**Biostat 626 Midterm 1**

**Human Activity Recognition using Wearable Sensor Technology**

1. Complete the Task 1 and 2 and submit your classification results via Canvas.

**Submitted.**

**Could find all classification results in the leaderboard at** [**https://github.com/xqwen/bios626**](https://github.com/xqwen/bios626)**. The SID is l1281.**

1. Set up a Github repository and upload all your code used for training, evaluation, and generating results of test data. Provide the url to your Github repository as the answer to this question.

**The GitHub repository could be found at** [**https://github.com/haisx/Biostat626ActivityRecognition**](https://github.com/haisx/Biostat626ActivityRecognition)**.**

3. Write a text file (name the file “README.md”) to provide necessary instructions, so

that other people can reproduce all your results.

**README.md file could be found at** [**https://github.com/haisx/Biostat626ActivityRecognition/blob/main/README.md**](https://github.com/haisx/Biostat626ActivityRecognition/blob/main/README.md)**.**

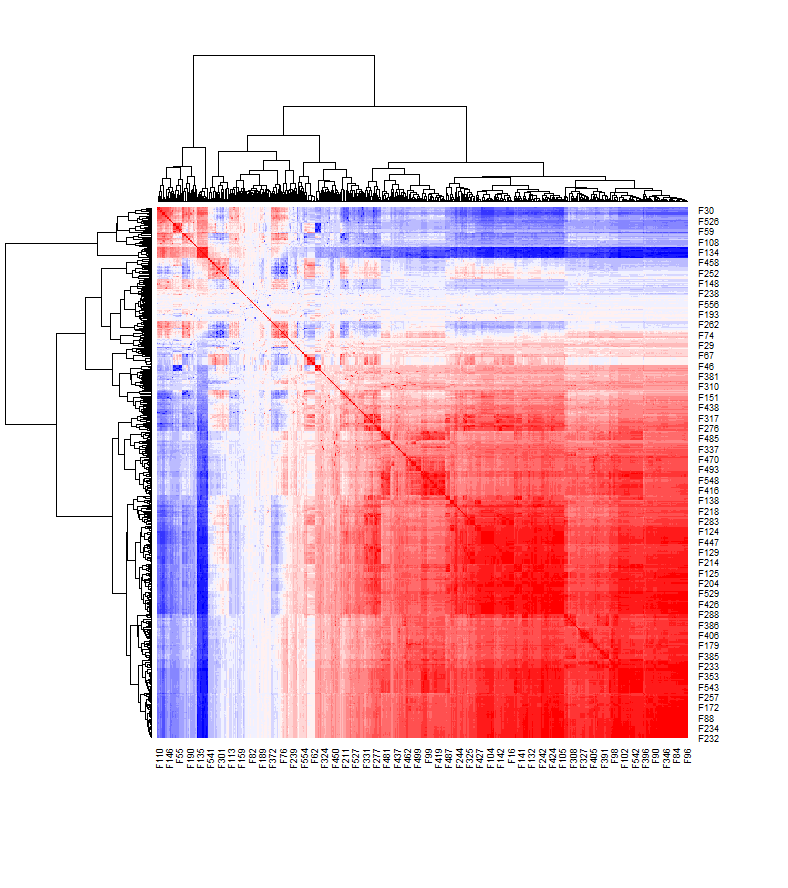
4. Describe your baseline algorithm and provide necessary tables and/or figures to sum-

marize its performance based on the training data.

**The dataset is a high dimensional dataset which contains numerous variables for classification. To get started, I first looked at the correlations between numeric variables in the training dataset. We can find that the dataset contains a lot of variables, and many variables are highly correlated with others. Thus, it is important to drop some useless variables and drop some highly correlated variables. The baseline algorithm here is to conduct logistic regression model with a few important variables based on the training data.**

**The most important variables were selected by the variables of important based on random forest method, variable importance based on Generalized cross validation, stepwise regression and so on.**

**Correlation heatmap on training data.**



===== Binary classification========

**Logistic regression with the 4 most important variables (F282, F141, F201, F182) in the training dataset.**

A picture containing text

Description automatically generated

**Accuracy: 0.964**

**Logistic regression with the 10 most important variables (F96, F103, F258, F272, F270, F259, F201, F405, F345) on training dataset.**

Diagram

Description automatically generated with low confidence

**Accuracy: 0.9952**

===== Multiclass classification========

**Multinomial logistic regression with the 10 most important variables (F96, F103, F258, F272, F270, F259, F201, F405, F345) on training dataset**

A picture containing text

Description automatically generated

**Accuracy = 0.7264**

==================================

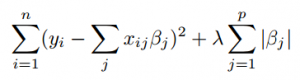
**More variables will lead to a higher accuracy on training data but may face overfitting issues on test data. LASSO regression may return a better result than hand choosing variables due to an automatically shrinkage of variables based on lambda.**

**Therefore, I tried LASSO regression on binary classification and returned an appropriate result (accuracy = 1.000 on training data). Next, for multiclass classification, I used multinomial logistic regression using LASSO regression as well since I got a good accuracy on binary classification using LASSO.**

5. Describe your final algorithm and provide necessary tables and/or figures to summarize

its performance based on the training data.

**Binary classification - LASSO Regression with a threshold 0.52 and a lambda = 4.636301e-05.**

The LASSO regression is trying to minimize 

**Lambda is a tuning parameter, which controls the strength of the L1 penalty.**

**The LASSO performance was based on training data.**

**Diagram

Description automatically generated**

**Accuracy = 1.000**

**A multinomial logistic regression using LASSO.**

**The LASSO performance was based on training data.**

A picture containing table

Description automatically generated

**Accuracy = 0.9947**

**Some misclassification happens when classifying class 4 and class 5 for training data.**

6. Use a figure or a table to show your leaderboard performance. Describe your efforts to

improve the performance.

**Binary Classification, SID = l1281, accuracy = 1.000**

**(Captured on 04/07/2023)**

Table

Description automatically generated

**Multi-class Classification, SID = l1281, accuracy = 0.955**

**(Captured on 04/07/2023)**

Table

Description automatically generated

**My efforts: Data Cleaning, Exploratory Data Analysis, Variables Selection, Regression methods selection.**

**To improve performance: Have tried different feature selection methods; Have tried different machine learning algorithms on classifications like glm, lasso, ridge, and so on; Have tried different cross validation parameters; Have tried different tuning parameters and threshold for classifications in algorithms; Have tried different seeds and see how they influence the misclassification rate on different algorithms.**

7. Comment on your final results and potential ways to further improve the classification

accuracy.

**In General,**

1. **Lasso could automatically choose the variables, but these variables chosen might not be the best subset of variables. Thus, hand choosing and changing some parameters and the set of variables may return a better result.**
2. **Overfitting issues may happen since the lasso still keep a lot of training data.**

**More specifically,**

1. **Binary classification has accuracy equals to 1.000.**
2. **Multiclass classification has an accuracy of about 0.955. According to the performance of training data, misclassification happens when classifying class 4 and class 5. One way to improve the results is to combine class 4 and class 5 into one class, class 45. And then doing the classification trying the methods other than LASSO to classify class 45 into class 4 and class 5.**