Depression Prediction Using Machine Learning

Pratik Datey

Abstract

Millions of people globally suffer from depression and it is a debilitating condition. At best it can be difficult for people to live their lives normally and happily, and at worst it leads to death by suicide. Primary care doctors are overwhelmingly finding that they are faced with the need to treat mental health conditions such as depression without any particular training of how to handle such cases. There is evidence that an integrated approach where physicians regularly screen patients for mental health disorders and work together with psychologists and other mental health professionals to treat patients leads to reduced costs and better patient outcomes. However, this approach can require a lot of buy-in from many individuals, require extra training, and is often not logistically feasible.

Using data from the CDC National Health and Examination Survey, machine learning was applied to predict patients who may have depression based on information that could typically be found in a medical file. These predictions could be used to put patients in touch with experienced mental health professionals sooner and easier.

1.   Problem Statement

Mental health issues are largely taboo in India. In 2017, mental disorders were the second leading cause in terms of years lived with disability worldwide. Never before has the world put mental health on a pedestal as high as it has now. Fortunately, the problem has pushed several national and international organisations to consider including mental health policies as a priority and make it a part of the Sustainable Development Goals (SDGs).**56 Million Indians Suffered** From Depression According to the World Health Organisation (WHO), India's mental health workforce is severely understaffed.[1]

It has predicted that owing to the pandemic and its influential factors, the number of patients who have any mental disorders would touch 20 per cent in the country. Roughly 56 million Indians suffer from depression, and 38 million suffer from some anxiety disorder. India has one of the highest prevalences of mental illnesses globally. The National Mental Health Survey 2016 found that close to 14 per cent of India's population required active mental health interventions. Every year, about 2,00,000 Indians take their lives. The statistics are even higher if one starts to include the number of attempts to suicide.

2.   Market/ Customer/ Business need Assessment

There is a massive shortage of psychiatrists and psychologists compared to the number of patients suffering from mental health issues. According to the international organisation's data, there are only three psychiatrists and psychologists for every 1,00,000 people. The needs assessment study highlights a number of interesting findings concerning the perceptions of depression and suicide among the four rural communities. While overall levels of awareness concerning depression and suicide appear to be high, there were significant differences across socio-demographic groups. Men, the under 40 age group and respondents with lower levels of education were significantly less concerned about current levels of suicide and depression, and men were less concerned than women about access to services. Clearly, given the increasing rates of suicide for younger males, the intervention programme needs to address levels of awareness among men and the younger members of the community. young adults between the ages of 18 and 25 had the highest prevalence of any mental illness and were more willing to talk openly about psychological well-being and to seek assistance.

3.   Target Specifications and Characterization

According to the World Health Organisation, more that 264 million people globally have depression. Many suicides each year are caused by depression with suicide being among the leading causes of death for young people especially. The National Institute of Mental Health found that the prevalence of a major depressive episode among U.S. adults in 2017 was 7.1% of people with young adults being the most affected.[2]

```
In [71]: plt.figure(figsize=(10,6))
         data['Age'].plot(kind='hist')
         plt.title('Age Counts')

Out[71]: Text(0.5, 1.0, 'Age Counts')
```
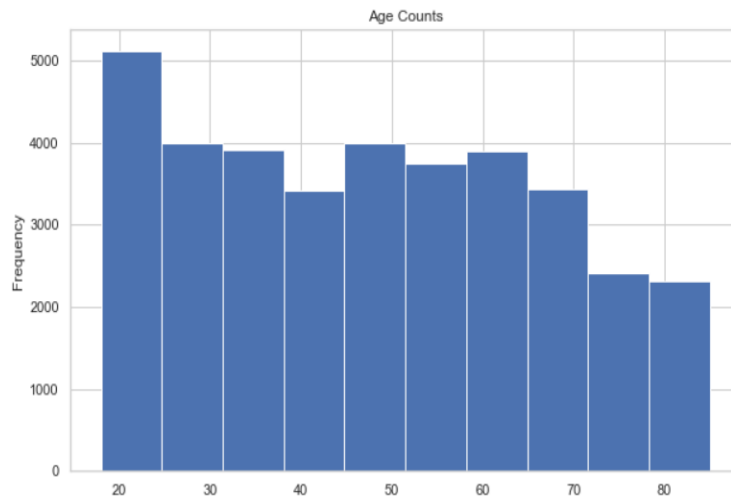

Fig No.01

The American Psychological Association identified that primary care physicians are often being asked to diagnose mental disorders such as depression without adequate training on how to handle such treatments. According to their numbers, 70% of primary care visits are because of patients' psychological problems, more than 80% of patients who have symptoms with no diagnosis receive psychological treatment by a physician, and only 10% follow up to a mental health professional. Patients are not getting the care they desperately need as 70% of individuals with depression go undiagnosed. Among people who commit suicide, 90% of people had a mental disorder and 40% of people had visited their doctor within the last month.

4.   External Search

The data for this project is from the Centers for Disease Control and Prevention National health and Nutrition Examination Survey. This data includes a vast array of health data done on a sample of the American population each year and is released every two years. For this project, data was taken from the years between 2005 and 2018 and comprised 36259 entries total of U.S. adults. Only data that was consistent across years was used and there was effort to only include data that would be reasonably found in a patient's medical file. Using as little data as possible while still being able to have accurate predictions is

desirable as it would catch more people who may not have a very deep medical history and also puts less burden on providers to have to capture so much information.
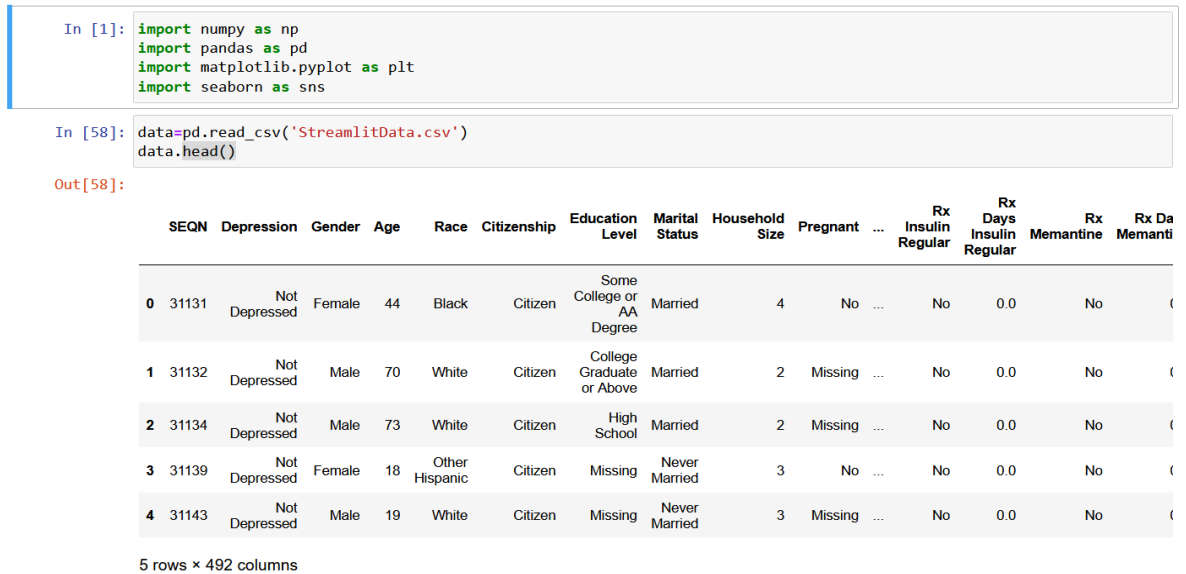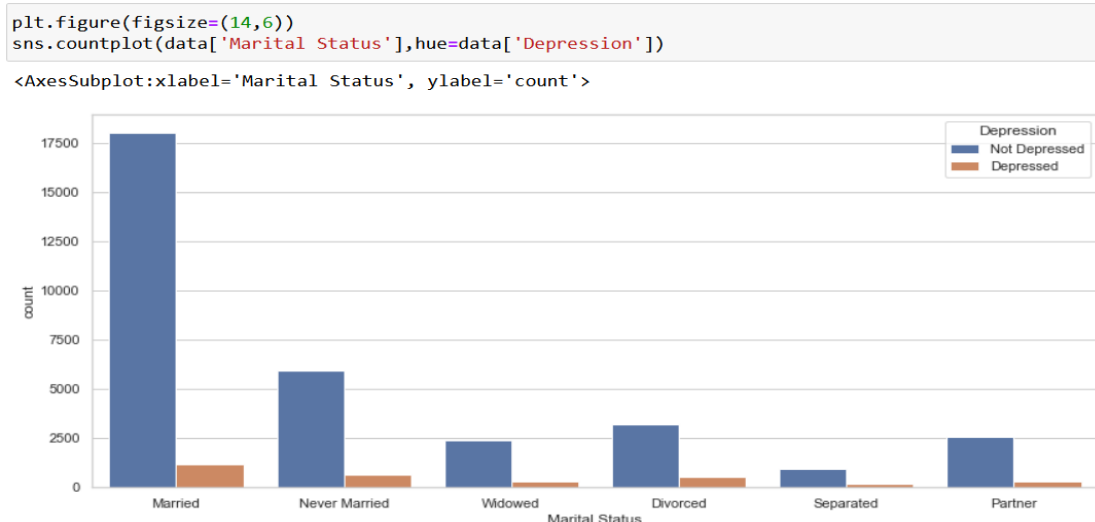
Dataset Origin :- https://wwwn.cdc.gov/nchs/nhanes/default.aspx.

Let's view our dataset :-

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [58]: data=pd.read_csv('StreamlitData.csv')
         data.head()
```

Out[58]:

	SEQN	Depression	Gender	Age	Race	Citizenship	Education Level	Marital Status	Household Size	Pregnant	...	Rx Insulin Regular	Rx Days Insulin Regular	Rx Memantine	Rx Da Memanti
0	31131	Not Depressed	Female	44	Black	Citizen	Some College or AA Degree	Married	4	No	...	No	0.0	No	
1	31132	Not Depressed	Male	70	White	Citizen	College Graduate or Above	Married	2	Missing	...	No	0.0	No	
2	31134	Not Depressed	Male	73	White	Citizen	High School	Married	2	Missing	...	No	0.0	No	
3	31139	Not Depressed	Female	18	Other Hispanic	Citizen	Missing	Never Married	3	No	...	No	0.0	No	
4	31143	Not Depressed	Male	19	White	Citizen	Missing	Never Married	3	Missing	...	No	0.0	No	

5 rows × 492 columns

Fig No.02

Some Visualisation of our data :-

```
plt.figure(figsize=(14,6))
sns.countplot(data['Marital Status'],hue=data['Depression'])
```

```
<AxesSubplot:xlabel='Marital Status', ylabel='count'>
```

```
: plt.figure(figsize=(10,6))
  sns.countplot(data['Gender'],hue=data['Depression'])

: <AxesSubplot:xlabel='Gender', ylabel='count'>
```
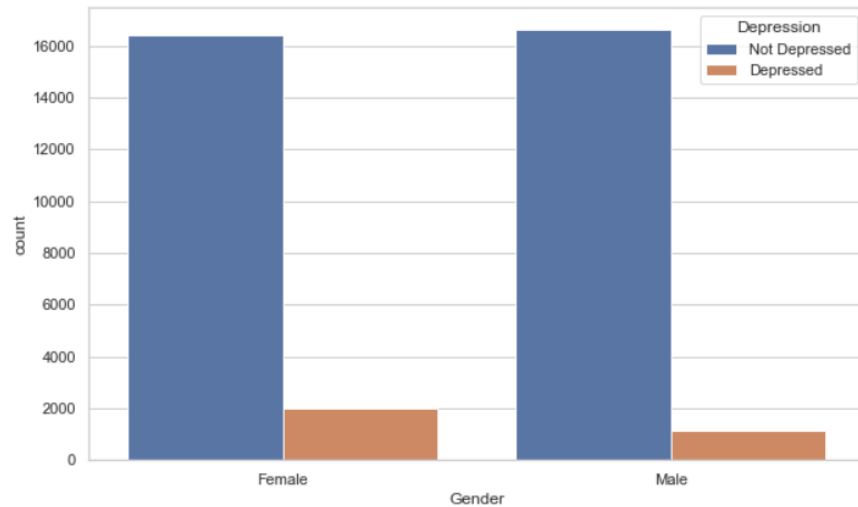

Fig No.03

5. Benchmarking

AI Can Now Detect Depression From Your Voice And It's Twice As Accurate As Human Practitioners. "Wouldn't it be great if we could detect and treat symptoms of depression in adults in the early stages, just the way mothers sense health challenges with their kids?" asks David Liu, CEO of Sonde Health. With a short voice sample, Sonde's vocal biomarker technology can provide insight and tell if you are at risk for conditions of mental health and respiratory health as shown in Fig No.4.
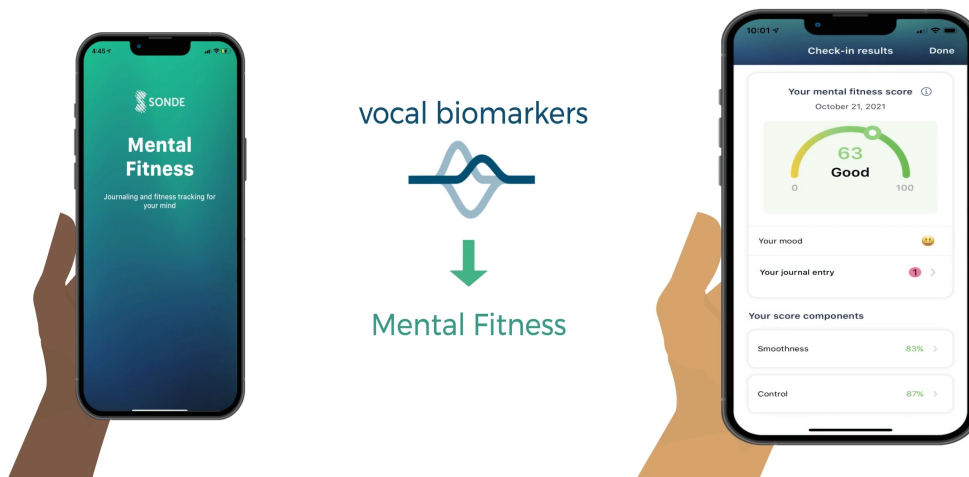
Fig No.04

Sonde's vocal biomarker gives a fitness score from your voice sample, that helps to understand your mental health.[3]

6.   Applicable Patents

6.1 The heart beat mood tracker

This application entitled 'Mood analysis method, system and apparatus' relates to a medical device which uses ECG (electrocardiogram) analysis to determine the user's heart rate. ECG analysis is used to determine the state of the sympathetic and parasympathetic nervous systems. These systems activate in response to depression and anxiety disorders and can be analysed to determine the user's mood. The application discloses a 'mood analysis system' that analyses and displays a user's mood according to the user's ECG. Fig No.05 shows the mood display coordinate system. Pretty cool, eh?[4]
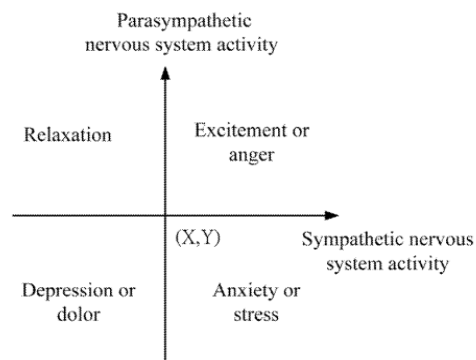
Fig No.05

6.2 Monitoring for a mental health crisis

This application, 'Mental health digital behaviour and monitoring system and method', shows that it's not just physiological behaviours that can be monitored to determine our state of mind. In the system, online 'activity data' from a user is collected and tracked. This data could relate to any online activity (social media, emails, browsing, and even Netflix consumption!). The data is analysed using AI algorithms. Irregularities are flagged by comparing the activity data to a baseline. An

alert may be sent to a mental health professional if an irregularity indicates the user is having a mental health crisis.[5]
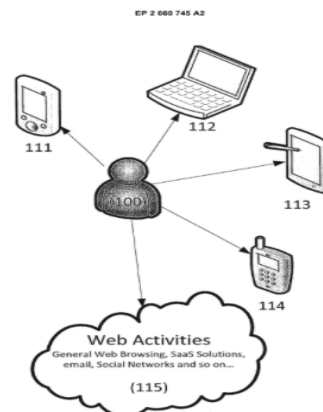

Fig No.06

Although it might feel a bit 'big brother', this system could fill an important gap in mental health treatment. As explained in the application, "currently [the caregiver] is unable to monitor a patient's condition between appointments". This can be critical if a patient's condition worsens, or when a patient changes medication. The disclosed monitoring system further doesn't require the patient to complete a questionnaire about their current mood, which can be a daunting task when in psychological distress.

7. Applicable Regulations (Government and Environmental)

- The Mental Healthcare Act, 2017 -
  The preamble of the **Mental Healthcare Act, 2017** (Act or MHCA) aims to provide mental healthcare and services for persons with mental illness and to promote, and fulfil the rights of such persons during delivery of mental healthcare and services. The act is progressive, patient-centric, and rights-based.
- Data protection and privacy regulations.
- Patents on ML algorithms developed.
- Laws related to privacy for collecting data from users Protection/ ownership regulations.
- Ensuring open-source, academic and research community for an audit of Algorithms.
- Review of existing work authority regulations.

8.   Applicable Constraints

● The quality of care depends heavily on the quality of the workforce. The significant variations between countries in the number of mental health professionals arouses concern:
● Requires a lot of research to obtain a universal dataset of depression patients in-order to provide more sophisticated and accurate results.
● Data privacy Constraints.
● Confidential data to be obtained to train the model.
● A lot of human resources are required to gain data.

9.   Business Opportunity

Indian mental health and wellness startups have garnered over US$ 28.24 with the decade's highest investment in 2021. McKinsey estimates the global wellness market worth more than $1.5 trillion in 2021, with an annual growth of 5 to 10 per cent. Indian mental health and wellness startups have garnered over US$ 28.24 with the decade's highest investment in 2021. The investment came in just three rounds. The highest investment made earlier in the segment was in 2019 of around US$ 12.36 in 7 rounds of funding.

Many people take so much time to realise that they have any symptoms of depression, but the sooner you realise, the sooner you can start to work on that. If you've felt depressed for a long time, you may feel like it'll always be part of your life.

Fig No.07

10.   Concept Generation

Using data from the CDC National Health and Examination Survey, machine learning was applied to predict patients who may have depression based on information that could typically be found in a medical file. These predictions could be used to put patients in touch with experienced mental health professionals sooner and easier.

11.   Final Product Prototype

The approach for this project was to create many different model types to see what performs the best and to compare and contrast the different types of models. The modelling effort was done starting with simpler models and moving to more complex models. The OSEMN process is the overarching structure of this project.[6]
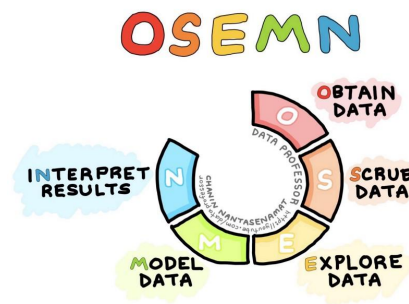
Fig No.08

So in order to generate the model based on the problem stated above, we need to use Machine learning. Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.
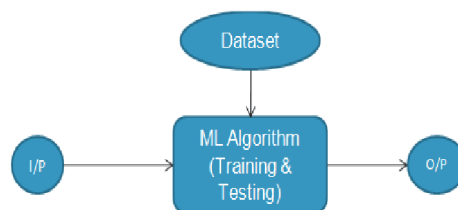

Fig No.09

The target was calculated using the PHQ-9 depression screening tool that was asked of all participants in the NHANES data.[7]

This is PHQ-9 test Interference

Over the last two weeks, how often have you been bothered by any of the following problems?

Little interest or pleasure in doing things?	**Not at all** Several days More than half the days Nearly every day
Feeling down, depressed, or hopeless?	**Not at all** Several days More than half the days Nearly every day
Trouble falling or staying asleep, or sleeping too much?	**Not at all** Several days More than half the days Nearly every day
Feeling tired or having little energy?	**Not at all** Several days More than half the days Nearly every day
Poor appetite or overeating?	**Not at all** Several days More than half the days Nearly every day
Feeling bad about yourself - or that you are a failure or have let yourself or your family down?	Not at all **Several days** More than half the days Nearly every day
Trouble concentrating on things, such as reading the newspaper or watching television?	**Not at all** Several days More than half the days Nearly every day
Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual?	**Not at all** Several days More than half the days Nearly every day
Thoughts that you would be better off dead, or of hurting yourself in some way?	Not at all Several days More than half the days **Nearly every day**
Total = 4 /27	Normal

Depression Severity: 0-4 none, 5-9 mild, 10-14 moderate, 15-19 moderately severe, 20-27 severe.

Fig No.10

A study showed that this screening tool has a specificity and sensitivity of 88% for major depression at a threshold score of 10 or more.[5] People were divided into "depressed" and "not depressed" categories based on the score for their answers in the screening tool with a score of 10 or more being "depressed".

Machine learning algorithms are trained and executed by use of python as its default programming language. Firstly depression dataset is prepared, in which data is extracted from the results of the patients who have undergone tests for depression. This dataset includes pregnancy, Marital Status, Education Level,etc. from the test and stored in data in variable (csv) format. Python contains some in-built libraries such as pandas, scikit learn, Seaborn which is exclusively used for training machine learning algorithms. Data is inputted into a depression ML system. After putting results of clients, the model will predict from past data, that this client is depressed or not.
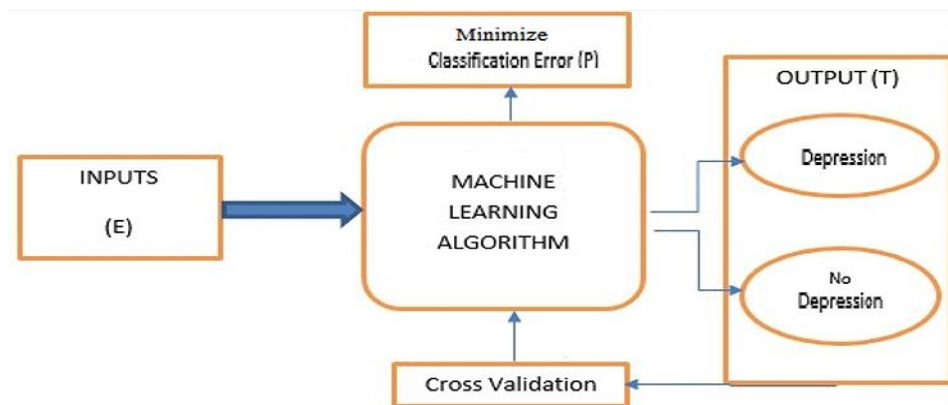

Fig No.11

The way the data was preprocessed with feature engineering, filling missing values, and scaling was done with the goal of increasing accuracy of the models. For each type of model, a model was first trained and fitted with default parameters as a base. Then, key parameters were chosen to tune using sklearn GridSearchCV and the best parameters were used to run the model. Finally, the tuned parameters were used to fit the same model using the resampled data for comparison. Performance was compared to the base model of each type, as well as between different model types.

12. Conclusion

In this paper, machine learning algorithms were applied to determine, a patient is in depression or not. The data was gathered using the PHQ-9 depression screening tool that was asked of all participants in the NHANES data. Subsequently, different types

of classification techniques were applied – Logistic Regression (LR), Decision Tree (DT), Random Forest Tree (RFT) .

The accuracy of Random Forest Tree (RFT) was found to be the highest and also Random Forest was identified as the best model. Due to the fact that this problem produced imbalanced classes, the best-model selection was made on the basis of the f1 score, which is used for cases of imbalanced partitioning.

13. References

1] The Mental Health Epidemic :-
https://thelogicalindian.com/mentalhealth/mental-health-indians-30811

2] Major Depression :- https://www.nimh.nih.gov/health/statistics/major-depression

3] Sonde Health :- Sonde Health

4] PHQ-9 Test :- https://pubmed.ncbi.nlm.nih.gov/11556941/

GitHub Link :- https://github.com/pratikdatey/Predicting-Depression.git