**WOA7015 ADVANCED MACHINE LEARNING**

Alternative Assessment

Academic Session 2022/2023: Semester 1

Jan 2023

**Project Proposal**

|  |  |  |
| --- | --- | --- |
| No | Team Members | Matric Number |
| 1 | Chong Chia Hsing | S2159070/1 |
| 2 | Darkhan Baibulat | S2172558/1 |
| 3 | Xue Ying Kung | S2149135/1 |

Table of Contents

[2 Introduction 3](#_Toc123298012)

[3 How to obtain the heart rate (target variable) from the dataset? 3](#_Toc123298013)

[4 How to handle the data (Pre-processing / Preparation)? 3](#_Toc123298014)

[5 What algorithm is used or built (model architecture)? 4](#_Toc123298015)

[6 Why can the stated algorithm solve the problem? 4](#_Toc123298016)

[7 What are the baseline models or benchmarks? 4](#_Toc123298017)

[8 What are the performance metrics used? 5](#_Toc123298018)

# Introduction

The project aim is to correlate and predict the heart rate from the respiratory signal. The datasets are obtained from the website, <https://physionet.org/content/picsdb/1.0.0/>, where there are 10 sets of data consist of both the respiratory and ECG signal for 10 different pre-term infants.

# How to obtain the heart rate (target variable) from the dataset?

For obtaining the Heart Rate, we use the function wfdb.processing.compute\_hr(sig\_len, qrs\_inds, fs) to calculate it, using the numpy and wdfb libraries. A brief description of the variables used to find the target variable:

* Sig\_len is size of the signal
* Qrs\_inds is Using file .qrsc
* Fs is the frequency

Web Link: <https://github.com/MIT-LCP/wfdb-python/blob/main/wfdb/processing/hr.py>

# How to handle the data (Pre-processing / Preparation)?

The data pre-processing involves the process of preparing in proper format and quality checking on the data before input to the model training process for better model prediction.

1. The 40+ hours dataset per infant are first split into a 1-hour bin for easier handling of the data and then it is visually inspected for the quality of the data.
2. Within the 1-hour bin data, split the data into training and test data sets at 70% ratio.
3. Perform heart rate calculation on the ECG data sets.
4. Perform some correlation test to investigate the correlation for the time base data using the covariance matrix, Pearson’s correlation (linear), and Spearman’s correlation (polynomial).
5. Perform Fourier Transform on Heart Rate and Respiratory Signal to obtain a frequency-based data.
6. Repeat the correlation test on frequency-based data.

# What algorithm is used or built (model architecture)?

Support Vector Machine (SVM) is used in this project. SVM is a supervised learning model with associated learning algorithms that analyse data for classification and regression. The basic idea of SVM is to find a hyperplane that best divides a dataset into two classes. For Support Vector Regression (SVR), the goal is to find the best fit line, which is the hyperplane that has the maximum number of points.

# Why can the stated algorithm solve the problem?

The main problem is to predict heartbeat rates of infants based on their respiratory rates, which is a regression problem. Support Vector Machine (SVM) is an algorithm that can solve regression problems.

SVM is chosen as the model for this project and the reasons are stated as below:

1. SVM has L2 Regularization feature, which has good generalization capabilities to prevent it from overfitting.
   * Preventing overfitting issues in health domain research is important because every individual has some small difference, which is normal.
2. Linear SVM handles outliers better, as it derives maximum margin solution.

* Even if there is any infant in the training data who is having health problems and has a significantly different relationship between his/her heartbeat rate and respiratory rate, his/her training data will not have too much negative impact on the model.

1. SVM can perform better than a few other algorithms when there is limited training data. Other algorithms, for example Neural Network needs large training data for sufficient accuracy.

* This is an important point because in our project, there is only data for 10 infants.

# What are the baseline models or benchmarks?

The baseline models are the simplest regression and data exploration technique used to do some initial inspection and testing on the dataset. Basic linear regression and polynomial regression will be used as the baseline models for the project to inspect the correlation of the 2 variables: ECG signal and respiratory signal.

|  |  |
| --- | --- |
| **Sklearn modules** | Description |
| [**linear\_model.LinearRegression**](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression)(\*[, ...]) | Ordinary least squares Linear Regression. |
| [**preprocessing.PolynomialFeatures**](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html#sklearn.preprocessing.PolynomialFeatures)([degree, ...]) | Generate polynomial and interaction features. |

# What are the performance metrics used?

There are 3 error or performance metrics that will be applied for the regression model:

* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* Mean Absolute Error (MAE)

|  |  |
| --- | --- |
| **Sklearn modules** | Description |
| [**metrics.mean\_absolute\_error**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html#sklearn.metrics.mean_absolute_error)(y\_true, y\_pred, \*) | Mean absolute error regression loss. |
| [**metrics.mean\_squared\_error**](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html#sklearn.metrics.mean_squared_error)(y\_true, y\_pred, \*) | Mean squared error regression loss. |