

MENTAL HEALTH PREDICTION USING MACHINE LEARNING

Submitted in partial fulfilment of the
requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

ADITYA SINGH (Reg.No - 40110035)
MURLIDHAR (Reg.No – 40110805)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

**Accredited with Grade “A++” by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119**

NOVEMBER - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with —A++ grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **ADITYA SINGH (Reg.No - 40110035)** who carried out the Project Phase-1 entitled “**MENTAL HEALTH PREDICTION USING MACHINE LEARNING**” under my supervision from June 2023 to November 2023.

Internal Guide

Dr. A.C SANTHA SHEELA M.E., Ph.D.

Head of the Department

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I, **ADITHYA SINGH (Reg.No- 40110035)**, hereby declare that the Project Phase-1 Report entitled **MENTAL HEALTH PREDICTION USING MACHINE LEARNING** done by me under the guidance of **Dr. A.C SANTHA SHEELA M.E., Ph.D.** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE:

PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph. D, Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. A.C Santha Sheela M.E., Ph.D.**, for her valuable guidance, suggestions, and constant encouragement paved way for the successful completion of my phase-1 project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

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 Kaggle.com and UCI repository , 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.

The results show that 71% of those who have depression and 80% of those who don't have depression can be correctly identified. Though more work needs to be done to create a more accurate model, this shows proof of concept that this is a realistic prediction task. Better results could be yielded by adding more patient information to the data or testing more types of models.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	i
	LIST OF FIGURES	iv
1.	INTRODUCTION	1 - 7
	1.1 INTRODUCTION	1
	1.2 OVERVIEW OF MACHINE LEARNING	1 - 2
	1.3 ESTIMATING DATASET INFORMATION	3
	1.4 PROBLEM LEARNING	3 - 4
	1.5 WORKING ENVIRONMENT	4 - 5
	1.6 ENVIRONMENT INSTALLATION	6
	1.7 INTRODUCTION TO PYCHARM	7
2.	AIM AND SCOPE OF THE PRESENT INVESTIGATION	8
	2.1 AIM	8
	2.2 SCOPE	8
	2.3 LITERATURE SURVEY	8

3.	EXPERIMENTAL OR AND METHODS, ALGORITHMS USED	9 - 16
	3.1 IMPORTED LIBRARIES	9 - 12
	3.2 DATA PREPARING	12 - 13
	3.3 MODEL TRAINING	14
	3.4 ALGORITHM USED	14 - 16
4.	CONCLUSIONS AND FUTURE WORKS	17
	5.1 CONCLUSION	17
	5.2 FUTURE WORKS	17
	REFERENCES	18

LIST OF FIGURES

FIGUR E NO.	FIGURE NAME	PAGE NO
1.2.1	FORMS OF ML	2
1.3.1	SURVEY DATASET	3
1.4.1	DATA DESCRIPTION	4
1.5.1	ANACONDA NAVIGATOR	5
1.5.2	JNB	5
1.6.1	CREATE ENV IN CONDA	6
1.7.1	PYCHARM	7
3.1.1	PYTHON LIBRARIES FORML	9
3.1.2	LIBRARIES	10
3.1.3	MATPLOTLIB	11
3.2.1	READING DATASET	12
3.2.2	DATA SHAPE	13

CHAPTER – 1

INTRODUCTION

1.1 Introduction

Mental Health First Aid teaches participants how to notice and support an individual who may be experiencing a mental health or substance use concern or crisis and connect them with the appropriate employee resources.

Employers can offer robust benefits packages to support employees who go through mental health issues. That includes Employee Assistance Programs, Wellness programs that focus on mental and physical health, Health and Disability Insurance, or flexible working schedules or time off policies. Organizations that incorporate mental health awareness help to create a healthy and productive work environment that reduces the stigma associated with mental illness, increases the organizations' mental health literacy, and teaches the skills to safely and responsibly respond to a co-worker's mental health concern.

We will be using classification algorithms such as Logistic Regression, KNN, Decision tree, Random Forest, AdaBoost, Gradient Boost, and XG Boost. We will train and test the data with these algorithms. From this, the best model is selected and saved in PKL format. We will also be deploying our model locally using Flask.

1.2 OVERVIEW OF MACHINE LEARNING (ML)

In general terms ML refers to enabling the machine to learn by itself from the experience like human do. As we know that machine understand only binary language (0 and 1), what ml does is it takes the raw input data and split the data into training and test set ,then trains the data and convert the output to binary form to machine understandable form . Here, machine understands the concept and learns

something new and takes decision or make prediction. In simple terms human like thinking. One example for ml is google maps it shows shortest route, when someone experienced with the route takes the shortest and reaches the destination in less time, this route is recorded in the google maps, the machine is learning from its experience. It focuses on studying and understanding data patterns and structures inorder to facilitate learning,reasoning, and decision-making without the involvement of a person.

The three forms of machine learning are supervised, unsupervised, and reinforcement .

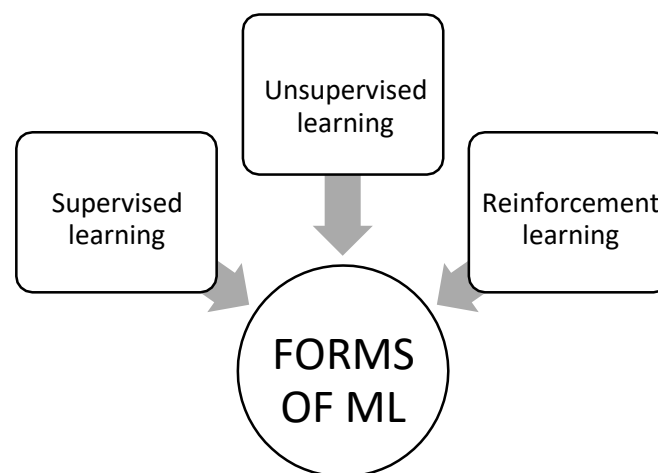


Fig 1.2.1 Forms of ml

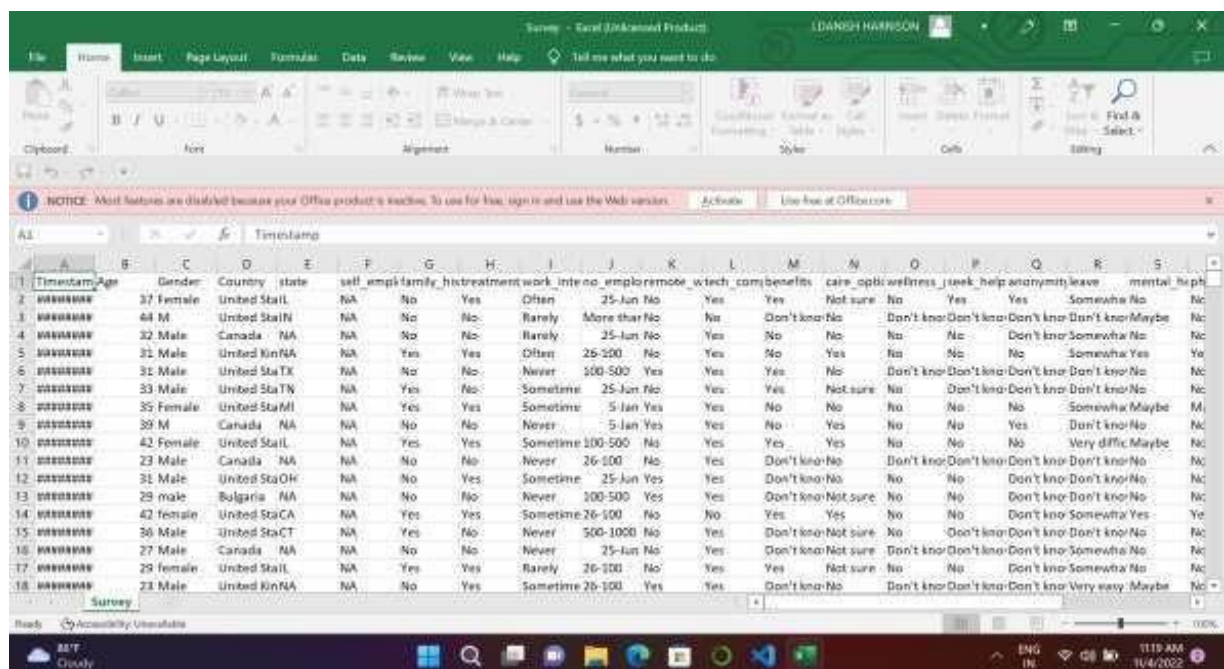
Supervised learning algorithms model the relationship between features (independent variables) and a label (target) from given set of observations. Then the model is used to predict the label of new observations using the features. Depending on the characteristics of the target variable i.e., it can be either be classification (discrete variable) or regression (continuous variable) the task is further engaged.

Unsupervised learning finds the structures in unlabeled data.

Reinforcement learning works on action-reward principle. An agent learns to react to the environment by continuously calculating the rewards that it gained from the actions.

1.3 ESTIMATING DATASET INFORMATION

This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. You are also encouraged to analyze data from the [ongoing 2016 survey found here](#).



Timestamp	Age	Gender	Country	State	self_employment	family_history	treatment_work	into_no_employment	mobile	which	com	benefits	care	opt	wellness	seek help	anonym	leave	mental health
1	37	Female	United States	IL	NA	No	Yes	Often	25-Jun	No	Yes	Yes	Not sure	No	Yes	Yes	Somewhat	No	No
2	44	M	United States	IN	NA	No	No	Rarely	More than 100-500	No	No	Don't know	No	Don't know	Don't know	Don't know	Don't know	Maybe	No
3	32	Male	Canada	NA	NA	No	No	Rarely	25-Jun	No	Yes	No	No	No	No	No	Don't know	Somewhat	No
4	31	Male	United Kingdom	NA	NA	Yes	Yes	Often	25-Jun	No	Yes	No	Yes	No	No	No	Don't know	Somewhat	Yes
5	31	Male	United States	TX	NA	No	No	Never	100-500	Yes	Yes	Yes	No	Don't know	Don't know	Don't know	Don't know	No	No
6	33	Male	United States	TX	NA	Yes	No	Sometime	25-Jun	No	Yes	Yes	Not sure	No	Don't know	Don't know	Don't know	No	No
7	35	Female	United States	MI	NA	Yes	Yes	Sometime	5-Jan	Yes	Yes	No	No	No	No	No	Somewhat	Maybe	MI
8	39	M	Canada	NA	NA	No	No	Never	5-Jan	Yes	Yes	No	Yes	No	No	Yes	Don't know	No	No
9	42	Female	United States	IL	NA	Yes	Yes	Sometime	100-500	No	Yes	Yes	Yes	No	No	No	Very difficult	Maybe	No
10	23	Male	Canada	NA	NA	No	No	Never	26-100	No	Yes	Don't know	No	Don't know	Don't know	Don't know	Don't know	No	No
11	31	Male	United States	OH	NA	No	Yes	Sometime	25-Jun	Yes	Yes	Don't know	No	No	No	Don't know	Don't know	No	No
12	29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes	Yes	Don't know	Not sure	No	No	Don't know	Don't know	No	No
13	42	Female	United States	CA	NA	Yes	Yes	Sometime	26-100	No	No	Yes	Yes	No	No	Don't know	Somewhat	Yes	Yes
14	36	Male	United States	CT	NA	Yes	No	Never	500-1000	No	Yes	Don't know	Not sure	No	Don't know	Don't know	Don't know	No	No
15	27	Male	Canada	NA	NA	No	No	Never	25-Jun	No	Yes	Don't know	Not sure	Don't know	Don't know	Don't know	Somewhat	No	No
16	29	female	United States	IL	NA	Yes	Yes	Rarely	26-100	No	Yes	Yes	Not sure	No	No	Don't know	Somewhat	No	No
17	23	Male	United Kingdom	NA	NA	No	Yes	Sometime	25-100	Yes	Yes	Don't know	No	Don't know	Don't know	Don't know	Very easy	Maybe	No

Fig 1.3.1 Survey Dataset

1.4 PROBLEM LEARNING

As we are going to classify the type of glass our problem type is **Classification**. From fig 1.2 dataset contains both features and label (output) and the type of problem of classification so the type of learning is going to be **Supervised learning**. This problem has to do with classifying the mental health based on its properties. Machine learning approaches such as Random Forest, ADA BOOST CLASSIFIER or any other classification strategy might be used to tackle this problem.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	Timestamp	Age	Gender	Country	state	self_empt	family_his	treatment	work_inte	no_employe	remote_w	tech_com	benefits	care_opti	xwellness	_seek	help	anonymi	leave	mental_high
2	#####	37	Female	United Sta	IL	NA	No	Yes	Often	25-Jun	No	Yes	Yes	Not sure	No	Yes	Yes	Somewha	No	No
3	#####	44	M	United Sta	IN	NA	No	No	Rarely	More than	No	No	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	Maybe	No
4	#####	32	Male	Canada	NA	NA	No	No	Rarely	25-Jun	No	Yes	No	No	No	No	Don't kno	Somewha	No	No
5	#####	31	Male	United Kin	NA	NA	Yes	Yes	Often	26-100	No	Yes	No	Yes	No	No	No	Somewha	Yes	Yes
6	#####	31	Male	United Sta	TX	NA	No	No	Never	100-500	Yes	Yes	Yes	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No
7	#####	33	Male	United Sta	TN	NA	Yes	No	Sometime	25-Jun	No	Yes	Yes	Not sure	No	Don't kno	Don't kno	Don't kno	No	No
8	#####	35	Female	United Sta	MI	NA	Yes	Yes	Sometime	5-Jan	Yes	Yes	No	No	No	No	No	Somewha	Maybe	MI
9	#####	39	M	Canada	NA	NA	No	No	Never	5-Jan	Yes	Yes	No	Yes	No	No	Yes	Don't kno	No	No
10	#####	42	Female	United Sta	IL	NA	Yes	Yes	Sometime	100-500	No	Yes	Yes	Yes	No	No	No	Very diffic	Maybe	No
11	#####	23	Male	Canada	NA	NA	No	No	Never	26-100	No	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Don't kno	No	No
12	#####	31	Male	United Sta	OH	NA	No	Yes	Sometime	25-Jun	Yes	Yes	Don't kno	No	No	No	Don't kno	Don't kno	No	No
13	#####	29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes	Yes	Don't kno	Not sure	No	No	Don't kno	Don't kno	No	No
14	#####	42	female	United Sta	CA	NA	Yes	Yes	Sometime	26-100	No	No	Yes	Yes	No	No	Don't kno	Somewha	Yes	Yes
15	#####	36	Male	United Sta	CT	NA	Yes	No	Never	500-1000	No	Yes	Don't kno	Not sure	No	Don't kno	Don't kno	Don't kno	No	No
16	#####	27	Male	Canada	NA	NA	No	No	Never	25-Jun	No	Yes	Don't kno	Not sure	Don't kno	Don't kno	Don't kno	Somewha	No	No
17	#####	29	female	United Sta	IL	NA	Yes	Yes	Rarely	26-100	No	Yes	Yes	Not sure	No	No	Don't kno	Somewha	No	No
18	#####	23	Male	United Kin	NA	NA	No	Yes	Sometime	26-100	Yes	Yes	Don't kno	No	Don't kno	Don't kno	Don't kno	Very easy	Maybe	No
Survey																				

Fig 1.4.1 Data Description

Now the major issue is dataset is unbalanced. These kinds of issues are known as unbalanced multiclass classification problems, and they need for both careful selection of machine learning models as well as appropriate assessment measure and predict mental health. The range are not same that's why the data is imbalanced.

1.5 WORKING ENVIRONMENT

ANACONDA NAVIGATOR is desktop GUI used to launch applications and also manage packages in one place.

Outlook of Anaconda Navigator:



Fig 1.5.1 Anaconda Navigator

CODING ENVIRONMENT: Jupyter notebook from the anaconda navigator is launched along with all the preinstalled packages for python.



Fig 1.5.2 J

1.6 ENVIRONMENT INSTALLATION

A conda environment is a directory that contains a specific collection of conda packages that you have installed. You cannot install packages into the default Anaconda environment. You must create your own environment before installing a new package into that environment. AEN is built on Anaconda, so you can install additional Python packages using conda or pip—both of which are included with Anaconda. A virtual environment is a tool that helps to keep dependencies required by different projects separate by creating isolated python virtual environments for them. This is one of the most important tools that most of the Python developers use. You can create a new environme.

Use the terminal or an Anaconda Prompt for the following steps:

1. To create an environment:

```
conda create --name myenv
```

Note

Replace `myenv` with the environment name.

2. When conda asks you to proceed, type `y` :

```
proceed ([y]/n)?
```

This creates the myenv environment in `/envs/` . No packages will be installed in this environment.

3. To create an environment with a specific version of Python:

```
conda create -n myenv python=3.9
```

4. To create an environment with a specific package:

```
conda create -n myenv scipy
```

Fig 1.6.1 create env in conda

1.7 INTRODUCTION TO PYCHARM

PyCharm is the most popular IDE for Python, and includes great features such as excellent code completion and inspection with advanced debugger and support for web programming and various frameworks. PyCharm is created by Czech company, Jet brains which focusses on creating integrated development environment for various web development languages like JavaScript and PHP.

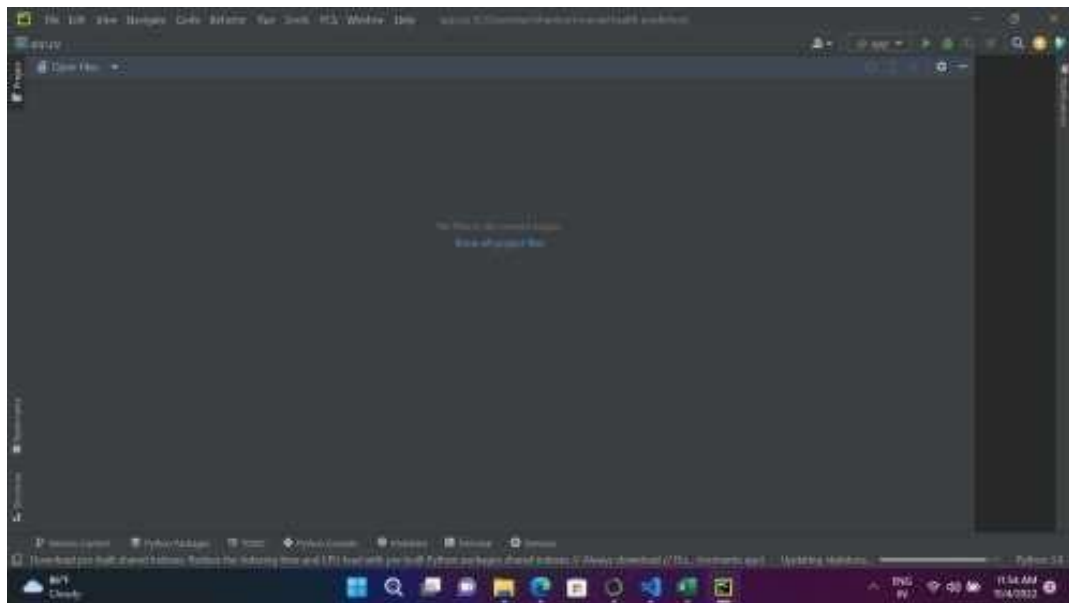


Fig 1.7.1 PYCHARM

CHAPTER – 2

AIM AND SCOPE OF THE PRESENT INVESTIGATION

2.1 AIM

To predict the 'Mental Health of Patients' in the given Survey of Data set using features from the given Data set

2.2 SCOPE

The SCOPE of this project is to gather data about people that would typically be in a patient's medical record to predict depression.

2.3 LITERATUTE SURVEY

Many clinics or doctors may find it impossible to have such integrated mental health services as cited in the previously mentioned study.

Having standard services where patients are constantly screened for mental health disorders and treatment is tightly integrated with teams of physicians and psychological professionals can be expensive, requires a lot of training, requires participation from many individual doctors that may feel too overwhelmed, and may also not be possible in certain areas due to various logistical factors.

Using machine learning and data that may otherwise be in a patient's medical file, the goal is to predict who may have depression in a way that requires very little human participation from doctors and has lower time and money costs associated

There are several methods for classification; thus, we must choose the best algorithms from the literature study and implement them in Python to categories various eyeglass kinds Vivencio et al [19] proposed utilizing a feature-weighted closest neighbour method with a KNN classifier.

CHAPTER – 3

EXPERIMENTAL OR MATERIALS AND METHODS, ALGORITHMS USED

3.1 IMPORTED LIBRARIES

Libraries are collections of prewritten code that users may utilise to optimise a process in computer languages. It has collections of code that may be utilised repeatedly in many projects.

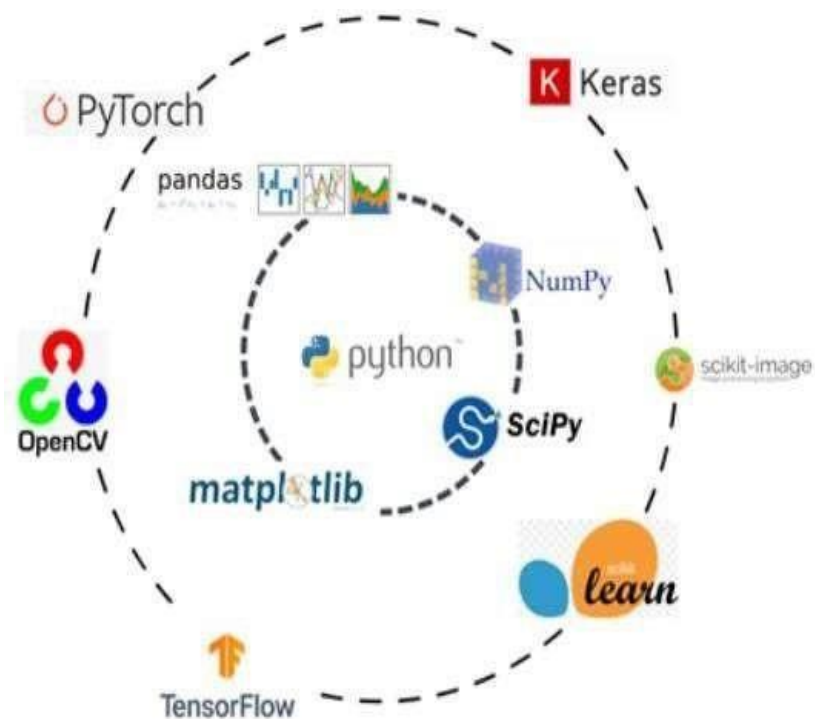


Fig 3.1.1 Python LibrariesforML

For the programmer, it simplifies and makes Python programming more simple and

easy. Since we don't have to write the same code for many apps repeatedly. Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries.

Python libraries that are used in Machine Learning for Glass Identification project are:

- Pandas
- Matplotlib
- Seaborn
- Numpy
- SciKit – Learn
- Collections

```
In [1]: #importing lib
import pandas as pd # read data
import numpy as np # to perform numerical calculation in pythonic way
import seaborn as sns # graphical plotting
import matplotlib.pyplot as plt # plot graph
from collections import Counter
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, confusion_matrix, f1_score, classification_report

import warnings
warnings.filterwarnings('ignore') # controls the warnings and check warnings are ignored
```

Fig 3.1.2 Libraries

PANDAS is a popular Python library for data analysis, It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for grouping, combining and filtering data.

MATPLOTLIB is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar charts, etc,

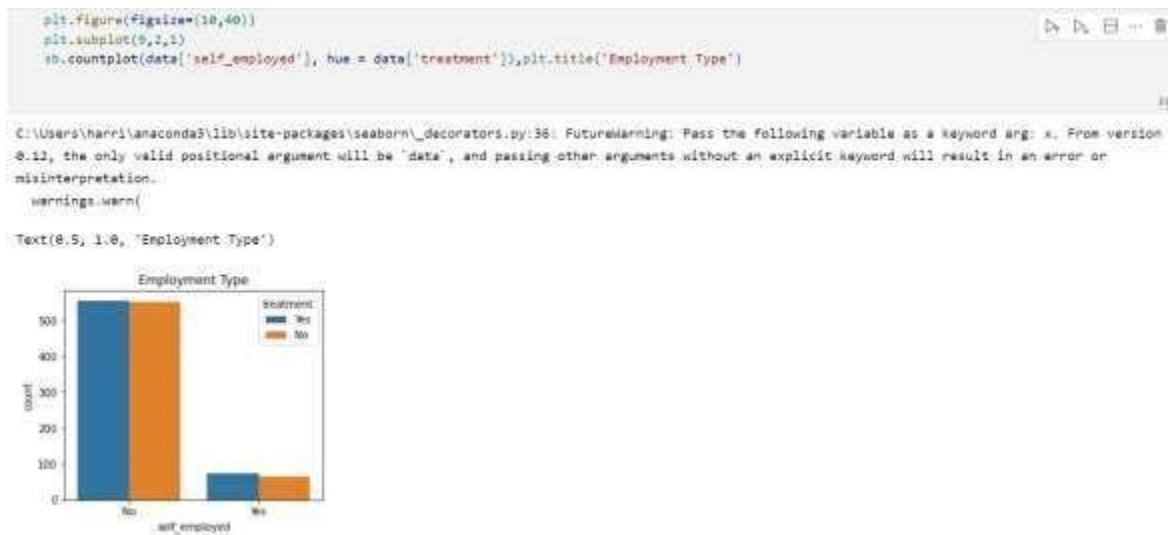


Fig 3.1.3 Matplotlib

NUMPY is a very popular python library for large multidimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities.

SCIKIT-LEARN is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with ML.

COLLECTIONS in Python are containers used for storing data and are commonly known as data structures, such as lists, tuples, arrays, dictionaries, etc. Python has a built-in collections module providing additional data structures for collections of data. Counter is a subclass of dict that's specially designed for counting hashable objects. Counter elements count can be positive, zero or negative integers. However there is no restriction on its keys and values. Although values are intended to be numbers but we can store other objects too.

3.2 DATA PREPARING

Python is employed as a backbone tool in this project to carry out machine learning ideas. A jupyter notebook is launched with the help of Anaconda Navigator, which is already installed alongside the navigator. Jupyter notebook is an open-source web tool that allows for the creation and sharing of documents, as well as the use of live code and visualisation. After importing the necessary libraries, the dataset will be read in the note book using data frame (two-dimensional labelled data structure with potentially multiple types of columns) and read csv (desired file type).The dataset Provided cannot always be a fully valuedset, in that case we need to prepare the data in such a way the machine understands the value that has been entered.



	Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	leave	mental_health_consequence
0	2014-08-27 11:29:31	37	Female	United States	IL	NaN	No	Yes	Often	6-25	Somewhat easy	No
1	2014-08-27 11:29:37	44	M	United States	IN	NaN	No	No	Rarely	More than 1000	Don't know	Maybe
2	2014-08-27 11:29:44	32	Male	Canada	NaN	NaN	No	No	Rarely	6-25	Somewhat difficult	No
3	2014-08-27 11:29:46	31	Male	United Kingdom	NaN	NaN	Yes	Yes	Often	26-100	Somewhat difficult	Yes
4	2014-08-27 11:30:22	31	Male	United States	TX	NaN	No	No	Never	100-500	Don't know	No

5 rows • 13 columns

Fig 3.2.1 Reading Dataset

Check for null values in the updated dataset using the `isnull()` or `info()` function.

```
data.info()

Output exceeds the size limit. Open the full output data in a text editor
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1259 entries, 0 to 1258
Data columns (total 27 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Timestamp              1259 non-null   object
1   Age                    1259 non-null   int64
2   Gender                 1259 non-null   object
3   Country                1259 non-null   object
4   State                  744 non-null    object
5   self_employed          1241 non-null   object
6   Family_History          1259 non-null   object
7   Treatment              1259 non-null   object
8   Work_Interface         995 non-null    object
9   No_Employees           1259 non-null   object
10  Remote_Work            1259 non-null   object
11  Tech_Company           1259 non-null   object
12  Benefits               1259 non-null   object
13  Care_Options           1259 non-null   object
14  Wellness_Program       1259 non-null   object
15  Seek_Help              1259 non-null   object
16  Anonymity              1259 non-null   object
```

The dataset is now ready for removing the outliers due to the unbalanced dataset.

Outlier - A value that is significantly smaller or greater than the majority of the other values in a set of data.

```
In [8]: data.shape
```

```
Out[8]: (1259, 27)
```

Fig 3.2.2 Data Shape

After the outlier removal the observation is reduced to 200 from 214. The data is now ready for creating ml models.

3.3 MODEL TRAINING

Train/Test is a technique for determining the accuracy of your model. It's named Train/Test because you divide the data set into two parts: training and testing. 80% for instruction and 20% for testing. The training set is used to train the model. The model is tested using the testing set. Create the model by training it. Testing the model entails determining the model's accuracy. The training set should be a random selection of 80% of the original data .The testing set should be the remaining 20%, splitting is done to reduce congestion and interference.

3.4 ALGORITHM USED

DecisionTreeAlgorithm

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decisiontrees are easy to understand and interpret.

The basic idea behind any decision tree algorithm is as follows:

- Select the best attribute using Attribute Selection Measures (ASM) to split the records.
- Create a decision node for that attribute and divide the dataset into smaller sections.

- Starts tree building by repeating this process recursively for each child until one of the conditions will match:
- All the tuples belong to the same attribute value.
- There are no more remaining attributes.
- There are no more instances.

Pros

- Decision trees are easy to interpret and visualize.
- It can easily capture Non-linear patterns.
- It requires fewer data pre-processing from the user, for example, there is no need to normalize columns.
- It can be used for feature engineering such as predicting missing values, suitable for variable selection.

Cons

- Sensitive to noisy data. It can over fit noisy data.
- The small variation (or variance) in data can result in the different decision tree. This can be reduced by bagging and boosting algorithms.
- Decision trees are biased with imbalance dataset, so it is recommended that balance out the dataset before creating the decision tree.

ADABOOST CLASSIFIER:

An ADABOOST algorithm can be used to boost the performance of any machine learning algorithm. Machine Learning has become a powerful tool which can make predictions based on a large amount of data. It has become so popular in recent times that the application of machine learning can be found in our day-to-day activities.

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called **Decision Stumps**.

AdaBoost stands for 'Adaptive Boosting', which transforms weak learners or predictors to strong predictors in order to solve problems of classification.

PROS:

- One of the many advantages of the AdaBoost Algorithm is it is fast, simple and easy to program.
- Also, it has the flexibility to be combined with any machine learning algorithm, and there is no need to tune the parameters except for T.
- It has been extended to learning problems beyond binary classification, and it is versatile as it can be used with text or numeric data.

CONS:

- AdaBoost also has few disadvantages, such as it is from empirical evidence and particularly vulnerable to uniform noise.
- Weak classifiers being too weak can lead to low margins and overfitting.

CHAPTER – 4

CONCLUSIONS AND FUTURE WORKS

5.1 CONCLUSION

We have modelled the Mental Health Prediction Dataset using AdaBoost classifier with 85% of accuracy to forecast Mental Health. The model with highest accuracy was deployed into a web app using streamlit app framework. The web app predicted the Mental health of patients using the input parameters. The web app created successfully predicted the mental health of patients.

5.2 FUTURE WORKS

Future work will be done to refine the model and make it more useful in realworld situations. Improve the suggested system's accuracy by

include more mental health questions in the model. The main difficulty now is that the dataset is imbalanced, and new features are needed to balance the dataset.

REFERENCES

- Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges([PDF](#))[Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges \(researchgate.net\)](#)
- B.German 2019). UCI Machine Learning Repository []vCentral Research Establishment Home Office Forensic Science Service Aldermeston
- <https://archive.ics.uci.edu/ml/datasets/Glass+Identification>
- For Learning Python Language <https://www.learnpython.org/>
- For library based Learning <https://www.coursera.org/articles/python-machine-learning-library>
- Project based learning [Mental Health in Tech Survey | Kaggle](#)