic regression classifier on test set: 0.7619


[[15  0]
 [ 3  3]]
              precision    recall  f1-score   support

           0       0.83      1.00      0.91        15
           1       1.00      0.50      0.67         6

    accuracy                           0.86        21
   macro avg       0.92      0.75      0.79        21
weighted avg       0.88      0.86      0.84        21

Accuracy of SVM classifier on test set: 0.8571
[[12  3]
 [ 3  3]]
              precision    recall  f1-score   support

           0       0.80      0.80      0.80        15
           1       0.50      0.50      0.50         6

    accuracy                           0.71        21
   macro avg       0.65      0.65      0.65        21
weighted avg       0.71      0.71      0.71        21

Accuracy of decision tree classifier on test set: 0.7143


 [[15  0]
 [ 2  4]]
              precision    recall  f1-score   support

           0       0.88      1.00      0.94        15
           1       1.00      0.67      0.80         6

    accuracy                           0.90        21
   macro avg       0.94      0.83      0.87        21
weighted avg       0.92      0.90      0.90        21

Accuracy of random forest classifier on test set: 0.9048
```

Figure 2: Accuracy of four different machine learning models

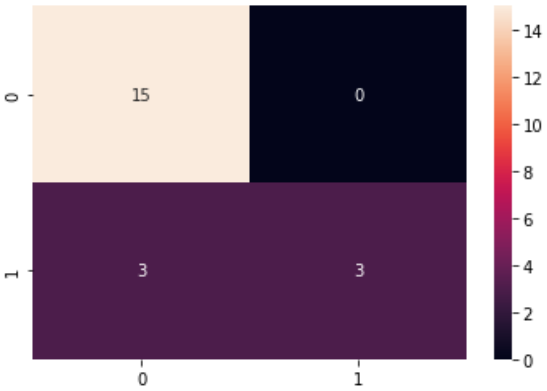
| | Predicted Negative | Predicted Positive |
|---|---|---|
| Actual Negative | $a$ | $b$ |
| Actual Positive | $c$ | $d$ |

Figure 3: Figure 3: The confusion matrix of a two-class classification problem

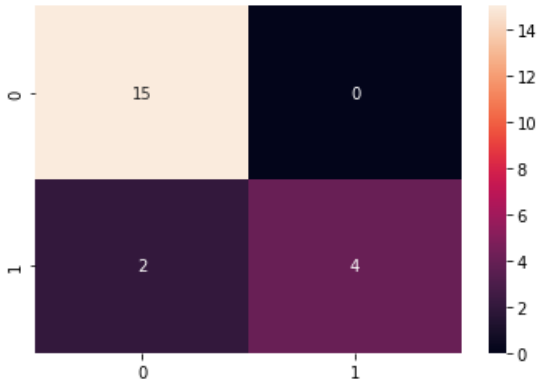
Actual Results



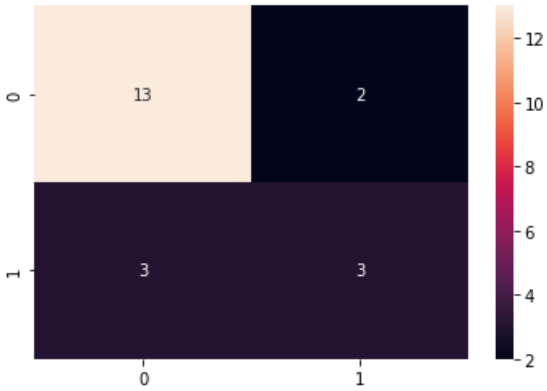

(a) Logistic Regression Confusion Matrix




(b) SVM Confusion Matrix




(c) Random Forest Confusion Matrix




(d) Decision Tree Confusion Matrix

Figure 4: Confusion Matrix of each Machine Learning model.

As seen with Figure 2; All of the models had over a 70% accuracy with Random Forest Classifier being the most accurate with 90.48% accuracy and Decision Tree Classifier being the least accurate at 71.43%. The SVM Classifier had an 85.71% accuracy and the Logistic Regression Classifier had an 76.19% accuracy. To find out the specifics of what our machine models were predicting, we can use confusion matrixes. With Figure 3; we can see how to read a confusion matrix with a being the number of correct negative predictions, b is the number of incorrect positive predictions, c being the number of incorrect negative predictions, and d being the number of correct positive predictions [3]. In our case, positive predictions means someone was predicted to have depression and negative predictions means they are not predicted to have depression. With our various models, we can see that generally, our different models had a high chance of correctly predicting a person not having depression, but struggled when predicting a person having depression.

## 3 Conclusion:

The machine learning models, especially the Random Forest Classifier, were able to achieve relatively high accuracy rates. However, more improvements to these models are necessary if these models are to truly be useful in determining if someone is depressed. There are many routes of improvements such as asking more types of questions such as income level and place of residency. We could also try other machine learning models such as K-Fold Cross Validation as well as trying to use a Recursive Feature Elimination to help reduce any unnecessary variables. The biggest need for improvement of these models however is simply more data. If this model wants to be feasible for the general population. The model needs to be able to train on more data and have more variety of data in order to avoid overfitting the model. The practical uses of an improved model can greatly help practitioners as a safety net as it's estimated that primary care physicians fail to recognize 30-50% of depressed patients [4]. Using the basis of the models created, with enough improvements, we can hopefully be able to successfully identify if someone has depression and are able to receive the help they need.

## References


[1] Williams S, Chung G, and Muennig P. "Undiagnosed depression: A community diagnosis. SSM -population health". In: 3 (2017), pp. 633–638. DOI: 10.1016/j.ssmph.2017.07.012.

[2] Garriga R et al. "Machine learning model to predict mental health crises from electronic health records". In: *Nature medicine* 28 (6 2022), pp. 1240–1248. DOI: 10.1038/s41591-022-01811-5.

[3] Visa Sofia et al. "Confusion Matrix-based Feature Selection". In: *CEUR Workshop Proceedings. 710. 120-127* (2011).

[4] Bansal V, Goyal S, and Srivastava K. "Study of prevalence of depression in adolescent students of a public school". In: *Industrial psychiatry journal* 18 (1 2009), pp. 43–46. DOI: 10.4103/0972-6748.57859.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5769115/#bib27 - Introduction, of depression undiagnosed

https://www.nature.com/articles/s41591-022-01811-5 - Introduction, timely intervention can help mitigate effects

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3016699/ - Conclusion, Usual care by primary care physicians fails to recognize 30-50% of depressed patients.

https://www.researchgate.net/publication/220833270_Confusion_Matrix-based_Feature_Selection