Prediction of Depression, Stress and Anxiety

Tejveer Singh*

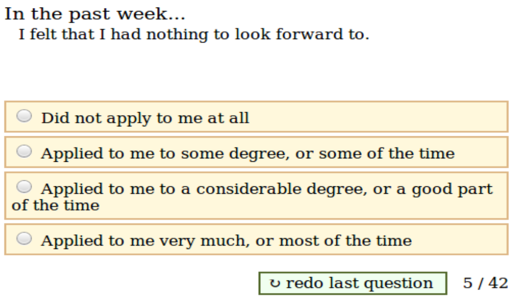
M.Sc. Student, Department of Mathematics and Computing, Indian Institute of Technology Dhanbad, Dhanbad, India

***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.**

***Keywords*: Depression, Stress, Anxiety.**

1. 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,



In the past week...
I felt that I had nothing to look forward to.

○ Did not apply to me at all

○ Applied to me to some degree, or some of the time

○ Applied to me to a considerable degree, or a good part of the time

○ Applied to me very much, or most of the time

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.

A. 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 people all over the world and were highly motivated towards taking this test. At the end they were even given a chance to do small research on the survey concluded. This dataset comes from those people who chose to complete the research. The dataset had nearly 170 attributes with 42 among them being the question regarding the conditions faced whereas some were about different personality moods and rest were some personal data.

The data was divided into three different datasets with each being Depression, Stress and Anxiety. Below is an example of how the table looked like.

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)

2. Methodology

The project is 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.

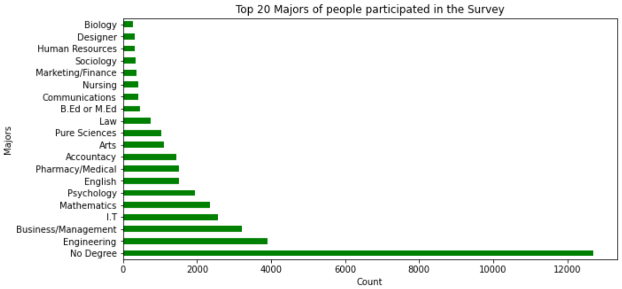


Fig. 1. Survey

The second part was making a model for prediction using

[*] Corresponding author: tejveers150@gmail.com

various machine learning methods and thereby evaluating their performances.

3. Model Creation

A. Machine Learning

Machine learning is field of study that gives computers to learn without being programmed. T gives the computer that makes it more similar to humans- The Ability to learn. Use of Machine learning has increased widely among different sectors. It has helped in making accurate futuristic decisions, predicting disease, stock markets and even more.

The iterative aspect of machine learning is important as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce accurate results. It has gained fresh momentum from precious decade. ML can be further divided into four different algorithms: Supervised, Semi-Supervised, Un-Supervised, Reinforcement.

1) Classification

Classification in machine learning is an example of supervised learning. It is the task of learning a target function f that maps each attribute set x to one of the predefined class labels y. Classification can be further divided into Descriptive i.e., where classification model serves as explanatory tool to distinguish between objects of different classes and Predictive modelling i.e., where classification model can also be used to predict the class label of unknown records. The work done in the dataset uses classification for modelling the data.

2) Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is collection of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters. In machine learning, clustering is example of unsupervised learning. Unlike classification, clustering doesn’t rely on predefined classes and class labelled training examples. For this reason, clustering is a form of learning by observation.

Different Machine Learning Algorithms:

a) Random Forest

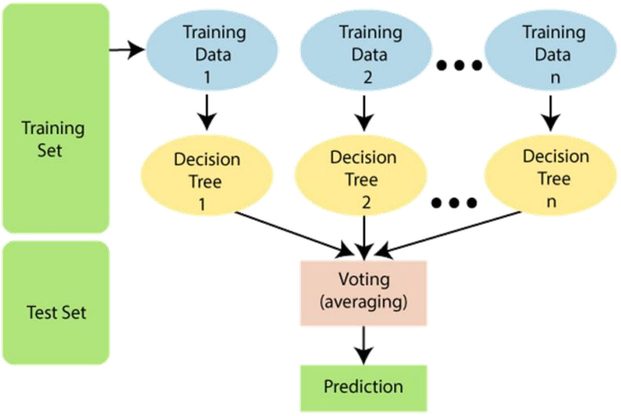


Fig. 2. Random Forest view

Random forests or ‘random decision forests’ is an ensemble learning method, that constructs a set of classifiers from the training data. It performs classification by taking vote on the predictions made by each base classifier. The improvement in classification accuracy is done by aggregating the predictions of multiple classifiers. Bootstrap sample is chosen to train decision trees as at every internal node randomly n attributes are selected for splitting.

b) Naïve Bayes

Naive Bayes classifiers are statistical classifiers based on Bayes theorem.

$$P(C|x) = \frac{p(C).p(x|C)}{p(x)}$$

A dataset with n attributes and m classes is chosen where x is a test tuple, prediction is made for a class if and only if $P(C(i)|X) > P(C(j)|X)$ i.e., to maximize P(C|X).

c) Support Vector Machines

SVM works very well with high dimensional data and as well avoids the dimensionality problem. It represents the decision boundary using subset of training examples that are known as Support vectors. The main goal is to find optimal separating hyperplane that maximizes the margin of training data.

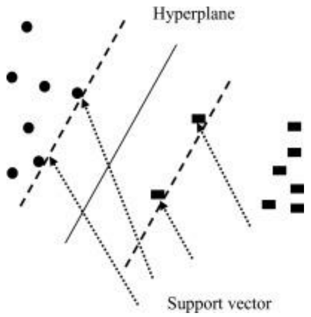


Fig. 3. Hyperplane for SVM

d) Decision Trees

Decision Tree is a flow chart like tree structure where the internal nodes represent the test on the attribute, the branch represents an outcome of the test and the leaf nodes represents the class labels. To find the best decision tree is a NP-hard problem following greedy search approach i.e., splitting based on attribute that optimizes certain criterions. Many different approaches are present to determine the split that are Entropy, Gini Index or Classification Error.

$$\text{Entropy} = -\sum p\left(\frac{i}{t}\right) logp(\frac{i}{t})$$

$$\text{Gini Index} = 1-\sum(p(\frac{i}{t}))^2$$

4. Hyper-Parameter Optimization

Machine learning involves predicting and classifying the data. Models are parametrized for better outputs of the given problem. These models may have different parameters and finding a perfect combination is treated as a search problem. There are different techniques to find the best pair of parameters for a particular model to perform the best e.g., Randomized Search CV or Grid Search CV.

5. Model Evaluation

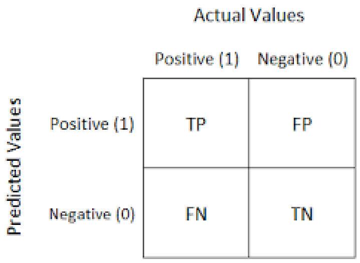
Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. There are different methods for evaluating a model's performance.

A. 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

B. Confusion Matrix

Confusion matrix can be referred as Error matrix. 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.



Predicted Values	Actual Values Positive (1)	Actual Values Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig. 4. Confusion Matrix

A Confusion matrix looks like:

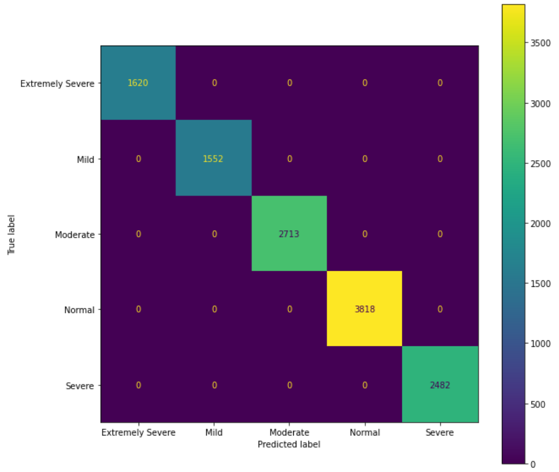


Fig. 5.

C. Precision and Recall

Precision is the proportion of observations that have been predicted to belong to the positive class and which being are actually positive.

Precision - $\frac{TP}{TP+FP}$

Where,

TP - True Positive
FP - False positive

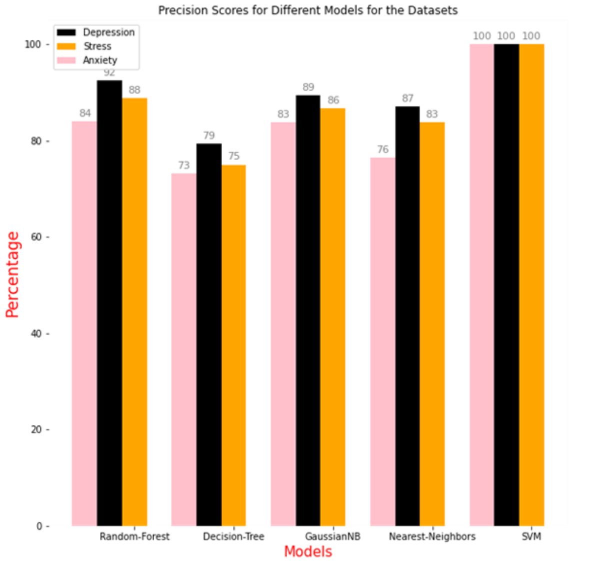


Fig. 6. Precision scores

Whereas 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 - $\frac{TP}{TP+FN}$, Where FN- False Negative

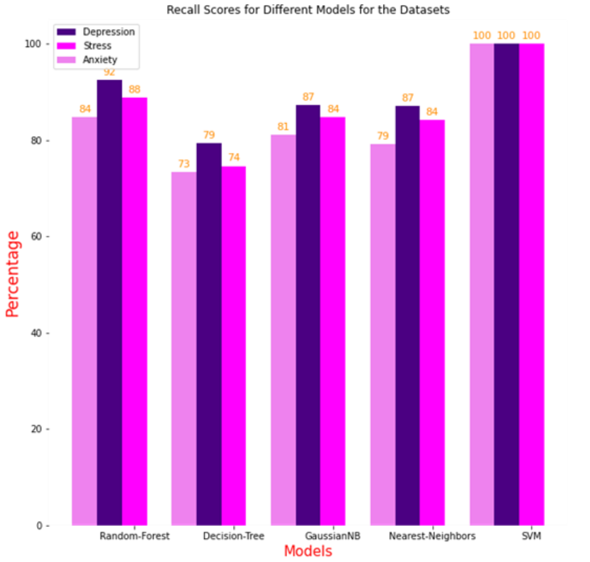


Fig. 7. Recall scores

D. 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

A. Count for people suffering from DASS

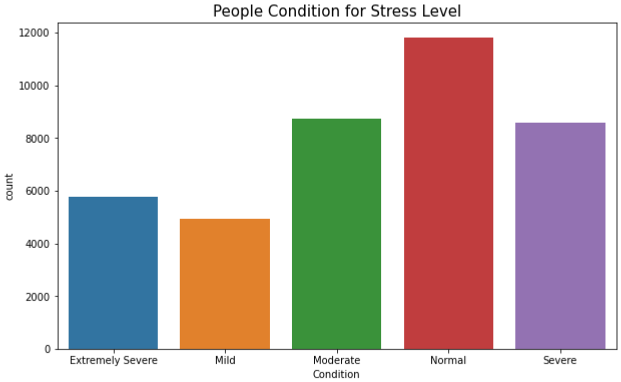


Fig. 8. Stress conditions for participated people

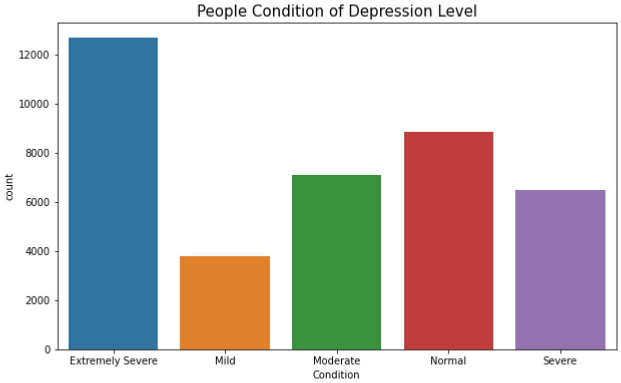


Fig. 9. Depression condition for participated people

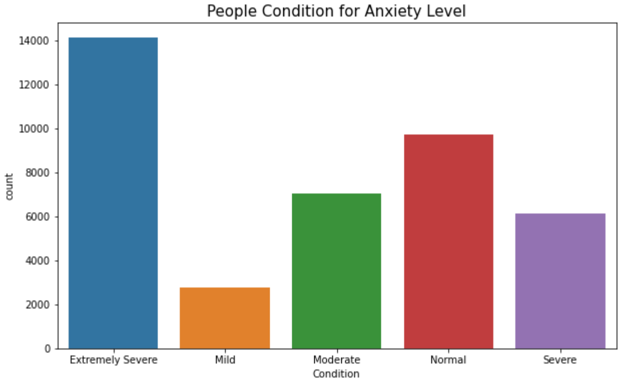


Fig. 10. Anxiety condition for participated people

B. Evaluation Scores

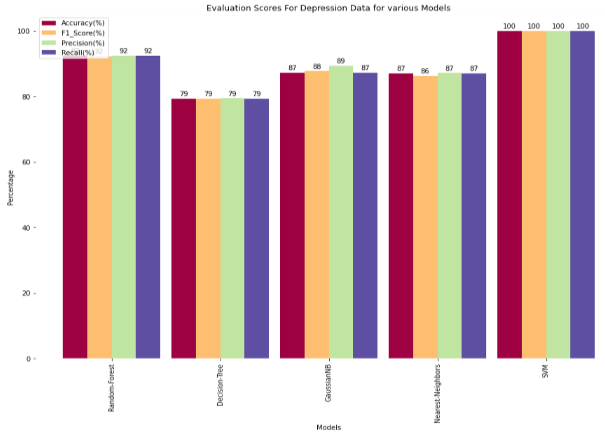


Fig. 11. Scores for depression table

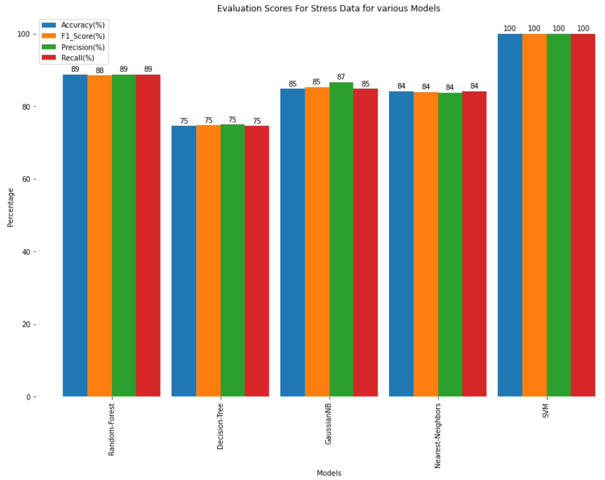


Fig. 12. Stress scores for models

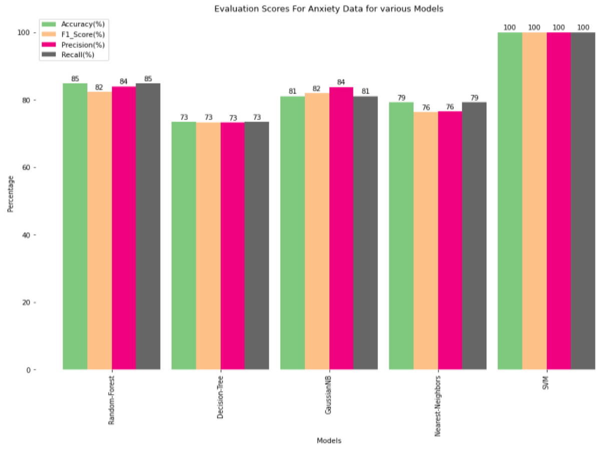


Fig. 13. Anxiety scores for models

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.

References

[1] DASS Dataset https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses

[2] https://www.kaggle.com/code/teju4405/das-prediction

[3] https://github.com/Tej752/WorkProjects/blob/master/Book1.ipynb

[4] das-prediction.ipynb

[5] codebook.txt

[6] DASS Questions https://maic.qld.gov.au/wp-content/uploads/2016/07/DASS-21.pdf

[7] DASS Scales, http://www2.psy.unsw.edu.au/dass/

[8] Machine Learning, https://en.wikipedia.org/wiki/Machine_learning

[9] Giuseppe Carleo et al., Machine learning and the physical sciences, Rev. Mod. Phys. 91, 045002.

[10] Anu Priya, Shruti Garg, Neha Prerna Tigga, “Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms,” Procedia Computer Science, vol. 167, pp. 1258-1267, 2020.

[11] Ryan C. Martin, Eric R. Dahlen, “Cognitive emotion regulation in the prediction of depression, anxiety, stress, and anger, Personality and Individual Differences, Volume 39, Issue 7, 2005, Pages 1249-1260.

[12] James Bergstra et al., Algorithms for Hyper-Parameter Optimization, Advances in Neural Information Processing Systems 24 (NIPS 2011).

[13] https://library.oapen.org/bitstream/handle/20.500.12657/23012/1007149.pdf?sequence=1

[14] https://towardsdatascience.com/hyperparameters-optimization-526348bb8e2d

[15] https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/

[16] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[17] https://scikit-learn.org/stable/modules/naive_bayes.html

[18] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code

[19] https://en.wikipedia.org/wiki/Support-vector_machine

[20] https://en.wikipedia.org/wiki/Random_forest

[21] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[22] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310-1315.

[23] Vladimir Nasteski, An overview of the supervised machine learning methods, 2017.

[24] Townsend, J.T. Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics 9, 40–50 (1971).

[25] https://en.wikipedia.org/wiki/Confusion_matrix

[26] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[27] https://machinelearningmastery.com/k-fold-cross-validation/

[28] https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

[29] https://en.wikipedia.org/wiki/Precision_and_recall

[30] Moskovitch, R. (2022). Multivariate temporal data analysis - a review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(1), e1430.

[31] Fatemeh Behrad, Mohammad Saniee Abadeh, An overview of deep learning methods for multimodal medical data mining, Expert Systems with Applications, Volume 200, 2022

[32] https://openpsychometrics.org/_rawdata/validity