Human Stress Detection During Sleep Based on Physiological Data

Arnav Kumar Palia
Dept. of computer Science, Indraprastha Institute of Information and Technology. Delhi
arnav20363@iiitd.ac.in

Mohd Faizan
Dept. of computer Science, Indraprastha Institute of Information and Technology. Delhi
faizan20313@iiitd.ac.in

Chirag Saraswat
Dept. of computer Science, Indraprastha Institute of Information and Technology. Delhi
chirag22002@iiitd.ac.in

Yearagra Paliwal
Dept. of computer Science,
Indraprastha Institute of Information and Technology. Delhi
yearagra22004@iiitd.ac.in

1.Abstract

As we know keep track of your mental health and regular mental health checkup is very important in this busy and stress full world but regular checkup quite expensive and time-consuming process, so, our main motive is to design a highly efficient machine learning model that will predict users stress level accurately and quickly by using users' physiological data.

2. Introduction

In this fast and day to day busy life almost every 2nd working adult is going through some kind of stress but the fact is that most them do not even know about mental health and stress, the reason is for a normal person its hard know about their mental health and stress level until they concern to a doctor.

So, in this project we will try to find a solution by implementing our knowledge and learning of machine learning.

3. Literature Review

To better stand our problem and find more prominent solution we referred following research paper:

- *Review on Psychological Stress Detection Using Bio signals [1]*
- *Machine Learning Based Solutions for Real-Time Stress Monitoring.* [2]
- *Requirement and Design of Mental Health System for Stress management of Knowledge.* [3]

3.1. Paper 1

The Title: *Review on Psychological Stress Detection Using Bio signals. [1]*

In this paper, the impact of stress to multiple bodily responses is surveyed. paper aims to provide a comprehensive review on bio-signal patterns caused during stress conditions and practical guidelines towards more efficient detection of stress. Emphasis is put on the efficiency, robustness and consistency of bio signal data features across the current state of knowledge in stress detection.

3.2. Paper 2

The Title: *Machine Learning Based Solutions for Real-Time Stress Monitoring.* [2]

Unlike other studies, the authors in this paper also evaluated extreme learning machines [ELM] algorithm with five different activation functions apart from Random Forest Classifier, Support Vector Machines, Artificial Neural Network, and Radial Based function network.

3.3. Paper 3

The Title: *Requirement and Design of Mental Health System for Stress management of Knowledge.* [3]

The research paper talks about how an average worker goes through the stress in their daily life and also factors that affects stress level. The insights we get is that paper provides ranked list of essential features help us to know considerable factors for stress level and also which machine learning algorithm will suitable for such classification.

TABLE I. RANKING OF THE STRESS MEASUREMENTS [5]

Rank	Primary Measure
1	Heart Rate Variability (HRV)
2	Galvanic Skin Response (GSR)
3	Electroencephalography (EEG)
4	Pupil Diameter (PD)
5	Voice
6	Eye Gaze
7	Facial Expression
8	Blood Pressure (BP)
9	Skin Temperature (ST)
10	Blood Volume Pulse (BVP)

TABLE II. RANKING OF MACHINE LEARNING ALGORITHMS FOR STRESS MODEL [7]

Rank	Primary Measure
1	Support Vector Machine (SVM)
2	Recurrent ANN
3	Adaptive Neuro-fuzzy System
4	Artificial Neural Network (ANN)
5	Hidden Markov Model (HMM)
6	Decision Tree
7	Naïve Bayesian Network
8	Fuzzy Clustering

4. Dataset

We have used physiological user collected data from Kaggle data repository [4]. The Dataset has total 1000+ entries with 9 attributes: snoring rate, respiration rate, body temperature, limb movement, blood oxygen, eye movement, sleeping hours, heart rate, stress level. Where stress level is the target attribute with in data type and rest are the features with float data type.

4.1. Data Pre-processing

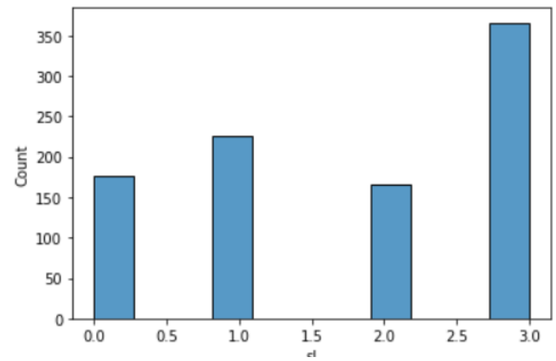
1) Feature selection and cleaning - We have found our features relevant, removed duplicate data samples and replaced the null values by mean value of column. We thus obtained 935 data samples and 9 features.

2) Normalization - In order to give equal weights to each feature in the dataset so that no single variable steers model performance in one direction, we performed data normalization using Minmax Scaler to fit between 0 and 1.

3) Oversampling data -

- ❖ The division of data into the four classes is imbalanced.
- ❖ We have used the random oversampling technique to avoid overfitting of the machine learning model on skewed classes by increasing the data samples of the classes with minority instances.

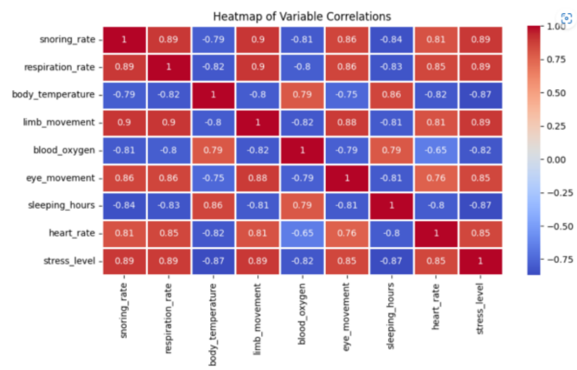
<matplotlib.axes._subplots.AxesSubplot at 0x7f68fc5a12e0>



The above graph shows the frequency of labels (1.0 to 3.0 represents stress level in increasing order) and as represented in graph stress level has highest count/occurance.

4.2. Exploratory Data Analysis

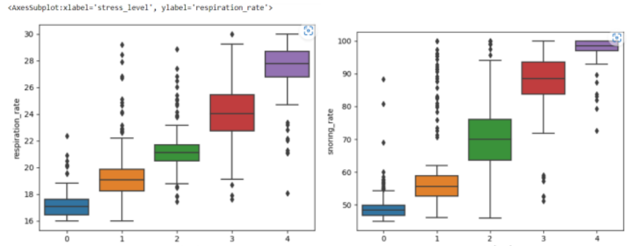
Heat Map:



The Heat map shows that all attributes are highly correlated to each other

Box Plots:

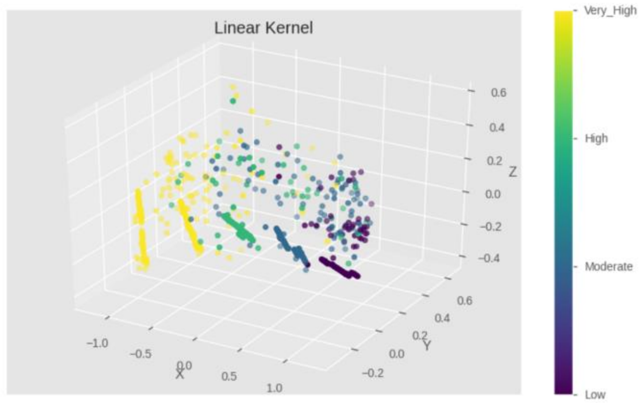
<AxesSubplot:xlabel='stress_level', ylabel='respiration_rate'>



The Box plots shows us that attributes have high number of outliers.

Linearly Seprability:

Since our dataset has 8 features, we used PCA to reduce the dimension of the dataset to 3 dimensions and plot a features vs labels graph.



From this fearture vs label we get to know that our dataset is linearly seprable in 3d.

5. Methodology

- ❖ For implementing model's dataset was split into training and testing set using a 80:20 split.
- ❖ The following models were implemented on the dataset using the sklearn library-
    - ➢ Logistic Regression
    - ➢ Naive Bayes (Gaussian & Multinomial)
    - ➢ Decision Tree
    - ➢ Random Forest
    - ➢ Adaboost
    - ➢ SVM
    - ➢ MLP
- ❖ We also performed hyperparameter tuning using GridSearchCV on some models and working on others to select out the best working model.
- ❖ We used matplotlib and seaborn plotting and visualization → histograms and matrices.

Applied Grid search technique on above models to find optimal hyperparameters for it and got the following model parameters:

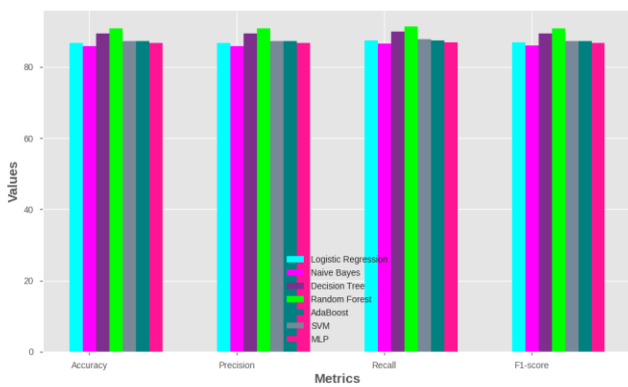
Classifier Algorithm	Optimal Parameters
Logistic Regression	C=166.81005372000558, max_iter=5000, multi_class='multinomial', penalty='l1', random_state=0, solver='saga'
Naïve Bayes	------------/-------------
Decision Tree	criterion='entropy', max_depth=31, max_features='sqrt', splitter='random
Random Forest	criterion='entropy', n_estimators=30, random_state=11
Adaboost	algorithm='SAMME', learning_rate=1, n_estimators=100
SVM	degree=4, kernel='poly', probability=True
MLP	'activation': 'logistic', 'alpha': 0.0001, 'hidden_layer_sizes': (350, 50), 'learning_rate': 'invscaling', 'solver': 'lbfgs'

6. Results and Analysis

Classifier Algorithm	Accuracy	Precision	Recall	F1 score
Logistic Regression	86.78	87.45	86.78	86.91
Naïve Bayes	85.90	86.48	85.90	86.04
Decision Tree	89.42	89.87	89.42	89.45
Random Forest	90.74	91.44	90.74	90.79
Adaboost	87.22	87.38	87.22	87.19
SVM	87.22	87.84	87.22	87.34
MLP	86.78	86.84	86.78	86.71

On observing the performance matrix results, highest accuracy has been observed in Random Forest (90.74%), and Naïve bayes gives the least accuracy (85.90 %).

- ❖ Since our dataset has only 936 inputs with all data in numeric form with multi dimension features (8 features and 4 labels) and also has outliers in it therefore, Random Forest performs better here than other algorithm because Random Forest Classifier performs better with more categorical data and robust to outliers since they get averaged out by the aggregation of multiple tree output. It works really well with non-linear data. There is a low risk of overfitting, as the results are calculated based on the output of multiple decision trees whereas Logistic Regression is a little confusing when comes to categorical data and for MLP, it large amount of data to train and updates the weights.
- ❖ Whereas Logistic Regression is a little confusing when comes to categorical data and for MLP, it needs large amount of data to train and updates the weights.
- ❖ For SVM our dataset was not so perfectly (as we can see above in PCA 3d graph) separable therefore the accuracy achieved in SVM is less than Random Forest.
- ❖ Since our dataset has outliers therefore the performance of Adaboost is less than Random Forest. As we know Adaboost performs well when there is No noise and Outliers.



The above graph also shows the comparison between all used 7 algorithm and we can see all algorithms performs really well (for algorithm accuracy was almost > 85%) and also the accuracies all algorithm is very close to each other where Random Forest gives the highest accuracy.

7.Conclusion

- ❖ By Working on real life data, we learnt that pre-processing plays a major role, followed with right machine learning model and tools can bring a satisfactory change in classification score analysis.
- ❖ Our methodology is not limited to data pre-processing and cleaning, but also includes feature selection and engineering methods like PCA used for dimensionality reduction.
- ❖ We can also use reduced dimensions dataset of PCA to train SVM model because PCA 3d graph for labels shows us that data is linearly separable up to some points.
- ❖ We have achieved almost fine accuracy, precision, recall, and F1 score metrics of around in 7 models.

References

[1] Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A. and Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, pp.1–1. doi:10.1109/taffc.2019.2927337.

[2] Nath, R.K., Thapliyal, H., Caban-Holt, A. and Mohanty, S.P. (2020). Machine Learning Based Solutions for Real-Time Stress Monitoring. *IEEE Consumer Electronics Magazine*, [online] 9(5), pp.34–41. doi:10.1109/MCE.2020.2993427.

[3] ieeexplore.ieee.org. (n.d.). *IEEE Xplore Search Results*. [online] Available at: https://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=Requirement%20and%20Design%20of%20Mental%20Health%20System%20for%20Stress%20management%20of%20Knowledge [Accessed 5 Dec. 2022].

[4] www.kaggle.com. (n.d.). *Human Stress Detection in and through Sleep*. [online] Available at: https://www.kaggle.com/datasets/laavanya/human-stress-detection-in-and-through-sleep.