DSS-740 - Analytics w/ Machine Learning

FINAL PROJECT

**on**

**Predicting heart rate to monitor stress level.**

By:

Aquina Panna

# Introduction

In today’s world, Anxiety and Stress have become a very common conditions among young and middle-aged people. According to an article, 40 million adults in the United States age 18 and older, or 18.1% of the population every year are getting affected by Anxiety. A survey shows that Gen Z teens (ages 13–17) and Gen Z adults (ages 18–23) are facing unprecedented uncertainty, are experiencing elevated stress, and are already reporting symptoms of depression.

Stress is the difficulty of an organism to maintain its homeostasis, often induced by external stimuli that cause mental or physical imbalance. It is known that when an individual is exposed to a stressor, the autonomic nervous (ANS) system is triggered resulting in the suppression of the parasympathetic nervous system and the activation of sympathetic nervous system. This reaction which is known as the fight-or-flight response can involve physiological manifestations such as: vasoconstriction of blood vessels, increased blood pressure, increased muscle tension and a change in heart rate (HR) and heart rate variability (HRV). Among these, HRV has become a standard metric for the assessment of the state of body and mind, with multiple markers derived from HRV being routinely used for identifying mental stress or lack thereof. HRV is a time series of the variation of the heart rate over time and is determined by calculating the difference in time between two consecutive occurrences of QRS-complexes, also known as the RR interval (RRI). An optimal HRV points to healthy physiological function, adaptability, and resilience. Increased HRV (beyond normal) may point to a disease or abnormality. Reduced HRV, on the other hand, points to an impaired regulatory capacity and is known to be a sign of stress, anxiety, and several other health problems. Identifying stress has been the focus of much research as an increasing body of evidence suggests a rising prevalence of stress-related health conditions associated with the stressful contemporary lifestyle.

With the rise of modern machine learning and deep learning methods, these methods have been applied in the study of heart rate variability. Machine learning and deep learning methods have previously been used with HRV and electrocardiography (ECG) data for various applications such as: fatigue and stress detection student stress prediction, congestive heart failure detection cardiac arrhythmia classification. Most prior arts, however, are supervised or mental stress detection.

Higher heart rate are usually connected to Stress, so we took this opportunity to work on the clinical data and to find out what all factors may affect the heart rates and how they are associated with Stress.

# Problem Definition and Algorithms

We collected ECG data and other nonlinear feature data for our analysis. There are **36 features** in the dataset and **369289 observations**. Here we are trying to predict the heart rate.

The issues we are focusing on with regards to dataset are:

* The dataset is very huge hence we are selecting the models which are fast and perform very well with large dataset, hence we are focusing on the models like: **Lasso, Ridge, Decision Tree, Gradient Boosting, and** **Neural Network**.
* Most of the features are skewed, this could affect the regression model so to normalize the skewness we used **Standard scaler.**
* The data set has 36 features hence we try to use **feature reduction technique- PCA** and **feature selecting techniques- Correlation matrix**.
* To evaluate the model we will try to keep **the Mean squared error to minimum** and **accuracy to maximum.**

A quick view on the features:

1. MEAN\_RR - Mean of RR intervals
2. MEDIAN\_RR - Median of RR intervals
3. SDRR - Standard deviation of RR intervals
4. RMSSD - Root mean square of successive RR interval differences
5. SDSD - Standard deviation of successive RR interval differences
6. SDRR\_RMSSD - Ratio of SDRR / RMSSD
7. pNN25 - Percentage of successive RR intervals that differ by more than 25 ms
8. pNN50 - Percentage of successive RR intervals that differ by more than 50 ms
9. KURT - Kurtosis of distribution of successive RR intervals
10. SKEW - Skew of distribution of successive RR intervals
11. MEAN\_REL\_RR - Mean of relative RR intervals
12. MEDIAN\_REL\_RR - Median of relative RR intervals
13. SDRR\_REL\_RR - Standard deviation of relative RR intervals
14. RMSSD\_REL\_RR - Root mean square of successive relative RR interval differences
15. SDSD\_REL\_RR - Standard deviation of successive relative RR interval differences
16. SDRR\_RMSSD\_REL\_RR - Ratio of SDRR/RMSSD for relative RR interval differences
17. KURT\_REL\_RR - Kurtosis of distribution of relative RR intervals
18. SKEW\_REL\_RR - Skewness of distribution of relative RR intervals
19. uuid - Unique ID for each patient
20. VLF - Absolute power of the very low frequency band (0.0033 - 0.04 Hz)
21. VLF\_PCT - Principal component transform of VLF
22. LF - Absolute power of the low frequency band (0.04 - 0.15 Hz)
23. LF\_PCT - Principal component transform of LF
24. LF\_NU - Absolute power of the low frequency band in normal units
25. HF - Absolute power of the high frequency band (0.15 - 0.4 Hz)
26. HF\_PCT - Principal component transform of HF
27. HF\_NU - Absolute power of the highest frequency band in normal units
28. TP - Total power of RR intervals
29. LF\_HF - Ratio of LF to HF
30. HF\_LF - Ratio of HF to LF
31. SD1 - Poincaré plot standard deviation perpendicular to the line of identity
32. SD2 - Poincaré plot standard deviation along the line of identity
33. Sampen - sample entropy which measures the regularity and complexity of a timeseries
34. higuci - higuci fractal dimension of heartrate
35. datasetId - ID of the whole dataset
36. condition - condition of the patient at the time the data was recorded
37. HR - Heart rate of the patient at the time of data recorded ---------------------- **To be Predicted.**

The dataset has no missing values. Also we removed 2 features- uuid and datasetid from the dataset because these 2 feature has no relevance in building model.

# Experimental Evaluation

We considered building model with selected feature using PCA and observed that the accuracy score of 0.68 which is much lower than linear regression model, hence we decided to reject PCA-regression model.

The base linear regression model gave an accuracy of 0.9809 on training data but miserably fails on test data, model gave the accuracy of -1.2734997593792958e.

Hence, we build hyper tuned Lasso regression model on all the 34 features to improve the performance and we achieve the accuracy of 0.9809 on training data and 0.9807 on test data. The MSE for this model is 2.074

We built hyper tuned Ridge model on full features and achieved accuracy of 0.9809 on training data and 0.9807 on test data. The MSE for this model is 2.072. From this we concluded that Lasso and Ridge work all most same.

We built decision tree on full features and achieved accuracy of 0. 0.9999 on training data and 0.9998 on test data. The tree recoded MSE of 0.021. This seems overfitted so we pruned the tree to 'max\_depth': 12, 'max\_features': 4, 'min\_samples\_leaf': 1 and model gave train accuracy of 0.9939 and test accuracy of 0.9928. The pruned tree has MSE of 0.769.

We build Gradient Boosting Model on full features and achieved accuracy of 0.9972 on training data and 0.9971 on test data. The MSE is 0.311.

Finally, we built **Neural Network** on full features and achieved **accuracy of 0.9998 on training data** and **0.9998 on test data**. This is the best performing model and the **MSE recorded of 0.018** which is good.

Other models we built on selected 24 features ('SKEW', 'KURT', 'higuci', 'SDRR\_RMSSD\_REL\_RR', 'LF\_NU', 'VLF\_PCT', 'LF', 'MEDIAN\_REL\_RR', 'SD1', 'LF\_PCT', 'HF\_NU', 'MEAN\_REL\_RR', 'pNN50', 'SD2', 'HF\_PCT', 'MEAN\_RR', 'HF', 'sampen', 'SDRR\_REL\_RR', 'condition\_time pressure', 'LF\_HF', 'condition\_no stress', 'VLF', 'RMSSD\_REL\_RR') are as below:

We built Decision Tree with parameters max\_depth=12,max\_features=7,max\_leaf\_nodes=6 and achieved accuracy of 0.7246 on training data and 0.7243 on test data. It has relatively higher MSE of 29.69

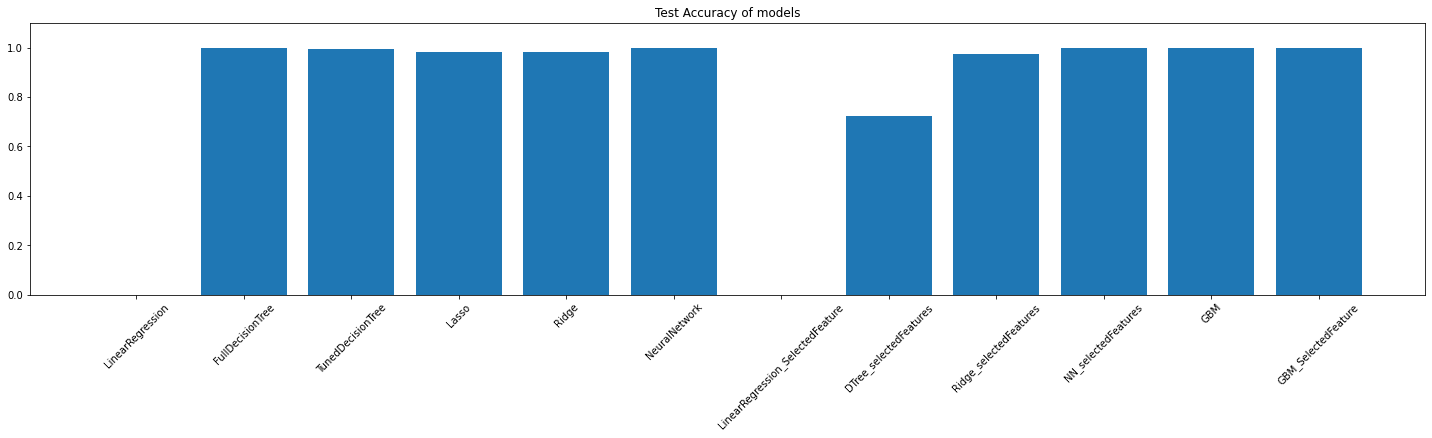
We built Ridge Model and achieved accuracy of 0.9759 on training data and 0.9755 on test data. The MSE id 2.634.The top 5 important features used in this model are: 'RMSSD\_REL\_RR', SD2, LF, 'SDRR\_RMSSD\_REL\_RR' and LF\_HF.

We built Gradient boosting model and achieved accuracy of 0.9984 on training data and 0.9983 on test data.

We built Neural network and achieved accuracy of 0.9998 on training data and 0.9997 on test data. The MSE is 0.03.

# Conclusion

The Neural Network built on all features outperformed other models in terms of accuracy and Mean squared Error. The accuracy of test and training data is almost same.

# Future Work

Though we have achieved test accuracy of 0.9998 and MSE of 0.018 with Neural Network we still believe the model can be improved by adding a good number of neurons and adding back proposition. Any model which deals with medicine and health needs to be fine tunned as much as we can. So, we will try to carry on optimizing the model.

# Appendix

The project is created in [google colab](https://colab.research.google.com/drive/1x1WLw8pR4Yz5DkRSMPIpDL-R_s_u5NDq?usp=sharing). Also the code is share in [github](https://github.com/Aquina-Panna/Data-Science-Portfolio)

Important screenshots:

