International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Physiological Stress Prediction using Machine Learning Classifiers

Nisitaa Karen, Anuja TR, Amirtha P, R. Angeline

Abstract: ***The aim of this study is to predict the stress of a person using Machine Learning classifiers. This system classifies the stress of a person as either High or Low. There are various classification algorithms present, out of which 9 classification algorithms have been chosen for this study. The algorithms implemented are K-Nearest Neighbor classifier, Support Vector Machine with an RBF kernel, Decision Tree algorithm, Random Forest algorithm, Bagging Classifier, Adaboost algorithm, Voting classifier, Logistic Regression and MLP classifier. The different algorithms are applied on the same dataset. The dataset is obtained from a GitHub repository labelled Stress classifier with AutoML. The different accuracies of each algorithm are found, and the classification algorithm with the best accuracy is determined. On comparison, it was found that the K-Nearest Neighbor algorithm has the best accuracy with an accuracy rate of 79.3% for physiological stress prediction. While other algorithms had varying accuracies, K-Nearest Neighbor algorithm was the most consistent.***

Keywords: KNN, Machine Learning Classification, Stress Prediction.

I. INTRODUCTION

Stress is defined as the psychological response triggered by external factors. These factors can be physical or mental. Stress triggers can be increased demands at the workplace or place of study, emotional triggers such as loss of a loved one, new environments, or any sudden changes in life. Studies have proven that stress can negatively impact a person's life by increasing the risk of psychiatric and mental illnesses, and can affect the day to day performance of a person. Stress also leads to bigger risk factors such as depression, anxiety, hypertension, cardiovascular diseases and cognitive dysfunctions.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Nisitaa Karen*, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: nisitaa_karen17@srmuniv.edu.in
Anuja TR, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: anuja_tr@srmuniv.edu.in
Amirtha P, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India. Email: amirtha_p17@srmuniv.edu.in
Angeline R, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology.

Continuous monitoring of stress is necessary in order to manage the stress experienced[1]. Continued stress observance could help provide ways to identify patterns in the increase or decrease of stress, and provide the best clinical interference method, if necessary. In recent years, Personal Health Monitoring systems have become a trend, with a steady increase in users every year. The data generated by these devices have enabled easier access to such psychological and physiological signals, and other behavioral data that can be used to create assumptions about a person's basal health.

In the past few years, there has been a surge in such studies on the analysis of mental health and other related disorders. Due to this, there is a surplus of data available for use, including physiological parameters such as Heart Rate Variability (HRV)[2], Blood Pressure (BP) [3], Electrocardiogram (ECG) [4], Electroencephalogram (EEG) [5], Galvanic Skin Response (GSR) [6], Body Temperature[7], Respiration Rate[8] and Electrodermal activity[9] used as classifiers on those studies which resulted to different levels of accuracy.

II. METHODOLOGY

The original data comes from a project conducted at MIT by Healey as a part of her PhD thesis[10], and consist of body measurements conducted on various young people driving in stressing environments, e.g. rush hour, highways, red lights, as well as a relaxation period to create a non-stressed base reading. The final data is obtained from a GitHub repository labelled Stress classifier with AutoML[11]. The obtained data doesn't contain any direct labels for stress, and had to be determined manually. The median of the galvanic skin response value was taken as the cut-off point to determine the stressed state, any value above the median value was labelled as stress, and any value below the median value was labelled as not stressed, and the problem was framed as a binary task. This is called as cleaning and preparing of data. The cleaned and prepared dataset can now be loaded, and the required algorithms can be applied.

Next, feature selection was performed. Originally, the dataset consisted of 23 parameters. From this, by trial and error, certain features were selected to obtain the most optimum results. 11 features were selected. Now, the algorithms are applied on this feature selected dataset.

The first algorithm that was applied was K-Nearest Neighbor (KNN). KNN is a supervised algorithm, that is non parametric in its function, and is commonly used for classification. It is used on a dataset in which the data is separated into classes to predict the class of a new data.

The structure of the model is based on the data. In KNN, the training phase is not explicit and is insignificant compared to other algorithms. Hence, the training phase is quick. A discrete variable is obtained as an output. The assigning of classes to an object is determined by the majority vote of the neighbors. KNN has a relatively high accuracy. Next, an SVM classifier was applied. Support Vector Machines (SVM) is a supervised machine learning algorithm that can be used for both classification and regression problems. Classification algorithm is used for the purpose of this study. Every value of data is plotted in a dimension space with n dimensions, where n is the number of features chosen. The value of a coordinate is determined by the value of each feature [13]. Then, Classification is performed so as to find the best hyper-plane differentiating the two or more classes distinctly. The SVM algorithm is implemented in practice using a kernel. Radial Basis Function (RBF) is used [14].

The next algorithm applied was Decision Tree classifier. A Decision Tree is considered one of the most popular and powerful algorithms for classification and regression problems. A decision tree is a tree where each node represents a feature, each branch represents a decision(rule) and each leaf represents an outcome(categorical or continuous value)[15]. The dataset is divided into many subsets based on an attribute value test and training phase is performed. This process is performed repetitively in a recursive fashion known as recursive partitioning. After the training phase, the system is tested using the test set.

A Random Forest Classifier was also applied. The Random Forest Classifier is considered to be near the top of the Classification Hierarchy. Random Forest algorithm consists of a set of Decision Trees that operate together as an ensemble. A large number of relatively uncorrelated trees operating as one individual will outperform any of the individual constituent models[16]. The trees protect each other from their individual errors. The predictions made by the individual trees need to have low correlation to each other. Decision Trees are very sensitive to the data they are trained on, i.e., a minor change in the given dataset can lead to a vast change in the overall result. The individual trees take advantage of this and randomly sample from the dataset to lead to an overall better result.

Bagging Classifier makes use of the Bagging technique. The technique can be explained in very simple terms as putting things in a bag under certain predefined conditions. It is an application of Ensemble Modelling, i.e., Divide and Conquer. They involve a group of predictive classifier models on the same dataset, sampled randomly, to achieve better stability and accuracy.[17] Here, the smaller set of models used are Random Forest classifiers. All these multiple trees are combined to form the final outcome of the Bagging Classifier.

Adaboost classifier is an another Ensemble Classifier. Adaboost classifier combines many different weak classifiers to form a more robust and stable classifier. It trains and retrains the system iteratively by choosing the training set based on the accuracy of the previous training set[18]. The weightage of each training set in each iteration depends on the accuracy achieved in the previous iteration. A weight is assigned to a classifier after training at every level. If a set is misclassified, it is given a higher weight, and this causes the set to appear in the successive iteration with a greater probability. A classifier with 50% accuracy is assigned a weightage of 0, and if it has an accuracy of less than 50%, it takes a negative weightage. Classifier with higher accuracy is given a larger weightage so that it may increase the impact it has on the final result.

Voting Classifier also falls under the Ensemble Classifier. Voting Classifier is one of the simplest ways to combine the results from various classification algorithms to form one whole result. Voting Classifier is not strictly a classifier but is actually a wrapper for different algorithms that are evaluated parallelly to exploit the differences among them[19]. We can use a set of different algorithms and combine them to predict the final result. The final output on prediction is taken according to two strategies: Hard Voting and Soft Voting. In this project, Hard Voting is used. Hard Voting is one of the simplest cases of majority voting. The class with the highest number of votes is chosen.

Logistic Regression is Machine Learning Classification algorithm that is used to predict the probability of a categorical dependent variable[20]. It requires that only meaningful variables be included, and the dataset be relatively large.

A Multilayer Perceptron (MLP) is a mathematical model that separates a set of data into two groups, even when no simple separation exists. It is an application of Neural Networks and Deep Learning. A Perceptron takes multiple binary inputs and produces a single binary output. In an MLP, there are several layers of perceptrons each having weighted binary inputs to make decisions at each level to produce the required output. Each perceptron layer takes the output of the previous layer as its input, and gives a binary output which is used as the input for its successive layer. Finally, a single binary output is produced which can be used for classification[21].

III. IMPLEMENTATION AND RESULTS

On the final dataset obtained after data cleaning and preprocessing, the different algorithms were applied. Prediction and classification was run on the data, and the respective accuracies of each classifier was found. For a K-Nearest Neighbor classifier, the accuracy obtained was 79.3%. Next, an accuracy of 61.86 was obtained for an SVM classifier with an RBF kernel. The Ensemble Classifiers, Decision Tree classifier, Random Forest Classifier, Bagging Classifier, Adaboost classifier, and Voting Classifier had the accuracies of 74.33%, 74.46%, 76.27%,75.91% and 63. 32% respectively. Logistic Regression gave an accuracy of 59.56%, while MLP gave 60.29%. The comparison of the accuracies is given in Fig 1.

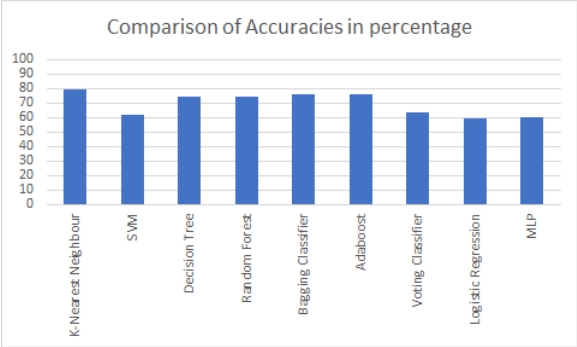

Fig 1. Comparison of Accuracies in percentage

Table I : Algorithms and their accuracies

Algorithms	Accuracy in Percentage
Logistic Regression	59.56
MLP	60.29
SVM	61.86
Voting Classifier	63.32
Decision Tree	74.33
Random Forest	74.46
Adaboost	75.91
Bagging Classifier	76.27
K-Nearest Neighbor	79.3

On comparison, it is found that K-Nearest Neighbor had the highest accuracy with 79.3%. Bagging Classifier had the second highest accuracy with 76.27%. All the Ensemble Classifiers had close accuracies in the range of 63%-77%. Logistic Regression Classifier had the lowest accuracy of 59.56%. The accuracies of the different classifiers are tabulated in Table I.

IV. CONCLUSION

We can conclude that the K-Nearest Neighbor classifier is the most efficient algorithm for Physiological Stress Prediction, with the highest accuracy of 79.3%. While other algorithms had varying accuracies, KNN produced the most consistent accuracies. Hence, for any future work conducted on Stress Prediction, it can be concluded that KNN can give optimal results. Since only the 9 most popular classifiers were analyzed in this paper, the accuracies of other classifiers may or may not be better than KNN. The accuracies can also vary depending on the features chosen.

ACKNOWLEDGMENT

A heartful thanks to Mrs. R. Angeline for her constant motivation and support throughout the tenure of this project. The development of this project would not have been possible without the Computer Science and Engineering department of our college SRM Institute of Science and Technology, Ramapuram.

REFERENCES

1. Rodney Karlo C. Pascual, John Paul D. Serrano, Jamie Mitchelle A. Soltez, John Christopher D. Castillo, Jumelyn L. Torres, and Febus Reidj G. Cruz, “Artificial Neural Network Based Stress Level Detection System using Physiological Signals”. IEEE, 2018.
2. Munla, N, “Driver stress level detection using HRV analysis”. International Conference on Advances in Biomedical Engineering (ICABME),2015.
3. Fernandez, A., et al., “Determination of stress using Blood Pressure and Galvanic Skin Response”. International Conference on Communication and Network Technologies (CCNT), 2014.
4. Karthikeyan, P., et al., “ECG signals based mental stress assessment using wavelet transform”. IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 2012.
5. Hu, B, “Signal Quality Assessment Model for Wearable EEG Sensor on Prediction of Mental Stress”. IEEE Transactions on NanoBioscience, 2015.
6. Atlee Fernandes, Rakesh Helawar, R. Lokesh, Tushar Tari and Ashwini V. Shahapurkar, “), Determination of Stress using Blood Pressure and Galvanic Skin Response”. International Conference on Communication and Network Technologies (ICCNT), 2014.
7. Caya, M.V.C., et al., “). Basal body temperature measurement using e-textile”. 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, 2018.
8. Gandhi, S, “Mental stress assessment – a comparison between HRV based and respiration based techniques”. Computing in Cardiology Conference (CinC), IEEE Conferences, 2016.
9. S Ward, M Brickleg, J Shany, G McDarby, C Heneghan, “Assessment of Heart Rate and Electrodermal Activity during Sustained Attention to Response Tests”. IEEE 2004.
10. Jennifer A. Healey, “Wearable and Automotive Systems for Affect Recognition from Physiology”. Massachusetts Institute of Technology, 2000.

AUTHORS PROFILE

Nisitaa Karen is currently pursuing her Bachelor degree in computer science and engineering from the prestigious SRM institute of science and technology. Her current interests lie in the domains of machine learning and data science. She is a meritorious student. She is planning to pursue her masters in one of the above domains.

Anuja T.R. is currently pursuing her Bachelor degree in Computer Science and Engineering from the prestigious SRM Institute of Science and Technology. Her current interests lie in the domains of Algorithms and Machine Learning and is looking out for a prospective career in her domain. She is a meritorious student.

Amirtha P is currently pursuing her Bachelor degree in Computer Science and Engineering from the prestigious SRM Institute of Science and Technology. She is interested in the domain of machine learning and also intrigued in the domain of Internet of Things. She is looking out for a promising career in her domain.

Angeline R is currently an assistant professor at the computer science and engineering department of SRM institute of science and technology. Her field of expertise is deep learning.