Machine Learning Techniques for Stress Prediction in Working Employees

U SRINIVASULU REDDY
Assistant Professor,
Machine Learning and Data Analytics Lab,
Department of Computer Applications,
National Institute of Technology, Trichy
usreddy@nitt.edu

ADITYA VIVEK THOTA
B.Tech., ECE,
National Institute of Technology, Trichy
adityavivek1998@gmail.com

A DHARUN
B.Tech., ECE,
National Institute of Technology, Trichy
dharun.anand97@gmail.com

***Abstract*- Stress disorders are a common issue among working IT professionals in the industry today. With changing lifestyle and work cultures, there is an increase in the risk of stress among the employees. Though many industries and corporates provide mental health related schemes and try to ease the workplace atmosphere, the issue is far from control. In this paper, we would like to apply machine learning techniques to analyze stress patterns in working adults and to narrow down the factors that strongly determine the stress levels. Towards this, data from the OSMI mental health survey 2017 responses of working professionals within the tech-industry was considered. 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. Boosting had the highest accuracy among the models implemented. 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.**

Keywords*- *Stress prediction, Boosting, Bagging, Decision Trees, Healthcare, Machine Learning

I. INTRODUCTION

Stress-related mental health disorders are not uncommon among the working class. Several studies in the past have raised concerns over the same. According to a study by the industry association, Assocham, more than forty-two percentage of working professionals in the Indian private sector suffer from depression or general anxiety disorder due to long work hours and tight deadlines. This portion of individuals is rising as stated in the 2018 Economic Times article based on the survey conducted by Optum[1] that half of the working professionals in India suffer from stress. The survey considered the responses of as many as 8 lakh employees from over 70 major companies, each with a workforce of 4,500 or above. Maintaining a stress-free workplace must be given a prime importance for greater productivity and well-being of the employees. Several steps can be taken to help working professionals cope up with stress for mental well-being like counselling assistance, career guidance, stress management sessions, and health awareness programs. Early identification of employees who will be needing such a help will improve the chances of such measures being successful.

We hope to ease this process by using machine learning methods to develop a model to predict the risk of stress experienced and if treatment is required by an individual by taking some of his/her professional and personal factors as parameters collected in the form of carefully drafted surveys. Such an approach will not only help HR managers to understand better about their employees but also help in taking preventive measures to decrease the chance of an employee leaving the company or underperforming. We can also perform early prediction if a person requires treatment for his mental health or not.

II. DATA DESCRIPTION

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.

As mentioned, OSMI Mental Health in Tech 2017 survey [2] was taken as the dataset, using which 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.

978-1-5386-1508-9/18/$31.00 ©2018 IEEE

The OSMI Mental Health in Tech 2017 survey containing 750 responses from various employees working in a wide range of tech divisions was used. These responses include both professional and personal factors of the individual and hence will give a complete view of the environment faced by professionals.

Data Cleaning: The original dataset contains 750 responses from different individuals and 68 attributes spanning both their personal and work life. The data has been cleaned using various standard methods that check for data consistency and validity of the survey responses [3]. In order to have a specific model some of these attributes have been neglected and finally 14 out of these above parameters were taken into consideration based on their relevance to our research. Whether an employee has taken treatment for stress-related disorders in past or not is used as a reference to be predicted by our trained models.

Furthermore, "one hot encoding" method is employed for certain responses in order to facilitate multiple fields that require individual parameters.

All the textual responses were given numerical weights according to their significance. 'Yes' is taken to be 1, 'no' to be 0 and a 'maybe' to be 0.5. All 'NaN' (not a number) cells were replaced with 0. The categorical data was converted into numeric using label encoder.

70 percent of the responses were used for training the model while the remaining 30 percent was utilized for testing.

III. MACHINE LEARNING TECHNIQUES USED

Machine learning is a subset of Artificial Intelligence that provides computers and computing systems the ability to independently learn and improve from previous experience without being explicitly programmed by a human. Machine learning is based on the development of computing programs that can retrieve data and learn for themselves. This is highly effective in healthcare as there is enormous amount of data and if this is properly fed to an intelligent system and trained accordingly, the resulting prediction model will be unparalleled and free from human errors and reduce the time required for diagnostics. Hence, the responses of the OSMI 2017 dataset were used to train the following ML models that are previously tested in healthcare based classification problems [4][5].

A. *Logistic Regression:* Like all regression methods, the logistic regression is a predictive analysis. It is used in scenarios where one binary variable is dependent on one or more independent variables. Here, we take the 14 relevant attributes to be independent variables and the possibility of an employee having stress and needing treatment as the dependent variable which is to be predicted by the trained model.

B. *KNN Classifier:* K-Nearest Neighbor (KNN) classifier is a supervised learning algorithm that can be implemented on labelled data. It was used here for predicting if a person needs treatment or not. KNN classifies the dependent variable based on how similar it's independent variables are to a similar instance from the already know data.

C. *Decision Trees:* A decision tree can be used to model multiple choices or if-else statements/decisions in a tree-like fashion. Here, decision trees are used to find the most contributing factors among the 14 features that are used. This is highly helpful, as now more attention can be given to these areas and necessary steps are taken on those lines.

D. *Random Forest Classifier:* Random Forests are a cluster of decision trees working together with each other and it has proved to be more effective than a single decision tree. Random Forest is a flexible, easy to use ML algorithm that produces a good result persistently, even without hyper tuning.

E. *Boosting:* We also implemented ensembled methods that augment the performance of existing models. Boosting is a highly effective and commonly used ensemble classifier. The main motive for boosting is to reduce bias in the model.

F. *Bagging*- Another effective ensemble method is, Bootstrap Aggregating (or bagging). It involves training of a model using the same algorithm but on different subsets of data from the dataset. This not only helps in improving the stability and accuracy of the model but also in reducing the variance of the model.

IV. PERFORMANCE METRICS

Different parameters were considered to evaluate the accuracy of our trained model as follows:

A. *Classification Accuracy:* It is a measure of the effectiveness of the classification model. Classification accuracy is the percentage of successful predictions out of the total predictions made. It can be used as a rank of performance among different models.

B. *False Positive Rate:* It is essentially a measure of how many instances the model classified a negative event as positive. In our scenario, it would mean the model classifying a person as needing help to

cope up with stress and mental health although the person is perfectly alright. Lower the false positive rate, better the model.

C. *Precision:* Precision in data mining is a measure of the fraction of positive predictions that are actually positive, that is, given an employee is stressed and needs treatment, the model classifies the same. Higher precision implies the model is good in identifying the people in need of help.

D. *AUC Score:* AUC stands for Area Under the Curve. It is used as an analysis criterion to measure the performance of different models to determine which model predicts the classes best. Receiver Operating Characteristics (ROCs) are a part of AUC and it is a plot of the true positive rate with the false positive rate. Area under a ROC curve determines how well the model can predict a positive classification correctly.

E. *Cross-validated AUC:* A major aspect of forming a prediction model is to ensure its stability. In order to increase the same, we go for cross-validation, in which we use a part of the training samples as test data and then fit our model.

V. RESULT AND PERFORMANCE ANALYSIS OF DIFFERENT METHODS

All the discussed models were implemented in Python using Scikit-learn [6] to test the prediction if a person needs treatment or not. The results are visualized and tabulated as follows:

From Figure 1, it can be inferred that all the trained models performed fairly well in classification with Boosting achieving highest accuracy of 75.13 and bagging achieved the least accuracy of 69.43.

Figure 2 depicts the feature importance of the 14 attributes considered, obtained by using a decision tree. The graph shows that gender has the highest influence on stress and mental health among the selected parameters. This can be reflected on the fact that women are generally found to be at greater mental stress than men. This supports the findings of a 2010 study published by American Psychological Association on the correlation between gender and stress [7].

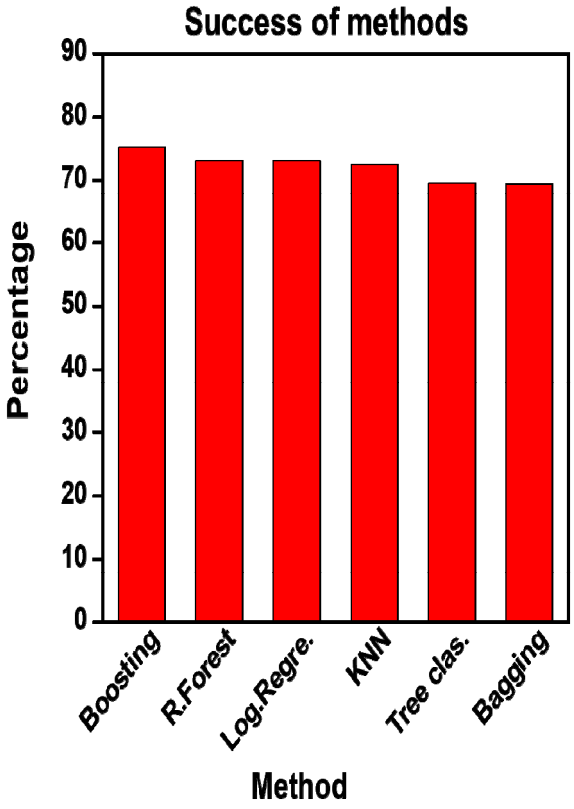

Figure 1: Percentage of accuracy for each method.

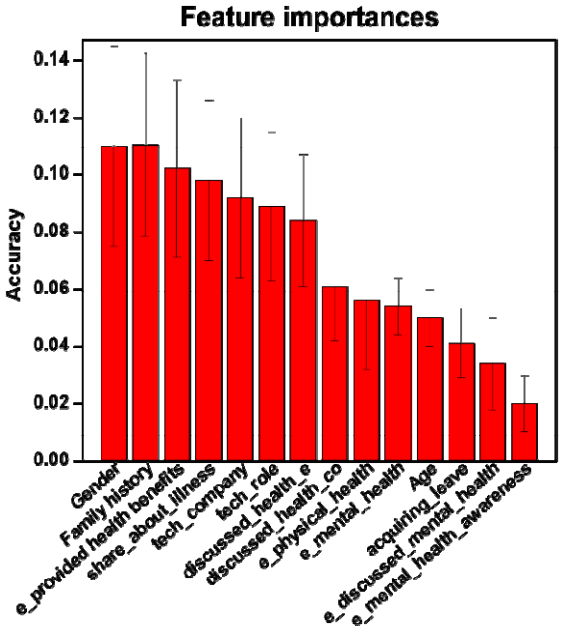

Figure 2: Feature importance of various attributes considered.

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

From the graph, it can also be observed that the people working in a tech company were slightly more at the risk of developing stress and mental health problems even if their role was not tech-based. The performance of various trained models was evaluated and tabulated in Table 1. The model was trained using various machine learning techniques and out of these, boosting performed better than the other models in terms of accuracy, false positive rate, and precision. However, in terms of cross-validated AUC score, random forest classifier scored higher indicating that this model is more stable. It can also be noted that both logistic regression and random forest classifier scored the same classification accuracy, but random forest outperforms the former in other parameters.

KNN classifier has the highest false positive rate indicating that it is highly unreliable to be used in the given scenario.

VI. CONCLUSION

Gender, family history of illness, and whether an employer gives mental health benefits to their employees had more importance than other parameters in determining whether a person can develop mental health issues.

From our study, it was found that people working in a tech company were slightly more at the risk of developing stress even if their role was not tech-based. These insights can be effectively used by corporates to frame better HR policies for their employees.

Also, ensemble methods like boosting produced the highest classification accuracy and precision followed by random forest. A 75.13% accuracy signifies that the application of Machine Learning methods for prediction of stress and mental health condition gives significant results and can be explored further, meeting the objective of this paper.

VII. FUTURE SCOPE OF WORK

Additional methods like the Naive Bayes classifier can be used to test the efficiency of the model. One can implement deep learning techniques like CNN (Convoluted Neural Networks) and verify how the model performs for the given dataset.

A much more specific and vast dataset can be used as a training model since the number of responses is limited in our case.

We can also customize the survey taken in order to procure responses in the right format and to increase the number of attributes as per relevance. Questionnaires from established institutions and organizations such as the World Health Organization relating to stress and mental health can be considered for conducting a survey. There is also a scope for formulating a scoring system for different attributes based on their importance to create a uniform scale to measure the stress levels of an individual.

VIII. ACKNOWLEDGMENT

The authors would like to acknowledge the Machine Learning and Data Analytics Lab, Department of Computer Applications, NIT Trichy for providing the infrastructure support.

REFERENCES

[1] Bhattacharyya, R., & Basu, S. (2018). India Inc looks to deal with rising stress in employees. Retrieved from 'The Economic Times'

[2] OSMI Mental Health in Tech Survey Dataset, 2017 from Kaggle

[3] Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS medicine*, *2*(10), e267.

[4] Shwetha, S, Sahil, A, Anant Kumar J, (2017) Predictive analysis using classification techniques in healthcare domain, *International Journal of Linguistics & Computing Research,* ISSN: 2456-8848, Vol. I, Issue. I, June-2017

[5] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[7] Gender and Stress. (n.d.). Retrieved from *APA* press release 2010