REMOVE PAGE NUMBERS

Project Report on

Mental Health Prediction using Machine Learning

Electronics & Telecommunication Engineering

(Semester VII)

Guided by

Dr. R.N. Awale

Submitted By

Kanchan Bhoyar 191091015

Isha Joshi 191091031

Madhura Raut 191091057

Mayuri Borkar 191091016

Tanvi Gurav 191091025

Electrical Engineering Department Veermata Jijabai Technological Institute

Mumbai 400 019

2022

CERTIFICATE

This is to certify that the project entitled **“Mental Health Prediction Using Machine Learning”** is a bonafide work of “**Madhura Raut” (191091057) ,Isha Joshi(191091031) , Tanvi Gurav(191091025) , Kanchan Bhoyar(191091015), Mayuri Borkar(191091016)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **“BTech”** in **“Electronics and Telecommunication Engineering”**.

full form EXTC B.Tech

(Name and sign)
Supervisor/Guide

(Name and sign)
Head of Department

Project Report Approval for B. E.

This project report entitled (**Mental Health Prediction Using Machine Learning**) by (**Madhura Raut" (191091057) ,Isha Joshi(191091031) , Tanvi Gurav(191091025) , Kanchan Bhoyar(191091015), Mayuri Borkar(191091016)**) is approved for the degree of BTech ExTC SEM VII .

B.Tech Electronics and telecommunicaton

Examiners

1.--

2.--

Date: 16/12/2022

Place: VJTI, Mumbai

Declaration

We declare that this written submission represents our my ideas in our my own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our my submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 16/12/2022

(Name of student and Roll No.)

Kanchan Bhoyar 191091015
Isha Joshi 191091031
Madhura Raut 191091057
Mayuri Borkar 191091016
Tanvi Gurav 191091025

Table of contents

ACKNOWLEDGEMENT

The desired project became a reality with the kind support and help of many individuals. We would like to extend my sincere regards to all of them.

We hereby take the opportunity to express deep gratitude to our Project Guide Dr. R.N. Awale and our Head of Department Dr. S J Bhosale for their cordial support, suggestions and for providing necessary guidance concerning project implementation.

Further, we would like to thank our institute Veermata Jijabai Technological Institute, Mumbai for giving us the opportunity of a semester long project in our final year providing us with the right platform to implement out practical knowledge. We would also like to thank our faculty, and lab assistants who supported us and gave their valuable suggestions.

The successful outcome of this project required a lot of guidance and assistance. We are extremely privileged to work on this project, which not only brushed up our existing skills but also helped us learn new concepts.

Finally, we wish to thank our colleagues for their constant help.

Abstract

Mental health has always been an important and challenging issue, especially in the case of working Professionals.

Reasons for mental health issues:

- Modernized (hectic) lifestyle and workload take a toll over people
- Prone to mental disorders like mood disorder and anxiety disorder
- Working professionals are under lots of pressure for reasons like peer pressure, short deadlines, competition. All these things contribute to building up mental stress.
- During the Covid-19 pandemic, mental health has been one of the most prominent issues, with loneliness and depression

We process data to find the features influencing the mental health of employees or features that can help to predict the mental health of the employee. The feature can be either personal or professional. We apply multiple machine learning algorithms to find the model with the best accuracy. We take precision and recall as the measure to check the performance of different ML models.

1. Introduction

In the US 18% of the working population which is 40 million people suffers from mental health disorder. We can classify mental health disorders into two types’ first Mood disorder and second Anxiety disorder. The mood disorder is serious changes in mood in the form of emotional inconsistency or abrupt changes/ amplification of certain specific emotions, which can be feeling extremely sad or feeling irritable. When they start interfering or disrupting normal life activities we call it mood disorder, in the case of working professionals the disruption is in form of work performance like having a hard time completing deadlines. We can further characterize mood disorder in forms depression disorder, mania, hypomania and bipolar disorder. Depression disorder as the name suggests is synonymous to depression, negative feelings start influencing the daily life of a person. Which includes mood swings, losing interest in daily life events, feeling apathy, hard time sleeping, losing appetite or feeling overly hungry. Depression in later phases may lead to suicidal tendencies. Mania is a hyperactive state of a person where a person has an excess of both physical and mental energy, whose symptoms are feeling restless and failing to sit still, easily getting distracted, not able focus on a particular thing, having a hard time sleeping, talking too much. These things may seem trivial but can prove fatal in work-life; mania makes us more susceptible to taking risks, for example taking up a large number of projects with the inability to complete them, the reason being the increase in grandiosity due to mania. In severe cases, mania can get people hospitalized. Hypomania is a milder version of mania. Bipolar disorder is the transition between depression and mania bipolar disorder is usually detected using mania and hypomania symptoms. Anxiety is an emotional response to a future event, which is usually in the form of fear. Anxiety is a normal emotion but an anxiety disorder is an entirely different case. In which we feel an excessive amount of fear and anxiety for no reason whatsoever, excess anxiety can make people skip meetings, avoid social interactions and much more. We can be further categorized as General anxiety disorder (GAD), phobia anxiety disorder, and social anxiety disorder.

- To deal with this problem industries provide mental health care incentives to their employees, but it is not enough to deal with the problem.
- We process data to find the features influencing the mental health of employees or features that can help to predict the mental health of the employee .The feature can be either personal or professional.
- Various Machine Learning techniques were applied to train our model after due data cleaning and preprocessing. The accuracy of the above models was obtained and studied comparatively. By using Decision Trees, prominent features that influence stress were identified as gender, family history and availability of health benefits in the workplace. With these results, industries can now narrow down their approach to reduce stress and create a much comfortable workplace for their employees.

- The dataset contains 27 attributes, which contain both personal and professional features of the employees. In the feature selection part of data processing, we select 24 attributes , which contribute most to the mental disorder or can be used to predict the mental health disorder. After the feature selection comes the classification part in which different ML algorithms are applied to the selected features to predict the mental disorders.

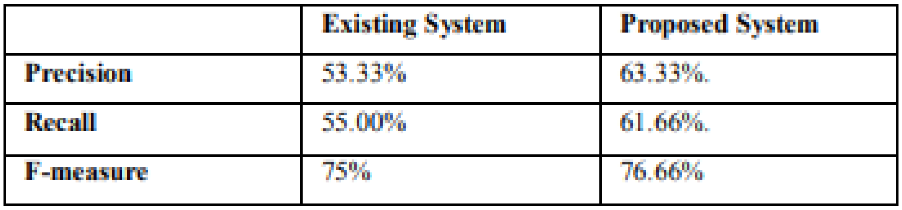
Open Sourcing Mental Illness (OSMI) is a non-profit organization that promotes awareness towards mental illness, disorders in the workspace and fights to eradicate the stigma surrounding them. They also help workplaces to identify the best resources to help their employees in this aspect.Using dataset we trained different machine learning models in order to analyze the patterns of stress and mental health disorders among tech professionals and to determine the most influential factors that contribute to the same.

2.Literature Survey

1. *S. Dmonte, G. Tuscano, L. Raut, and S. Sherkhane, "Rule generation and prediction of Anxiety Disorder using Logistic Model Trees," 2018 Int. Conf. Smart City Emerg. Technol. ICSCET 2018, 2018, doi: 10.1109/ICSCET.2018.8537258.*

   In a paper for predicting Anxiety disorder, the author proposed rules based on factors like a person's working place(home or office) and some other personal factors, and the prediction is done using logistic model trees. Which is a hybrid model based on logistic regression and decision trees, and provides better accuracy.

Table 1: Comparison between existing system and proposed system



	Existing System	Proposed System
Precision	53.33%	63.33%.
Recall	55.00%	61.66%.
F-measure	75%	76.66%

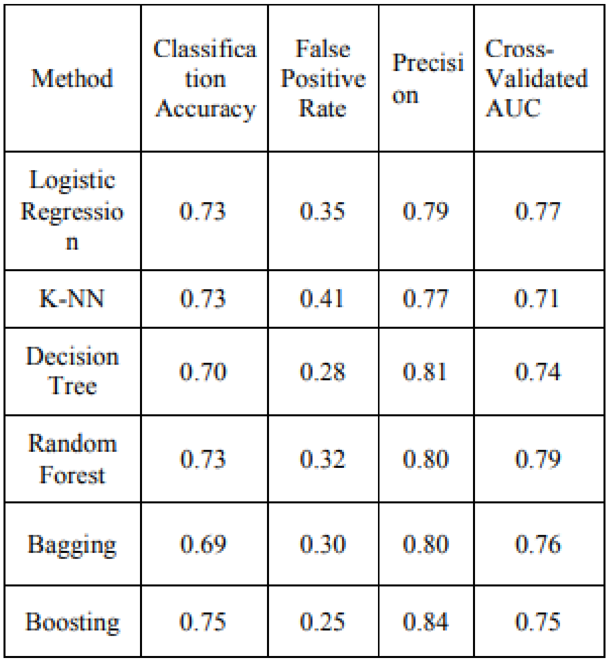
2. *Theodor Chris Panagiotakopoulos, Dimitrios Panagiotis Lyras, Miltos Livaditis, Kyriakos N. Sgarbas, "A Contextual Data Mining Approach Towards Assisting the Treatment of Anxiety Disorders", IEEE transactions on information technology in biomedicine, vol. 14, no. 3, pp. 567-581, May 2010*

   In a contextual data mining approach towards assisting the treatment of anxiety disorders, an application that archiving and retrieving patient's health records was developed. Four treatment supportive services were developed. The three of them concentrated on the discovery of possible associations between the patient's contextual data using apriori association rule algorithm whereas the last service focused on predicting the stress level a patient might suffer from, in a given context. Stress level of the patient has been determined based on their context parameters using a Bayesian network. This author has proposed to use decision trees instead of association rule which can be used in order to better distinguish between each patient's stress provoking contexts and environmental settings that result in serenity.

3. *U. S. Reddy, A. V. Thota, and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," 2018 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2018, pp. 1–4, 2018, doi: 10.1109/ICCIC.2018.8782395.*

   In this paper the author used ML algorithms to detect stress in working employees and found features which contribute to mental stress. Boosting was found to have the highest precision and accuracy with 84%

Table 1: Performance of different models trained.



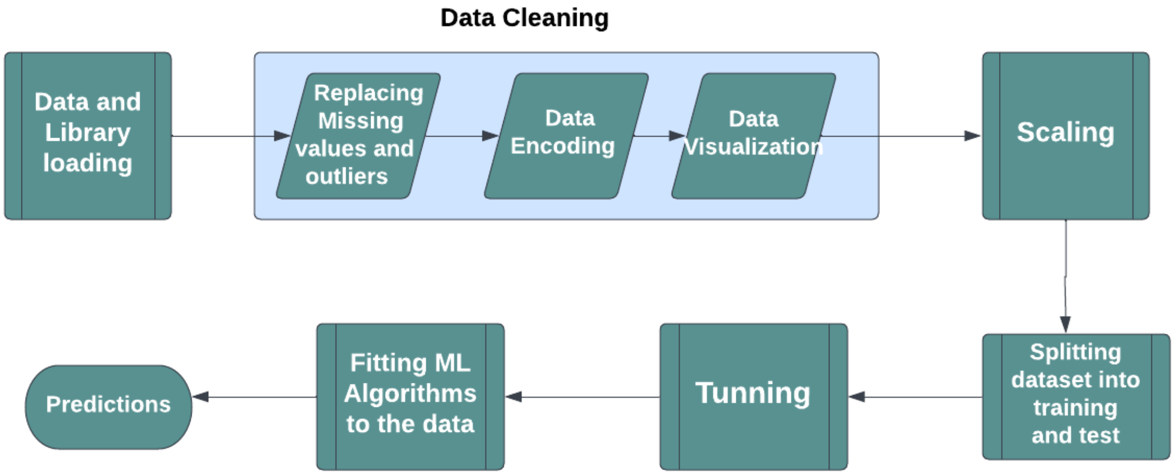
Method	Classification Accuracy	False Positive Rate	Precision	Cross-Validated AUC
Logistic Regression	0.73	0.35	0.79	0.77
K-NN	0.73	0.41	0.77	0.71
Decision Tree	0.70	0.28	0.81	0.74
Random Forest	0.73	0.32	0.80	0.79
Bagging	0.69	0.30	0.80	0.76
Boosting	0.75	0.25	0.84	0.75

3. Problem Statement

- Changing lifestyles and work culture increases the risk of stress among employees.The risk of mental health problems increases due to depression and anxiety.
- During and after the covid 19 pandemic an increasing amount of people are facing mental health related situations
- Research has shown that there are a lot of indicators of deteriorating mental health like Change in feelings or demeanor, Loss of interest, Change in sleeping habits .
- Early detection of a mental illness can prove to be essential
- Hence this projects aims to predict whether a person needs treatment for a mental illness or not based on given input features

4. Proposed Solution

4.1 Block Diagram


Data Cleaning
Data and Library loading
Replacing Missing values and outliers
Data Encoding
Data Visualization
Scaling
Splitting dataset into training and test
Tunning
Fitting ML Algorithms to the data
Predictions

1. Data and library loading :

   In this step all the necessary datasets and libraries are imported

2. Data cleaning:

   Replacing missing values and outliers

   Data encoding

   Data visualization

3. Scaling Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
4. Splitting the data in training set and test set
5. Tuning
6. Fitting ML algorithms to the dataset
7. Predictions.

4.2 Software requirements

- This project uses python as the main programming language .
- Python libraries used in this project are :
- NumPy:

1. NumPy is a Python library used for working with arrays.It also has functions for working in the domain of linear algebra, fourier transform, and matrices.
2. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.
3. NumPy stands for Numerical Python.NumPy is a Python library and is written partially in Python, but most of the parts that require fast computation are written in C or C++.
4. NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently.
5. NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.
6. Arrays are very frequently used in data science, where speed and resources are very important.This behavior is called locality of reference in computer science.
7. This is the main reason why NumPy is faster than lists. Also it is optimized to work with the latest CPU architectures.

- Pandas:

1. Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data.
2. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.
3. Pandas allows us to analyze big data and make conclusions based on statistical theories.Pandas can clean messy data sets, and make them readable and relevant.Relevant data is very important in data science.It offers data structures and operations for manipulating numerical tables and time series.

- Seaborn:

1. Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive.
2. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.
3. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for the same variables for better understanding of the dataset.
4. Different categories of plot in Seaborn

- Plots are basically used for visualizing the relationship between variables. Those variables can either be completely numerical or a category like a group, class or division. Seaborn divides plot into the below categories –
- Relational plots: This plot is used to understand the relation between two variables.
- Categorical plots: This plot deals with categorical variables and how they can be visualized.
- Distribution plots: This plot is used for examining univariate and bivariate distributions
- Regression plots: The regression plots in seaborn are primarily intended to add a visual guide that helps to emphasize patterns in a dataset during exploratory data analyses.
- Matrix plots: A matrix plot is an array of scatterplots.
- Multi-plot grids: It is a useful approach to draw multiple instances of the same plot on different subsets of the dataset.

- Sklearn:

1. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

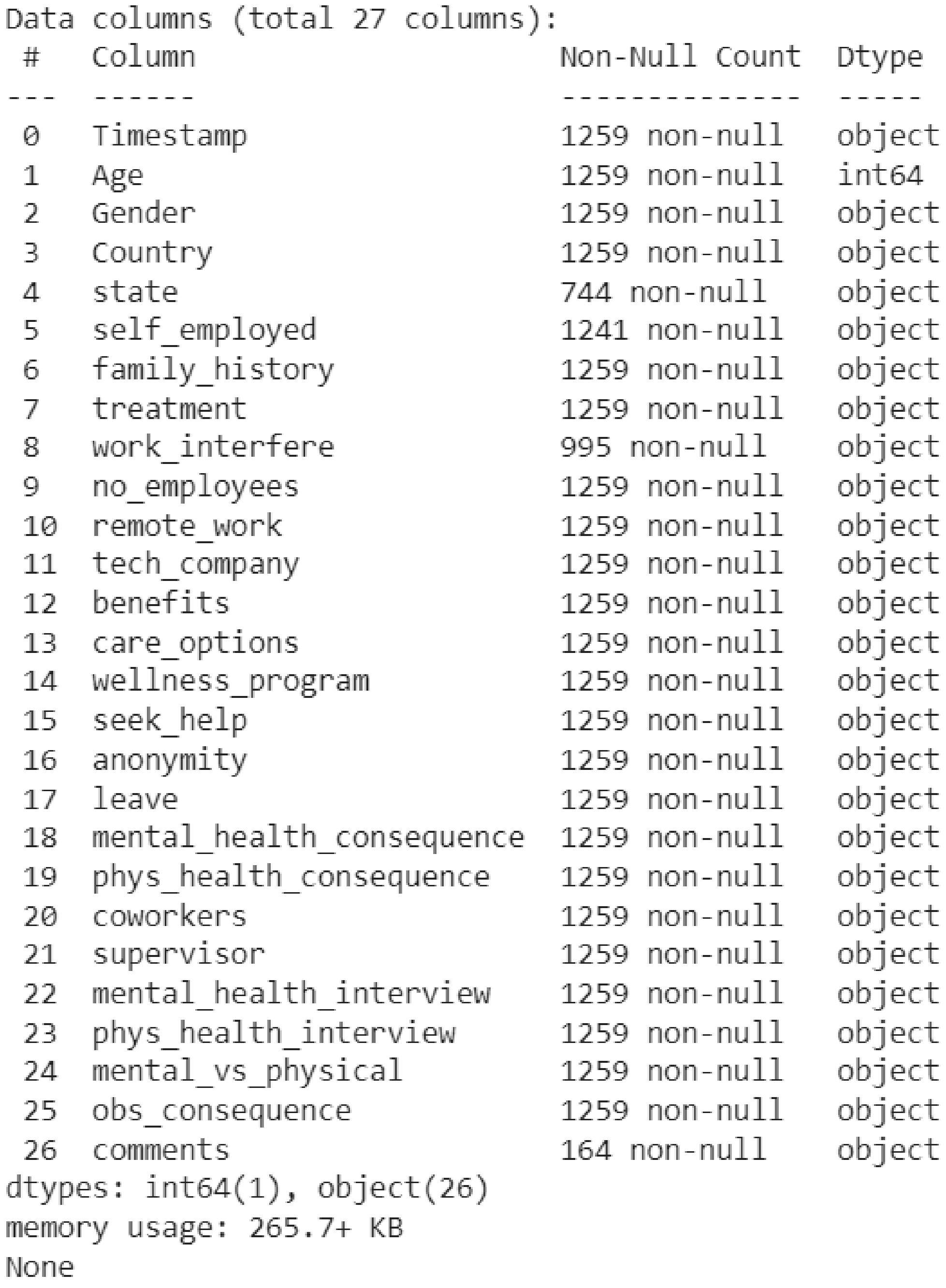
2. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.The scikit-learn project started as scikits.learn, a Google Summer of Code project by French data scientist David Cournapeau.
3. Its name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately-developed and distributed third-party extension to SciPy.
4. The original codebase was later rewritten by other developers. In 2010 Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort and Vincent Michel, all from the French Institute for Research in Computer Science and Automation in Saclay, France, took leadership of the project and made the first public release on February the 1st 2010. Of the various scikits, scikit-learn as well as scikit-image were described as "well-maintained and popular" in November 2012 Scikit-learn is one of the most popular machine learning libraries on GitHub.
5. Scikit-learn is largely written in Python, and uses NumPy extensively for high-performance linear algebra and array operations. Furthermore, some core algorithms are written in Cython to improve performance.
6. Support vector machines are implemented by a Cython wrapper around LIBSVM; logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR. In such cases, extending these methods with Python may not be possible.
7. Scikit-learn integrates well with many other Python libraries, such as Matplotlib and plotly for plotting, NumPy for array vectorization, Pandas dataframes, SciPy, and many more.

- matplotlib:

1. Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.
2. It was introduced by John Hunter in 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals.
3. Matplotlib consists of several plots like line, bar, scatter, histogram etc.Matplotlib comes with a wide variety of plots.

4.3 Dataset Preprocessing

- The dataset currently used in this project is taken from an anonymous survey conducted within a corporate organization after covid-19 pandemic .The dataset contains a total 27 columns and 1259 entries.
- The 27 features are :



```
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_interfere             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
 17  leave                      1259 non-null   object
 18  mental_health_consequence  1259 non-null   object
 19  phys_health_consequence    1259 non-null   object
 20  coworkers                  1259 non-null   object
 21  supervisor                 1259 non-null   object
 22  mental_health_interview    1259 non-null   object
 23  phys_health_interview      1259 non-null   object
 24  mental_vs_physical         1259 non-null   object
 25  obs_consequence            1259 non-null   object
 26  comments                   164 non-null    object
dtypes: int64(1), object(26)
memory usage: 265.7+ KB
None
```

4.4 Data Cleaning

Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn’t the fanciest part of machine learning and at the same time, there aren’t any hidden tricks or secrets to uncover. However, the success or failure of a project relies on proper data cleaning. Professional data scientists usually invest a very large portion of their time in this step because of the belief that “Better data beats fancier algorithms”.

If we have a well-cleaned dataset, there are chances that we can achieve good results with simple algorithms also, which can prove very beneficial at times especially in terms of computation when the dataset size is large.

Obviously, different types of data will require different types of cleaning. However, this systematic approach can always serve as a good starting point.

- Removal of unwanted observations
  This includes deleting duplicate/ redundant or irrelevant values from your dataset. Duplicate observations most frequently arise during data collection and Irrelevant observations are those that don’t actually fit the specific problem that you’re trying to solve.
  - Redundant observations alter the efficiency by a great extent as the data repeats and may add towards the correct side or towards the incorrect side, thereby producing unfaithful results.
  - Irrelevant observations are any type of data that is of no use to us and can be removed directly.

- Fixing Structural errors

  The errors that arise during measurement, transfer of data, or other similar situations are called structural errors. Structural errors include typos in the name of features, the same attribute with a different name, mislabeled classes, i.e. separate classes that should really be the same, or inconsistent capitalization.

  - For example, the model will treat America and America as different classes or values, though they represent the same value or red, yellow, and red-yellow as different classes or attributes, though one class can be included in the other two classes.
  - Managing Unwanted outliers

    Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models. Generally, we should not remove outliers until we have a legitimate reason to remove them. Sometimes, removing them improves performance, sometimes not. So, one must have a good reason to remove the outlier, such as suspicious measurements that are unlikely to be part of real data.
- Handling missing data

  Missing data is a deceptively tricky issue in machine learning. We cannot just ignore or remove the missing observation. They must be handled carefully as they can be an indication of something important. The two most common ways to deal with missing data are:
- Dropping observations with missing values.The fact that the value was missing may be informative in itself.Plus, in the real world, you often need to make predictions on new data even if some of the features are missing. Imputing the missing values from past observations.Again, “missingness” is almost always informative in itself, and you should tell your algorithm if a value was missing.

- So, missing data is always informative and an indication of something important. And we must be aware of our algorithm of missing data by flagging it. By using this technique of flagging and filling, you are essentially allowing the algorithm to estimate the optimal constant for missingness, instead of just filling it in with the mean.

4.5 Data Encoding

- The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model. Since most machine learning models only accept numerical variables, preprocessing the categorical variables becomes a necessary step. We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information.
- Categorical variables are usually represented as ‘strings’ or ‘categories’ and are finite in number. Further, we can see there are two kinds of categorical data-
- Ordinal Data: The categories have an inherent order
- Nominal Data: The categories do not have an inherent order

- In Ordinal data, while encoding, one should retain the information regarding the order in which the category is provided. Like in the above example the highest degree a person possesses, gives vital information about his qualification. The degree is an important feature to decide whether a person is suitable for a post or not.
- While encoding Nominal data, we have to consider the presence or absence of a feature. In such a case, no notion of order is present.
- Label Encoding or Ordinal Encoding
- We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence.
- In Label encoding, each label is converted into an integer value.

4.6 Data splitting

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model.

How data splitting works?

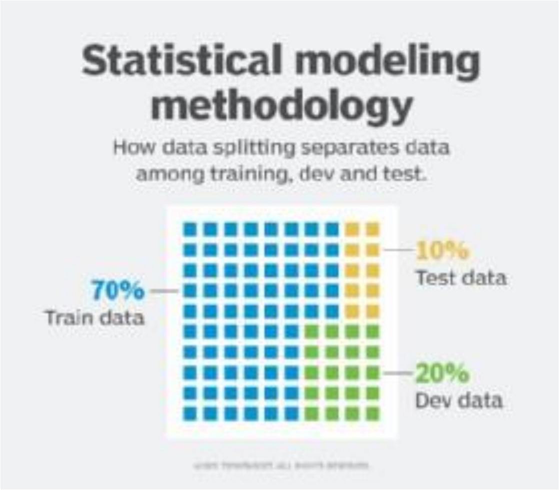
In a basic two-part data split, the training data set is used to train and develop models. Training sets are commonly used to estimate different parameters or to compare different model performance.

The testing data set is used after the training is done. The training and test data are compared to check that the final model works correctly. There is no set guideline or metric for how the data should be split; it may depend on the size of the original data pool or the number of predictors in a predictive model.

Data splitting in machine learning

In machine learning, data splitting is typically done to avoid overfitting. That is an instance where a machine learning model fits its training data too well and fails to reliably fit additional data. The original data in a machine learning model is typically taken and split into three or four sets. The three sets commonly used are the training set, the dev set and the testing set:

1. The training set is the portion of data used to train the model. The model should observe and learn from the training set, optimizing any of its parameters.
2. The dev set is a data set of examples used to change learning process parameters. It is also called the cross-validation or model validation set. This set of data has the goal of ranking the model's accuracy and can help with model selection.
3. The testing set is the portion of data that is tested in the final model and is compared against the previous sets of data. The testing set acts as an evaluation of the final mode and algorithm.



Data should be split so that data sets can have a high amount of training data. For example, data might be split at an 80-20 or a 70-30 ratio of training vs. testing data. The exact ratio depends on the data, but a 70-20-10 ratio for training, dev and test splits is optimal for small data sets.

4.7 Hyperparameter Tuning

All machine learning algorithms have a “default” set of hyperparameters, which Machine Learning Mastery defines as “a configuration that is external to the model and whose value cannot be estimated from data.” Different algorithms consist of different hyperparameters. For example, regularized regression models have coefficients penalties, decision trees have a set number of branches, and neural networks have a set number of layers. When building models, analysts and data scientists choose the default configuration of these hyperparameters after running the model on several datasets.

While the generic set of hyperparameters for each algorithm provides a starting point for analysis and will generally result in a well-performing model, it may not have the optimal configurations for your particular dataset and business problem. In order to find the best hyperparameters for data, need to tune them.

Model tuning allows you to customize your models so they generate the most accurate outcomes and give you highly valuable insights into your data, enabling you to make the most effective business decisions.

Parameters which define the model architecture are referred to as hyperparameters and thus this process of searching for the ideal model architecture is referred to as *hyperparameter tuning*.

Hyperparameter tuning methods:-

Grid search

Grid search is arguably the most basic hyperparameter tuning method. With this technique, we simply build a model for each possible combination of all of the hyperparameter values provided, evaluating each model, and selecting the architecture which produces the best results.

Random search

Random search differs from grid search in that we longer provide a discrete set of values to explore for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values may be randomly sampled.

Bayesian optimization

The previous two methods performed individual experiments building models with various hyperparameter values and recording the model performance for each. Because each experiment was performed in isolation, it's very easy to parallelize this process. However, because each experiment was performed in isolation, we're not able to use the information from one experiment to improve the next experiment. Bayesian optimization belongs to a class of sequential model-based optimization (SMBO) algorithms that allow one to use the results of our previous iteration to improve our sampling method of the next experiment.

4.8 Algorithms Used

ml approches

A machine learning (ML) algorithm is a process or a set of instructions that help create and adapt a model suitable for achieving a given objective. Algorithms are designed to train the model to adapt to values given in the dataset to produce the desired results.

1. KNN (K- Nearest Neighbors)

   KNN is a non-parametric, lazy learning algorithm. Its aim is to predict a new sample point using the dataset which is divided into different classes. In other words, KNN is based on feature similarity i.e. how close do the sample features resemble each other of the training set helps

   determine the given data point. KNN classifier is a distance-based classifier that classifies instances on the basis of likeness. Being non-parametric means that it does not make any assumptions on how the data is classified, also it is a lazy learning algorithm implying that it does not have a training phase or is minimum.

2. Logistic Regression

   Logistic regression is the suitable regression analysis to use when the dependent variable is dichotomous (binary). Logistic regression is a predictive analysis method. It is used when the dependent variable(target) is of categorical nature. It uses maximum likelihood estimation as a method of approximation. Types of logistic regression:

   1. Binary Logistic Regression: target variable has only two possible outcomes.
   2. Multinomial Logistic Regression: target variable has three or more nominal categories.
   3. Ordinal Logistic Regression: target variable has three or more ordinal categories.

3. Decision Tree

   A decision tree is a supervised learning algorithm and is widely used in classification problems. Both input and output variables can be categorical or continuous, and the decision tree can be implemented with both types of variables. In this method, the basic foundational idea is to segregate the whole sample into multiple sub-samples based on a splitting criterion. Target variable determines the type of tree. It is classified into two types:

   1. . Categorical variable decision tree: When the target variable is categorical in nature, the resulting decision tree is called categorical variable decision tree.
   2. Continuous variable decision tree: When the target variable is continuous in nature, the resulting decision tree is called a continuous variable decision tree.

   A tree's accuracy is majorly affected by the way it determines strategic splits Decision trees use multiple algorithms to decide on how to divide the input samples via the differentiator. Each internal node in a tree is called an attribute, whereas each leaf node is called a class label.

4. Random Forest

   Random Forest is an adaptable machine-learning algorithm. It can perform classification and regression problems both. It can also help in dimension reduction, missing values, and extreme values and undergoes other important processes of data exploration, and does a decent task. It is an ensemble learning algorithm, in this, a group of other models assembles to get a powerful more accurate model. It can identify the most important variables therefore it is considered a dimensions reduction method. Further, the model results the significance of variables. In this project we compare the results obtained by training and testing on all features of the obtained dataset and then by training and testing them only using important features.

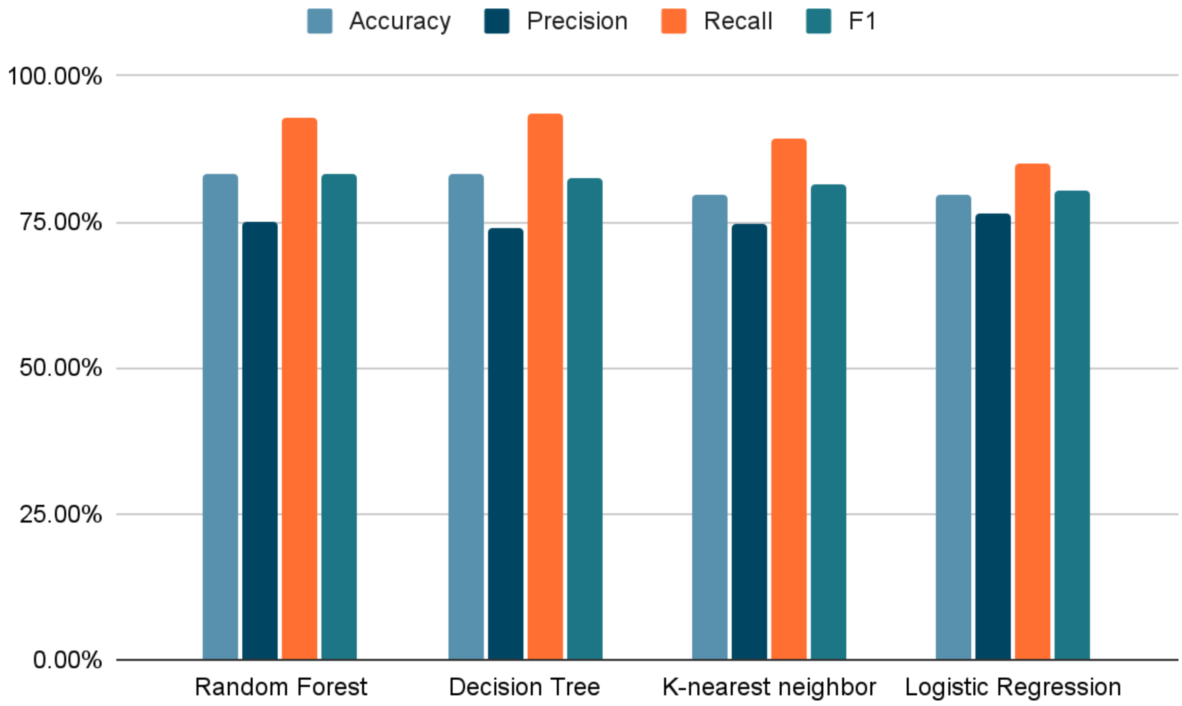
5. Conclusion

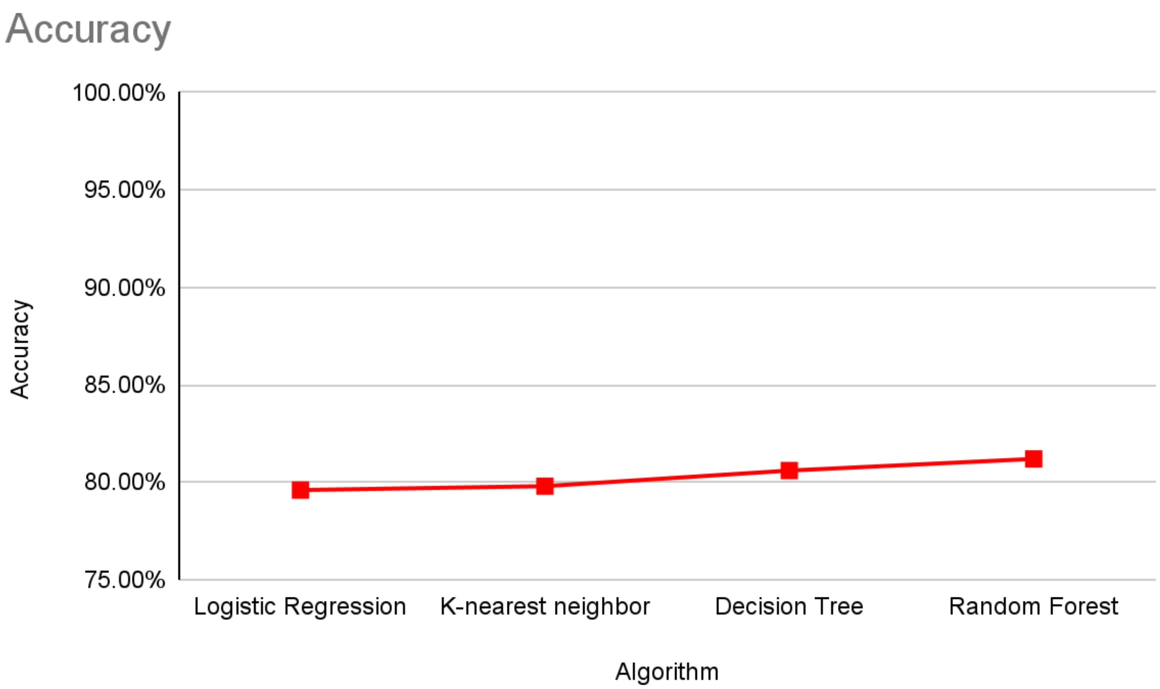
After applying various machine learning algorithms when predicting mental health disorders with selected attributes, we got following results:

Sr.no	Algorithm	Accuracy	Precision	Recall	F1
1	Logistic Regression	79.60%	76.40%	85.00%	80.47%
2	K-nearest neighbor	79.80%	74.80%	89.30%	81.41%
3	Decision Tree	80.60%	74.10%	93.50%	82.68%
4	Random Forest	81.20%	75%	93%	83.04%

Comparison of models

- By comparing the models used so far,it has been observed that Random Forest and has the highest accuracy followed by Decision Tree Algorithm
- We also found that the highest precision value is given by Random forest while decision tree shows higher recall.
- Precision considers false positives while recall considers false negatives.
- In this case study, our main focus is on precision and recall tradeoff as we have to detect mental health conditions.
- Ideally in a good classifier, we want both precision and recall to be one which also means FP and FN are zero. Therefore we need a metric that takes into account both precision and recall. F1-score is a metric which takes into account both precision and recall.
- So, Random Forest has the best performance out of all classifiers as it has higher accuracy and higher F1 score.


Accuracy
Precision
Recall
F1
100.00%
75.00%
50.00%
25.00%
0.00%
Random Forest
Decision Tree
K-nearest neighbor
Logistic Regression


Accuracy
100.00%
95.00%
90.00%
85.00%
80.00%
75.00%
Accuracy
Logistic Regression
K-nearest neighbor
Decision Tree
Random Forest
Algorithm

6. Future Scope

- Mental health is a critical element of public health.
- It is integral for living a healthy and balanced life.
- Mental health impacts one’s thoughts, behavior and emotions. It can affect the productivity and effectiveness of an individual.
- Early forecast and identification is a crucial step towards actualizing avoidance and clinical mediation policies.
- Statistical models, similar to the one created in this project, can provide a technically advanced approach to the medical science for sensing, as well as treating mental health related disorders among the patients, which is mainly dependent on data that is reported by the patient itself.
- After analyzing , we found that Random Forest has the highest accuracy
- Also, Feature importance of selected features showed that Age contributes the most during disorder prediction followed by Gender.
- It was found that rest of features contribute to bare minimum to the prediction
- But to further prove this we need more data
- For future scope, neural networks, deep learning and hybrid classifiers can be used for improving the accuracy when predicting mental health conditions.

7. References

1. S. Dmonte, G. Tuscano, L. Raut, and S. Sherkhane, “Rule generation and prediction of Anxiety Disorder using Logistic Model Trees,” 2018 Int. Conf. Smart City Emerg. Technol. ICSCET 2018, 2018, doi: 10.1109/ICSCET.2018.8537258.

2. Theodor Chris Panagiotakopoulos, Dimitrios Panagiotis Lyras, Miltos Livaditis, Kyriakos N. Sgarbas, “A Contextual Data Mining Approach Towards Assisting the Treatment of Anxiety Disorders”, IEEE transactions on information technology in biomedicine, vol. 14, no. 3, pp. 567-581, May 2010

3. U. S. Reddy, A. V. Thota, and A. Dharun, “Machine Learning Techniques for Stress Prediction in Working Employees,” 2018 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2018, pp. 1–4, 2018, doi: 10.1109/ICCIC.2018.8782395.