PREDICTION OF DEPRESSION STRESS AND ANXIETY

Tejveer Singh

IIT(ISM) Dhanbad MSc Mathematics and Computing

ABSTRACT

Over the generations mental health has become a major concern among people worldwide. In this paper a try has been made to predict the situation of people suffering from Depression, Stress, Anxiety and the level of suffering i.e., severe, mild or not suffering using different Machine Learning Algorithms on the dataset from an online questionnaire in which nearly 42K people among different parts of world participated. For the model to create firstly the data was cleaned and visualized properly and therefore the model was created and later checked by different evaluation methods.

CONTENT

1. INTODUCTION

2. DATASET COLLECTION AND DESCRIPTION

3. METHODOLGY

4. MODEL CREATION

4.1 MACHINE LEARNING

4.1.1 SUPERVISED LEARNING

4.1.2 SEMI-SUPERVISED LEARNING

4.1.3 UN-SUPERVISED LEARNING

4.1.4 REINFORCEMENT LEARNING

4.2.1 RANDOM FOREST

4.2.2 NAÏVE BAYES

4.2.3 SUPPORT VECTOR MACHINES

4.2.4 ARTIFICIAL NEURAL NETWORK

5 HYPER PARAMETER OPTIMIZATION

6 MODEL EVALUATION

6.1 CROSS VALIDATION

6.2 CONFUSION MATRIX

6.3 PRECISION AND RECALL

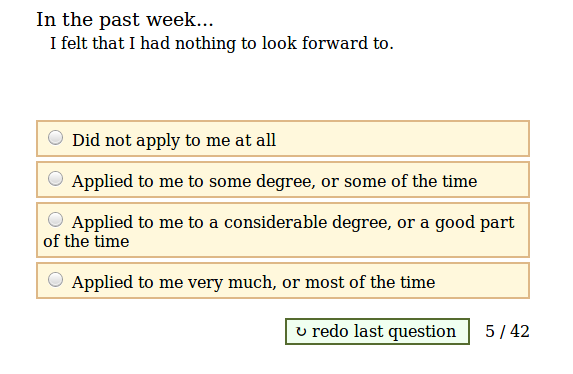
6.4 AUC SCORE

7 PLOTS

8 CONCLUSION

9 REFERENCES

INTRODUCTION

Depression Anxiety and Stress are some of the major reasons for the past suicides and even presently. There are people in parts of the world who are suffering but won’t agree and thereby the health deteriorates gradually leading to different outcomes, one of that is Death. There are many clinics but some people are too shy to share their problems with strangers so different online questionnaires are present online mainly being the DASS Questionnaire from which the data being used is taken as through these types of the forms people can know their current mental situation and take precautions as per needed. The Questionnaire contains 42 questions regarding different feelings and some more questions are asked as education, gender, age and some more. The sample question is

Firstly, different visualizations and methods are used for getting to know the data properly and make the data ready for model creation appropriately and different machine learning methods are used for prediction of the conditions of DAS. Only 5 Machine Learning Methods were applied to the dataset and later hyper-parameters were tuned using methods as Randomized Search or Grid Search for better.

2 Dataset Collection and Description

The Data was collected from an online version of Depression Anxiety Stress Scales [http://www2.psy.unsw.edu.au/dass/]. In total there were nearly 42k instances recorded and the model was created based on those.

The survey was open to anyone and people were motivated to take it to get personalized results. At the end of the test, they also were given the option to complete a short research survey. This dataset comes from those who agreed to complete the research survey and answered yes to the question "Have you given accurate answers and may they be used for research?" at the end.

Tables are of the form:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Q3 | Q5 | Q10 | Q13 | Q16 | Q17 | Q21 | Q24 | TOTAL  COUNT | CONDITION |
| 1 | 3 | 0 | 3 | 3 | 2 | 0 | 3 | 27 | Mild |
| 1 | 3 | 1 | 3 | 2 | 3 | 1 | 1 | 24 | Mild |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 39 | Extremely Severe |
| 1 | 2 | 2 | 0 | 1 | 2 | 0 | 0 | 16 | Normal |

* Condition of Depression, Stress was computed by the total count (count of values given to question -1)

The Ten Item Personality Inventory was administered (see Gosling, S. D., Renfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528.) e.g.:

E.g.; TIPI1 Extraverted, enthusiastic, TIPI2 Critical, quarrelsome, TIPI3 Dependable, self-disciplined.

The following values were derived from technical information:

Country- ISO country code of where the user connected from

screen-size:

1 - device with small screen (phone, etc)

2 - device with big screen (laptop, desktop, etc)

unique work location:

1 - only one survey from user's specific network in dataset,

2 - multiple surveys submitted from the network of this user

source - how the user found the test

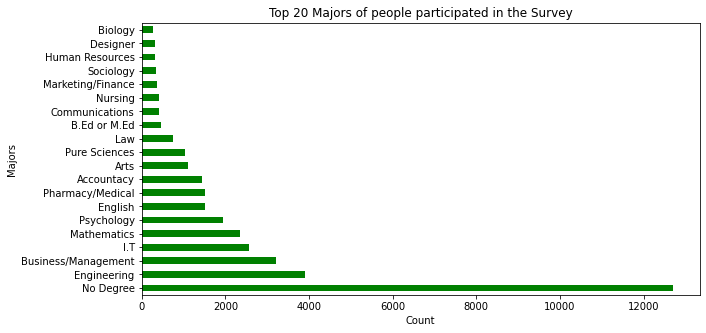
1 - from the front page of the site hosting the survey

2 - from google

0 - other or unknown

2. Methodology

The project was divided into two parts where first part was analysis of the dataset and clearing all the unwanted and making the data set perfect for the modelling and plotting some useful plots for future purposes.



The second part was making a model for prediction using various machine learning methods and thereby evaluating their performances.

3 MODEL CREATION

3.1 Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The Ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect. The term ‘machine learning’ is often, incorrectly, interchanged with Artificial Intelligence, but machine learning is actually a sub field/type of AI. Machine learning is also often referred to as predictive analytics, or predictive modelling. Coined by American computer scientist Arthur Samuel in 1959, the term ‘machine learning’ is defined as a “computer’s ability to learn without being explicitly programmed”.The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results. It’s a science that’s not new – but one that has gained fresh momentum. There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement.

DIFFERENT LEARNING METHODS

3.1.1. Supervised Learning

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance. Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

3.1.2 Semi-Supervised Learning

Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this combination, machine learning algorithms can learn to label unlabelled data.

3.1.3. Unsupervised Learning

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analysing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organise that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organised. As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined. Under the umbrella of unsupervised learning, fall: Clustering, Dimension-Reduction

3.1.4. Reinforcement Learning

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

Different Machine Learning Algorithms

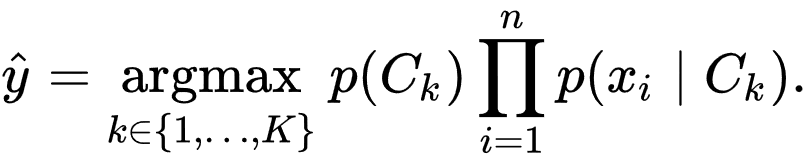
3.1.2.1 Random Forest

Random forests or ‘random decision forests’ is an ensemble learning method, combining multiple algorithms to generate better results for classification, regression and other tasks. Each individual classifier is weak, but when combined with others, can produce excellent results. The algorithm starts with a ‘decision tree’ (a tree-like graph or model of decisions) and an input is entered at the top. It then travels down the tree, with data being segmented into smaller and smaller sets, based on specific variables.



Figure Random Forest view

3.1.2.2 Naïve Bayes

Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes Theorem with naïve independence assumptions between the features. Using Bayes theorem the conditional probability can be decomposed as:

3.1.2.3 Support Vector Machines

Supervised Models with associated learning algorithms that analyze data for classification and regression analysis. SVM maps training examples to points in space so as to maximize the width of the gap between the two categories. New Examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin.

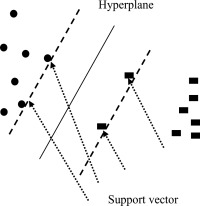


Figure plane for svm

3.1.2.4 Artificial Neural Network

An artificial neural network (ANN) comprises ‘units’ arranged in a series of layers, each of which connects to layers on either side. ANNs are inspired by biological systems, such as the brain, and how they process information. ANNs are essentially a large number of interconnected processing elements, working in unison to solve specific problems.  
  
ANNs also learn by example and through experience, and they are extremely useful for modelling non-linear relationships in high-dimensional data or where the relationship amongst the input variables is difficult to understand.

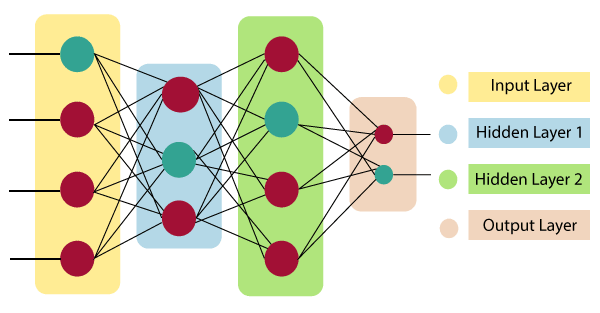


Figure ANN Layers

4 Hyper-Parameter Optimization

Machine Learning involves predicting and classifying data and to do so, we employ various machine learning models according to datasets. Machine learning models are parametrized so that their behaviour can be tuned for a given problem. These models can have parameters and finding the best combination of parameters can be treated as a search problem. A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data whereas a hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data. These are often used in processes to help estimate model parameters. There are many sets of method for finding best hyperparameters such as Randomized Search (trains and evaluates a series of models by taking random draws from predetermined set of hyperparameter distribution) or Grid Search (uses different combination of all the specified hyperparameters and their values and calculates the performance for each combination and select the best value).

5 Model Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. There are different methods for evaluating a model’s performance.

5.1 Cross Validation

Sometimes called rotation estimation or out of sample testing, is any of various similar model validation techniques for assessing how the results of statistical analysis will generalize to an independent data set. Cross Validation is resampling method that uses different portions of data to test and train a model on different iterations and is mainly used in settings where the goa is prediction, and want to estimate how accurately a predictive model would perform. Types can be as exhaustive (learns and test on all possible ways to divide the original sample into a training and a validation set) and non-exhaustive (Don’t compute all ways of splitting the original sample) cross validation

5.2 Confusion Matrix

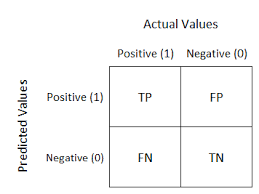
Confusion matrix may be referred as Error matrix. It has a high level of interpretability. It comprises a singular tabular format, which is often generated and visualized as heatmap. A perfect confusion matrix will have values only along the main diagonal. It not only shows us where the machine learning model faltered but also how it reached its conclusion.

Figure Confusion Matrix

5.3 Precision and Recall

Precision refers to the proportion of all observations that have been predicted to belong to the positive class and are actually positive.

Precision - Where TP- True Positive and FP- False positive

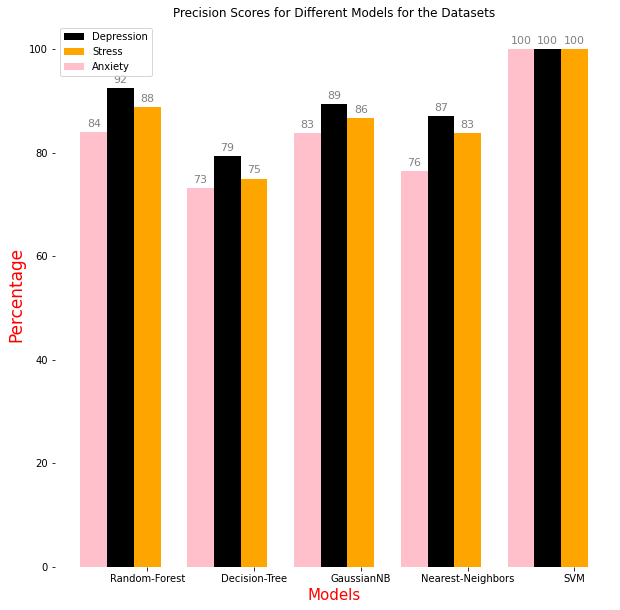


Figure Precision Scores

Recall is the proportion of observation predicted to belong to the positive class, that truly belongs to positive class. It indirectly tells us the model’s ability to randomly identify an observation that belong to positive class. Recall - Where FN- False Negative

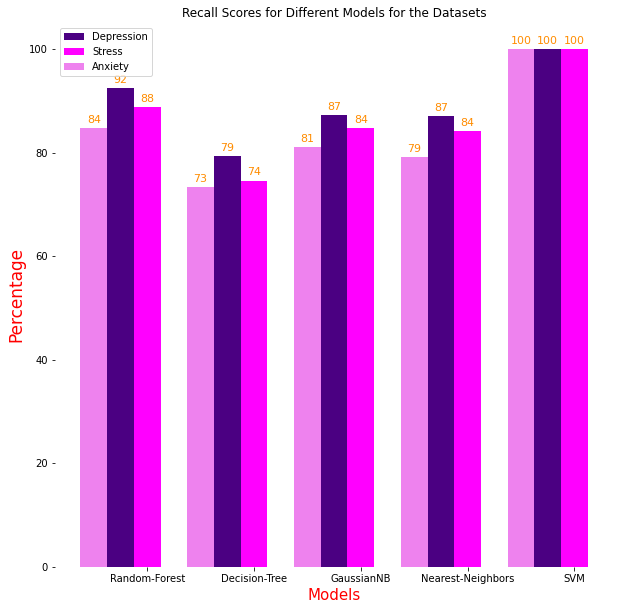


Figure Recall Scores

5.4 Area under ROC Curve

It is performance measurement for classification problems at various threshold settings. It tells how much a model is capable of distinguishing between classes. The higher the AUC better the model is at predicting when a 0 is actually a 0 and 1 is actually a 1.

6 PLOTS

6.1 Count for people suffering from DASS

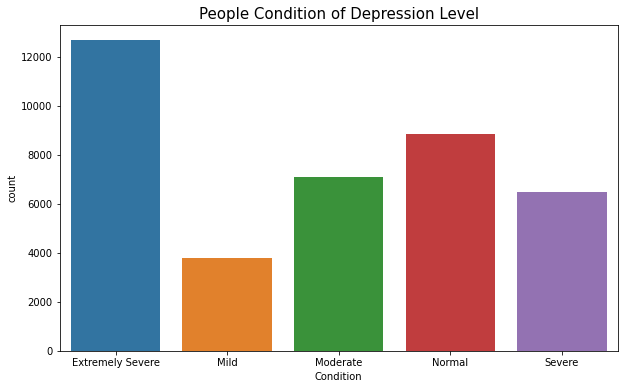
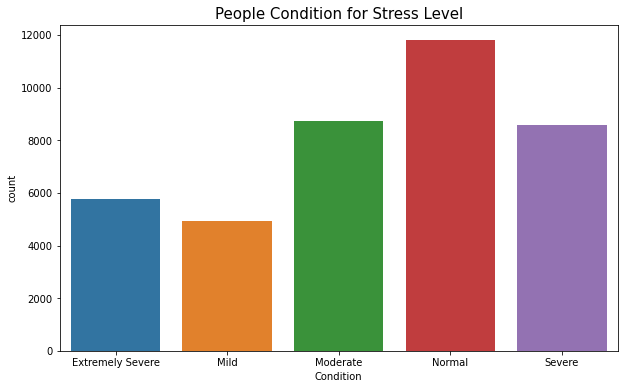


Figure Depression Condition for Participated People

Figure Stress Conditions for Participated People

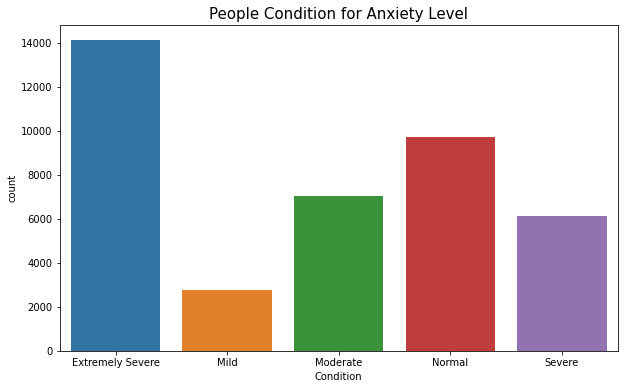


Figure Anxiety Condition for Participated People

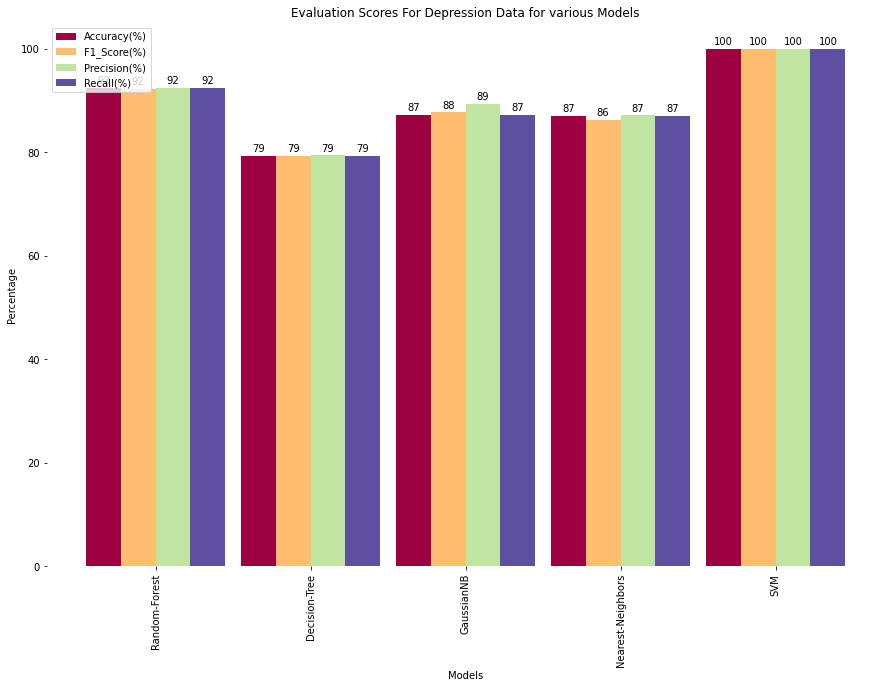
6.2 Evaluation Scores

Figure Scores for Depression table.

Figure Scores for Stress table

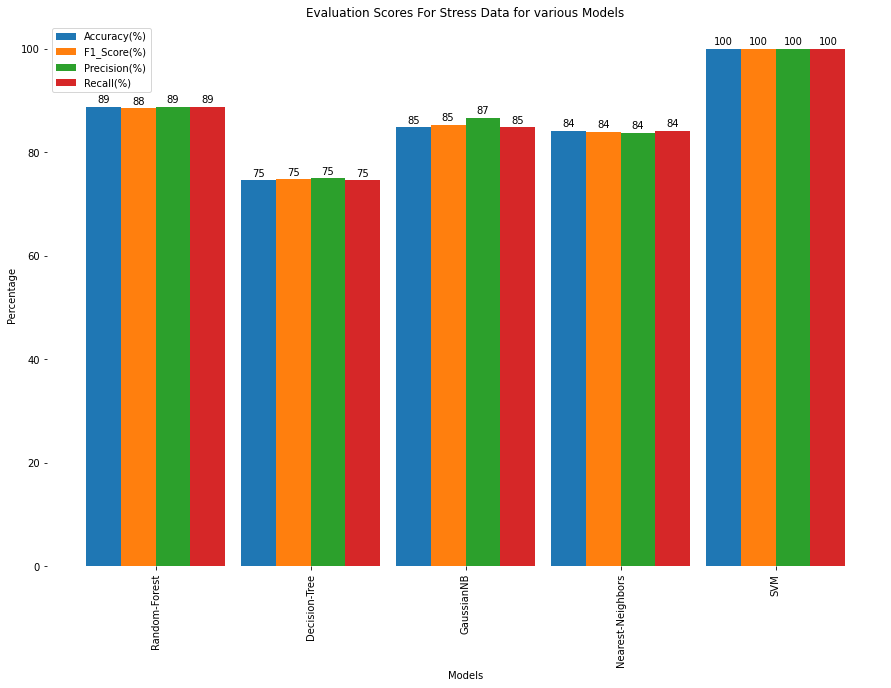


Figure Stress Scores for Models

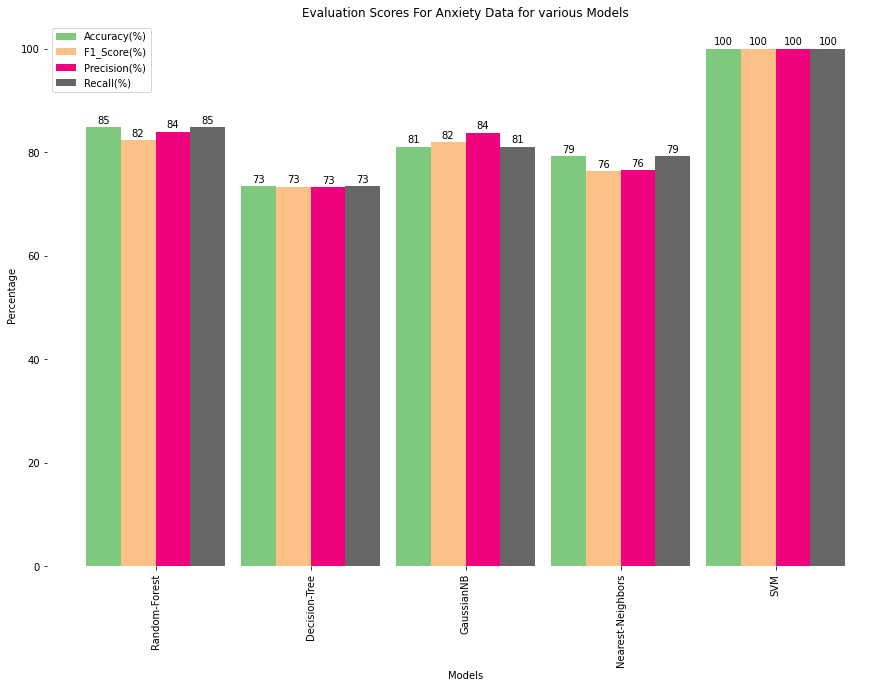


Figure Anxiety Scores for Models

7. CONCLUSION

In this dataset lots of information was taken to predict the condition of Depression, Anxiety and Stress for various people worldwide using various Machine learning methods that are Naïve Bayes, SVM, Random Forest, Decision Tree and Nearest Neighbours. Parameters were tuned using Randomized search or grid search for better selections. The dataset was divided into two parts one for the training and the other part for testing the model with SVM performing the best among all.

8. REFERENCES

DASS Dataset-> <https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses>

<https://www.kaggle.com/code/teju4405/das-prediction>

<https://github.com/Tej752/WorkProjects/blob/master/Book1.ipynb>

<das-prediction.ipynb>

<codebook.txt>

DASS Questions-> <https://maic.qld.gov.au/wp-content/uploads/2016/07/DASS-21.pdf>

DASS Scales-> <http://www2.psy.unsw.edu.au/dass/>

Machine Learning-> <https://en.wikipedia.org/wiki/Machine_learning>

<https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.91.045002>

Papers-> <https://www.sciencedirect.com/science/article/pii/S1877050920309091>

<https://www.sciencedirect.com/science/article/pii/S0191886905001765?casa_token=H_f7VD4Lp8QAAAAA:EbxVVNzb2RCmZ6Dh35208DSlOkhLjSMpI8vHSUvdSB_8zWmyu5x8Z6SXZiMXpw1EDrVlGOi6kg>

Hyper-parameter Optimization-> <https://proceedings.neurips.cc/paper/2011/hash/86e8f7ab32cfd12577bc2619bc635690-Abstract.html>

<https://library.oapen.org/bitstream/handle/20.500.12657/23012/1007149.pdf?sequence=1>

<https://towardsdatascience.com/hyperparameters-optimization-526348bb8e2d>

Reinforcement Learning-> <https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/>

Different Machine learning methods-> <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>

<https://scikit-learn.org/stable/modules/naive_bayes.html>

<https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code>

<https://en.wikipedia.org/wiki/Support-vector_machine>

<https://en.wikipedia.org/wiki/Random_forest>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<https://ieeexplore.ieee.org/abstract/document/7724478>

<https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf>

Evaluation-> <https://link.springer.com/article/10.3758/BF03213026>

<https://en.wikipedia.org/wiki/Confusion_matrix>

<https://en.wikipedia.org/wiki/Cross-validation_(statistics)>

<https://machinelearningmastery.com/k-fold-cross-validation/>

<https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f>

<https://en.wikipedia.org/wiki/Precision_and_recall>

Analysing-> <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1430>

<https://www.sciencedirect.com/science/article/pii/S0957417422004249>

[< https://openpsychometrics.org/\_rawdata/validity](%3c%20https://openpsychometrics.org/_rawdata/validity)>