2020 IEEE INTERNATIONAL CONFERENCE ON
ADVANCES AND DEVELOPMENTS IN ELECTRICAL AND ELECTRONICS ENGINEERING (ICADEE 2020)

Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies

Rahul Katarya
Department of computer science
Delhi Technological University (DTU)

New Delhi, India
rahuldtu@gmail.com

Saurav Maan
Department of computer science
Delhi Technological University (DTU)

New Delhi, India
sauravmaan_2k19swe14@dtu.ac.in

2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE) | 978-1-7281-9251-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICADEE51157.2020.9368923

***Abstract*— mental health has always been an important and challenging issue, especially in the case of working Professionals. The modernized (hectic) lifestyle and workload take a toll over people over time making them more prone to mental disorders like mood disorder and anxiety disorder. Thus, the risk mental health problems increase in working professionals. To deal with this problem industries provide mental health care incentives to their employees, but it is not enough to deal with the problem. In this paper, we utilize the data from mental health survey 2019 that contains the data of working professionals for both tech and non-tech company employees. 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.**

Keywords— SVM, KNN, Regression, Decision tree, Random forest.

I. INTRODUCTION

In the modernized world, working professionals are under lots of pressure for reasons like peer pressure, short deadlines, competition. All these things contribute to building up mental stress, which slowly leads to mental health disorders. In the US 18% of the working population which is 40 million people suffers from mental health disorder[1]. 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 case of working professionals the disruption is in form of work performance like having a hard time completing deadlines [2]. We can further characterize mood disorder in forms depression disorder, mania, hypomania and bipolar disorder. Depression disorder as the name suggests is synonym 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 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 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 [3]. 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. In this paper, we use the OSMI mental health survey 2019 as dataset and apply multiple Ml algorithms to find the attributes or features, which contribute to mental disorders or can be used to detect mental health disorders.

OSMI (Open source mental illness) is a non-profit organization whose motive is to raise awareness about mental health disorders. They teach the economic community how mental disorders influence the productivity of workers and contribute to making the workplace a safe and better place for employees. They provide guidelines to executives and HR for mental wellness in the workplace. The dataset contains more than 70 attributes, which contain both personal and professional features of the employees. In the feature selection part of data processing, we select seven attributes out of 70 attributes, which contribute most to the mental disorder or can be used to predict the mental health disorder. The feature selection is done by carefully scrutinizing done on the base of past works/papers and some research on disorders. After the feature selection comes the classification part in which different ML algorithms are applied to the selected features to predict the mental disorders and at last we apply feature importance to find the attributes which are most useful in predicting the mental health disorders.

II. RELATED WORK

In paper [4] for predicting Anxiety disorder author proposed rules based on factors like persons 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.

In [5] author used smartphone-based sensor system to monitor or find the changes in states of bipolar disorder

978-1-7281-9251-2/20/$31.00 ©2020 IEEE

patient. Also, develop an early warning system with recall and precision of about 97%.

In [6] author proposed a wearable headband for multimodal stress monitoring and detection.

In [1] the author predicted generalized anxiety disorder among women and proposed that women are more prone to GAD (generalized anxiety disorder) almost two times than men are; the author was able to predict GAD with 90% accuracy with random forest.

In [7] the author used wearable sensors for Bipolar based prediction using Heart rate variability.

In [8] the author used ML algorithms to detect stress in working employees and found features which contribute to mental stress. Random forest was found to have the highest precision and accuracy with 75.13%.

In [9] the author used HRV to detect stress with multiple ML algorithms to predict the stress.

In [10] the author used heartbeat data to identify stress and used ML algorithms for classification, data is collected at a 5-minute interval such that after 5 minutes of heartbeat data collection a relaxation interval of 5 minutes to the user.

III. MACHINE LEARNING TECHNIQUES USED

Machine learning is a subset of AI, which allows the system to learn from previous actions and improve without giving explicit commands. Machine learning can be categorized into two types based on the type of learning or the data used. In Supervised learning, we have labelled data where labelled means both input and output is known to us. Whereas in Unsupervised learning we have unlabeled data, which means we are given values of independent variables and value of dependent variables are absent. We will only use supervised learning algorithms as the dataset used is labelled.

A. Support vector machine(SVM)

SVM takes data points as input and gives the output as a hyperplane. It divides the classes using a plane(hyperplane) also known as the decision boundary[11]. Where the decision boundary must maximize the distance of the nearest element of each class. Decision boundary separates the points into different classes.

Let n be number of data points

$$(\overrightarrow{x_1}.y_1) \dots \dots \dots (\overrightarrow{x_n}.y_n) \quad (1)$$

Here x is the real vector and y represents the class (0 or 1)
The maximum margin hyperplane can be defined as

$$\vec{w}.\vec{x} - b \quad (2)$$

Where $\vec{w}$ is the normal vector and $\frac{b}{[\![\vec{w}]\!]}$ is the offset of hyperplane along $\vec{w}$.

B. Logistic Regression

It uses the sigmoid function when predicting the dependent variable using one or more dependent variables[12]. The dependent variable usually is a binary number that ranges between 0 and 1. The sigmoid function is represented as follows :

$$\frac{1}{1+e^{-value}} \quad (3)$$

The logistic regression can be represented as follows:

$$\ln \frac{p(y=1)}{1-p(y=1)} = w_0 + w_1x_1 + \dots + w_nx_n \quad (4)$$

C. K-nearest neighbours (KNN)

As it is a supervised learning algorithm we will need labelled data for training the model. When classifying a data point we look at K nearest neighbours of the data the majority will decide which class the data point will belong[13]. Where nearest distance is in form of Euclidean distance, whose formula is

$$(x, y) = \sum_{j=1}^{k} \sqrt{(x_j - y_j)^2} \quad (5)$$

After this, the input data point will be assigned to the class which have the highest probability. The probability can be represented as

$$P(y = j|X = x) = \frac{1}{k} \sum_{y \in A} I(y^i = j) \quad (6)$$

In KNN we have to be prudent when selecting the value of K. A smaller value of k usually leads to an irregular decision boundary and high value of k leads a smoother decision boundary.

D. Decision Tree

A decision tree is a supervised machine-learning algorithm so that means the data set should be labelled. In Decision tree algorithm, the classification is done based on a set of rules. In a decision tree, a node will represent a feature, the branch will represent a rule and the leaf node will represent the outcome. It can be represented in a tree-like structure which provides higher stability and accuracy[14].
In Decision tree Algorithm following steps are taken in the first step a tree will be constructed which will have input features as its nodes. In the next step, it will select a feature from the input features for predicting the output, which gives the highest information gain. Now use the above steps for the creation of subtrees by making use of the features, which are not used earlier. In the paper, we use the decision tree to find features, which contribute to the mental health disorders and the degree to which they contribute to the mental health disorder, by using feature importance.

E. Random Forest

It is an ensemble model which means it uses many machine learning algorithms to increase its performance as compared to other machine learning algorithms[15]. Random forest randomly picks up a subset from the training data set and

generate various decision trees. It will predict the class of test class objects using the decision trees.

F. Naïve Bayes

It is a probability-based classifier which applies Bayes theorem. For Naïve Bayes features should have strong independence[16].

IV. FLOW CHART

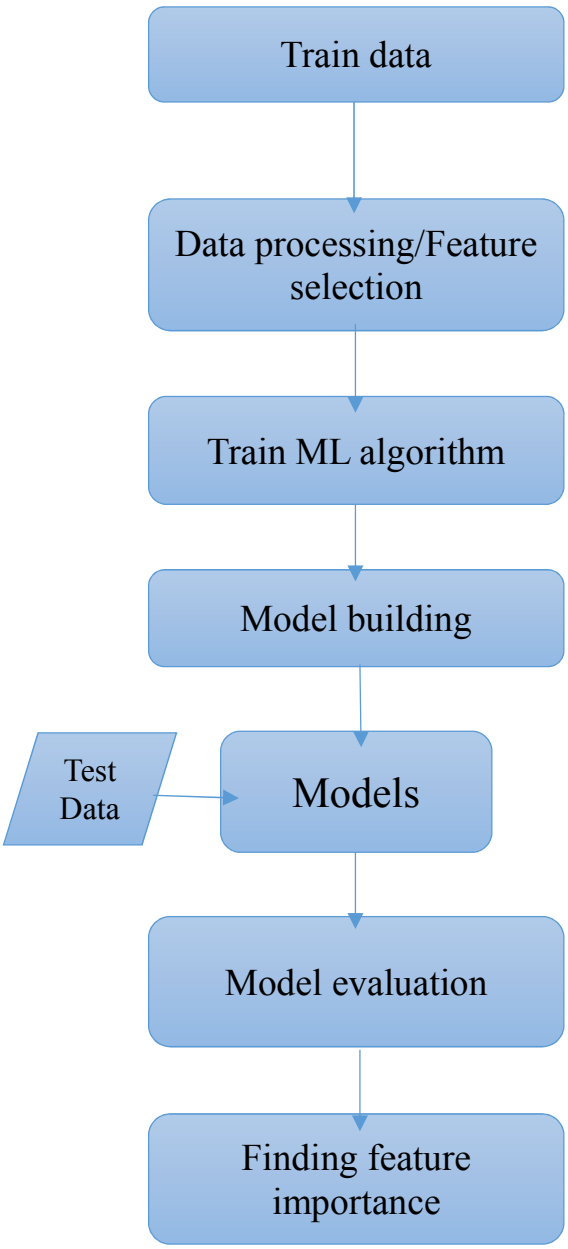
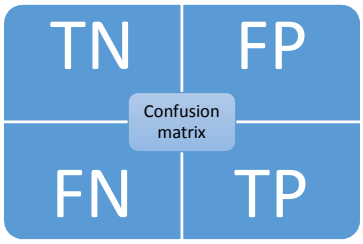


Fig. 1 Data flow chart for the framework

V. CONFUSION MATRIX

The confusion matrix can be described in the form of the form of



Where,
TP stands for True positive and TN stands for true negative. TN means the output was zero or negative and it was predicted correctly. TP means the output was positive or 1 and was predicted correctly. FP means the output is negative or 0 but it is predicted as 1 or positive. FN(False negative) means the output is positive or 1 but it is predicted as 0 or negative.
We can use these value to check the performance of a classifier by calculating precision and recall of the classifier
Where,

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Precision gives the percentage of correctly predicted outputs

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

Recall gives the percentage of positive outputs correctly predicted by the classifier

$$F1 = 2.\frac{Precision.Recall}{Precision+Recall} \quad (9)$$

F1 score considers both false positive and false negative and is the average of precision and recall.

VI. RESULT

After applying various machine-learning algorithms when predicting mental health disorder with selected attributes which are

1. Whether the company is a Tech company or not.
2. Age of the employee.
3. Gender of the employee.
4. Family history of mental health disorders if any.
5. Personal history of a mental health disorder.
6. Mental health benefits or care provided by the employer.
7. Discussing mental health status with the employer.

We got the following results

TABLE I. COMPARISON OF CLASSIFICATION MODELS

S.no	Algorithms	Accuracy (%)	Precision	Recall	F1
1.	KNN	74%	76	82	79
2.	SVM	76%	75	88	81
3,	Logistic Regression	84%	82	94	87
4.	Decision tree	84%	83	92	87
5.	Random Forest	77%	81	80	81
6.	Naïve bayes	79%	78	90	83

As shown in table 1 and fig 1 logistic regression and decision tree shows the highest accuracy when predicting mental health disorders in employees of both tech and non-tech companies using the selected 7 attributes of the employees.

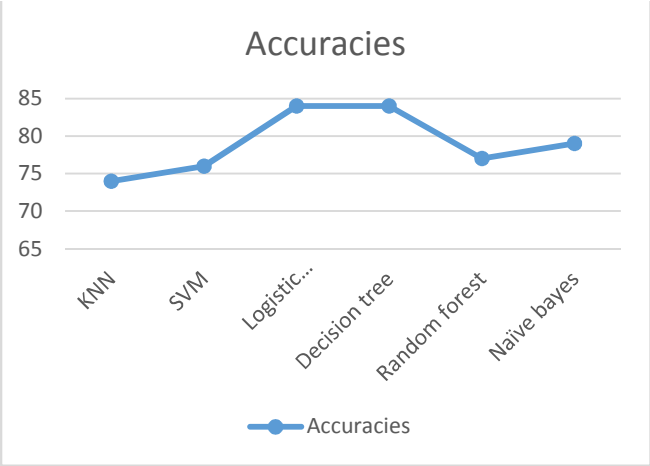


Fig.1 accuracy comparison of different classifier

In fig 2. By comparing precision and recall of logistic regression and decision tree. We found that the decision tree has shows higher precision while logistic regression shows higher recall. Precision considers false-positive while recall considers false-negative. In this case study, our main focus is on precision and recall tradeoff as we want to detect mental health disorders. So Decision tree has the best performance out of all the classifiers as it has higher accuracy. We used feature importance using decision and found that the Family history of the employee and history of mental health of the employee is the most contributing features when predicting mental health disorders in working employees.

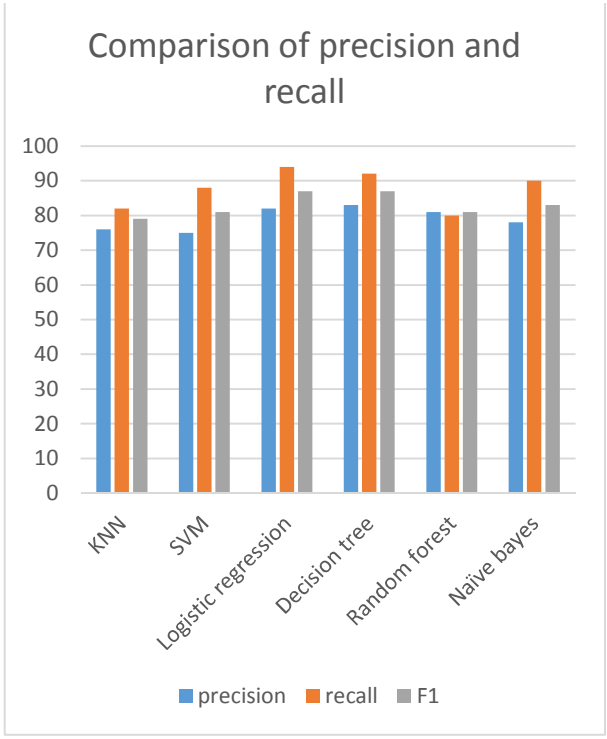


Fig.2 accuracy, precision and recall comparison of all the classifiers.

VII. Conclusion and Future Scope

After analyzing, we found that decision tree classifier has the best performance. As it has the best accuracy and precision with accuracy 84% and precision 83 followed by logistic regression with 84% accuracy and 82 precision followed by Naïve Bayes with 79% accuracy and 78 precision, random forest with 77% accuracy and 81 precision, SVM with 76% accuracy and 75 precision and KNN has the worst performance with 74% accuracy and 76 precision. Also, Feature importance of the selected features showed that a history of mental health disorder contributes most during disorder prediction followed by family history. It was also found that rest of the features contributes bare minimum to the prediction with gender as their top rest of the features which includes mental health benefits or care provided by the employer, age and discussing mental health status with the employer barely makes any contribution to the prediction of mental health disorder. Thus this paper contributes to proving that gender and company type that is the tech and non-tech even though have some influence on mental disorders. It is not prominent enough to come to a conclusion people of a gender or company type are more prone to mental disorders than the rest. But to further prove this, we need more data. For future scope deep learning and hybrid classifiers can be used for improving the accuracy when predicting mental health disorders like anxiety and mood disorder. We can improve feature selection by consulting a psychiatrist as features are directly related to the accuracy of prediction. The features selected will give a brief idea to the HR department on how to improve the working conditions of employees and provide mental health care to the employees. Such that employees with mental health disorders can work, like normal people and does not face any discrimination.

VIII. REFERENCE

[1] W. Husain, L. K. Xin, N. A. Rashid, and N. Jothi, "Predicting Generalized Anxiety Disorder among women using random forest approach," *2016 3rd Int. Conf. Comput. Inf. Sci. ICCOINS 2016 - Proc.*, pp. 37–42, 2016, doi: 10.1109/ICCOINS.2016.7783185.

[2] U. J. Frp *et al.*, "Mental disorder detection," pp. 304–308, 2019.

[3] N. F. Jie *et al.*, "Discriminating Bipolar Disorder from Major Depression Based on SVM-FoBa: Efficient Feature Selection with Multimodal Brain Imaging Data," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 4, pp. 320–331, 2015, doi: 10.1109/TAMD.2015.2440298.

[4] 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.

[5] A. Grünerbl *et al.*, "Smartphone-based recognition of states and state changes in bipolar disorder patients," *IEEE J. Biomed. Heal. Informatics*, vol. 19, no. 1, pp. 140–148, 2015, doi: 10.1109/JBHI.2014.2343154.

[6] U. Ha *et al.*, "A Wearable EEG-HEG-HRV Multimodal System with Simultaneous Monitoring of tES for Mental Health Management," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 6, pp. 758–766, 2015, doi: 10.1109/TBCAS.2015.2504959.

[7] G. Valenza *et al.*, "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 5, pp. 1625–1635, 2014, doi: 10.1109/JBHI.2013.2290382.

[8] 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.

[9] G. Giannakakis, K. Marias, and M. Tsiknakis, "A stress recognition system using HRV parameters and machine learning techniques," *2019 8th Int. Conf. Affect. Comput. Intell. Interact. Work. Demos, ACIIW 2019*, pp. 269–272, 2019, doi: 10.1109/ACIIW.2019.8925142.

[10] S. P. L. A. Pramanta, A. S. Prihatmanto, and M. G. Park, "A study on the stress identification using observed heart beat data," *Proc. 2016 6th Int. Conf. Syst. Eng. Technol. ICSET 2016*, pp. 149–152, 2017, doi: 10.1109/FIT.2016.7857555.

[11] A. Goel and S. Mahajan, "Comparison: KNN & SVM Algorithm," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 887, no. Xii, pp. 2321–9653, 2017, [Online]. Available: www.ijraset.com.

[12] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019*, pp. 135–139, 2019, doi: 10.1109/ICCSNT47585.2019.8962457.

[13] J. Huang, Y. Wei, J. Yi, and M. Liu, "An improved knn based on class contribution and feature weighting," *Proc. - 10th Int. Conf. Meas. Technol. Mechatronics Autom. ICMTMA 2018*, vol. 2018-January, pp. 313–316, 2018, doi: 10.1109/ICMTMA.2018.00083.

[14] S. S. Gavankar and S. D. Sawarkar, "Eager decision tree," *2017 2nd Int. Conf. Converg. Technol. I2CT 2017*, vol. 2017-January, pp. 837–840, 2017, doi: 10.1109/I2CT.2017.8226246.

[15] K. A. Kaplan, P. P. Hardas, S. Redline, and J. M. Zeitzer, "Correlates of sleep quality in midlife and beyond: a machine learning analysis," *Sleep Med.*, vol. 34, pp. 162–167, 2017, doi: 10.1016/j.sleep.2017.03.004.

[16] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, no. xxxx, p. 105361, 2020, doi: 10.1016/j.knosys.2019.105361.