Modelling of Stress data for Nurses

April 25, 2023

Registration number: 2238
Link to GitHub: https://github.com/Tharunillendula/Stress-Detection

Executive summary (max. 200 words)	118
Main findings and Discussion (max. 600 words)	490
Conclusions (max. 300 words)	220
Total word count	828

Contents

Abstract

Your executive summary/abstract goes here.

1 Main Findings

1.1 Exploratory data analysis

The total data is extremely large, and data analysis and modeling cannot be done without proper computing resources. Reducing the total amount of data analysis on the summary report helps us choose specific IDs of people to have maximized information gain.
The figures below show the pots of each column against the data. The following figures show that temperature and heart rate directly relate to the class labels.

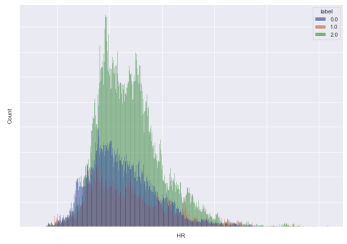

Figure 1: .

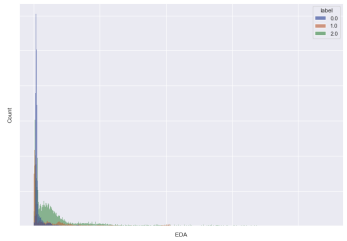

Figure 2:

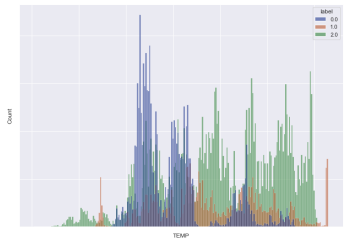

Figure 3:

Conducting the PCA on the data shows evident proof for the above and below is the table with contributions to the data variation in eigen values.
For the purpose of building a classifier I have taken multiple appraches including Random Forest approach(including various columns),SVM(linear,radial),Neural Network(2 layers).

COLUMN	EIGEN VALUES
X	0.00188695
Y	0.00824138
Z	0.05859564
EDA	0.18567169
HR	0.35180726
TEMP	0.39379707

2 Discussion

Data normalization is an essential preprocessing step that can improve the accuracy, stability, and generalization of machine learning models. Applied normalization because each column of the data is on a different scale and to get the same amount of representational power from each input this was done.
A Random Forest Classifier is a popular machine learning algorithm used for classification and regression tasks. It is an ensemble method that combines multiple decision trees, where each tree is trained on a random subset of the training data and a random subset of the features.
To gain a deeper understanding of the given data and identify potential patterns or relationships, a random classifier approach was used as a starting point. The aim was to explore the data and obtain initial insights that could inform subsequent analyses. Through the process of cross-validation, it was determined that using 100 estimators, or 100 decision trees, was the most effective approach to achieving high accuracy levels.

Further analysis revealed that using all the available columns in the data did not result in the highest accuracy levels, highlighting the importance of feature selection in machine learning. Using only the highest PCA eigenvectors, which were Temperature and Heart rate, led to significantly better results. This finding underscores the importance of selecting the most relevant variables when training machine learning models.

To further explore classification options, a Support Vector Machine (SVM) was applied to the data. SVM is a popular machine learning algorithm that can be used for classification or regression tasks. The kernel function, which enables the algorithm to function in high-dimensional feature spaces and capture complicated interactions between the data points, is a crucial component of SVM. While the linear kernel function provided the best results, it did not perform better than the Random classifier approach.

Finally, a neural network with two layers was developed for the classification task. The first layer was a fully connected layer that represented the input in a higher dimension. The final layer applied a sigmoid function to the downsize layer to obtain the class probabilities of each label. The learning rate was set to 1e-3, and a batch size of 32 was used for the training with 8 epochs.In addition to the hyperparameters mentioned, there are other important factors to consider when developing a neural network. One such factor is the choice of activation functions for the hidden layers. Common activation functions including sigmoid, ReLU (Rectified Linear Unit) and tanh worked well for our use case, each with its own advantages and disadvantages depending on the task. Due to limited resources, many hyperparameters in the model were limited to search, but further training could definitely improve the model's performance. Batch normalization and weight initialization played a key factor in minimizing the overfitting on a class and getting better generalization. With its high representational power, the neural network has the potential to outperform other approaches and achieve the best results.

3 Conclusions

The research, which looked at whether the sensors on a worn watch might be utilized to detect and/or forecast stress levels, was completed successfully. To find the best reliable strategy for identifying stress levels, a variety of machine learning algorithms were used to the data set of nurse stress levels. When employing a Random Forest Classifier, the first study suggested that temperature and heart rate were the most crucial signals to take into account. It was essential to choose the right features since utilizing just the highest PCA eigenvectors considerably improved the outcomes. Support Vector Machines and a two-layer neural network were used for more research. While the linear kernel function produced the greatest results for SVM, the neural network showed significant promise for improving outcomes even further with more training and hyperparameter tweaking. These models have higher representational power and when more data is obtained on a wider range we get better and more accurate results. Overall, the experiment has demonstrated that it is feasible to identify stress levels using the wearable watch's built-in sensors. To obtain high levels of accuracy, it is crucial to choose the most pertinent signals and apply the right machine-learning methods. This is especially important when trying to persuade a hospital to accept a contract or when false negatives are a major problem.

Model	Accuracy	Precision(avg)	Recall(avg)
Random Forest	90.2	88.46	92.32
Random Forest(HR,TEMP)	98.39	97.60	95.13
SVM Radial	94.7	90.12	95.88
Neural Network	84.22	76.13	78.29

Table 1: Final Summary.