ME315: Machine Learning in Practice
Final Project

London School of Economics & Political Science
LSE ID: 202391272

Jay Sakarvadia*

July 4, 2024

*I would like to express my deepest gratitude to Dr. Kostas Kalogeropoulos for delivering an outstanding lecture series, providing us with a comprehensive and focused introduction to machine learning. My sincere thanks also go to Dr. Phil Chan for guiding my subsection, where I gained hands-on experience in applying theory, building, testing, and validating models. Finally, I extend my appreciation to Dr. Francesca Panero for her invaluable instruction on neural networks and fairness in machine learning.

Contents

1 Introduction

For my final project, I explore the area of disease prognosis and diagnosis by performing exploratory data analysis on the SUPPORT2 dataset, which was developed to aid the creation and validation of a model capable of predicting survival and disease progression for seriously ill hospitalized adults. In this study, I cleaned, processed, and split the dataset. I then trained, validated, and tuned various classification and regression models to address the following two key empirical questions:

1) How well can we classify a patient's disease given vital signs and predictive markers?

2) Can we predict a patient's future functional disability levels within a time horizon of two months?

By systematically approaching these questions, I aim to uncover insights that could enhance predictive modeling in medical prognostics and diagnostics. Improving the accuracy of disease classification can help healthcare providers make earlier and more precise diagnoses, potentially reducing the incidence of prolonged, painful end-of-life processes.

2 Data

The dataset used in this study comprises records of 9,105 critically ill patients from five United States medical centers, collected over two periods: 1989-1991 and 1992-1994. The patients met specific inclusion and exclusion criteria for nine disease categories: acute respiratory failure, chronic obstructive pulmonary disease, congestive heart failure, liver disease, coma, colon cancer, lung cancer, multiple organ system failure with malignancy and multiple organ system failure with sepsis. The dataset includes sensitive information such as race, gender, income, and education level, and contains 9,105 instances and 44 features.

3 Cleaning & Preprocessing

The dataset contains missing values for several baseline physiological variables. Professor Frank Harrell, the creator of the dataset, recommended imputation values for missing data:

Baseline Variable	**Normal Fill-in Value**
Serum albumin (alb)	3.5
PaO2/FiO2 ratio (pafi)	333.3
Bilirubin (bili)	1.01
Creatinine (crea)	1.01
BUN	6.51
White blood count (wblc)	9 (thousands)
Urine output (urine)	2502

Table 1: Imputation Values for Baseline Physiological Variables

Despite performing data imputation, numerous instances still contained significant amounts of missing data. Specifically, the two target variables, 'dzgroup' (classification) and 'sfdm2' (regression), had multiple missing values. Additionally, many physiological variable columns also had missing values. To address this, I decided to drop all rows with any missing or incomplete data. Consequently, the dataset size was reduced to 549 instances, representing a 93.97% decrease from the original size. This substantial reduction raised concerns about potential biases, the representativeness of the dataset and the statistical power of the analysis.

For linear models such as Linear Regression and Logistic Regression, standard scaling of the data was employed. Standard scaling transforms features to have a mean of 0 and a standard deviation of 1, improving logistic and linear regression by speeding up convergence, enhancing model performance, ensuring numerical stability, and making coefficients more interpretable. Standard scaling was omitted for the other tree-based models as they don't require it.

4 Classification

In an effort to classify a patient's disease given predictive markers, I employed five different models to determine which performed best in classifying a patient's disease using 44 features. My target variable, 'dzgroup', was categorical and consisted of 8 different disease classes that a patient could be classified as.

After pre-processing, 'dzgroup' exhibited class imbalance which can be seen below.

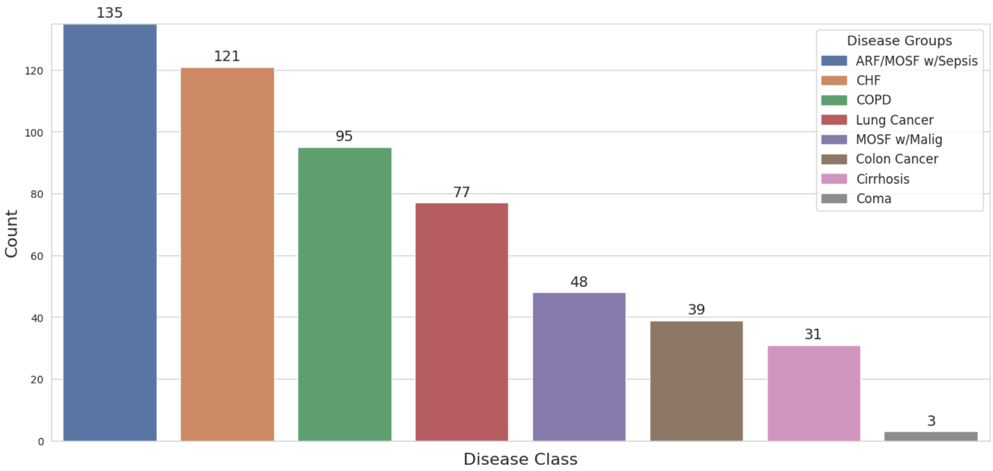

Figure 1: Count of Different Disease Classes

In an effort to remedy this class imbalance, I included class weight balancing in each of the models to ensure that all classes are given equal importance when training:

Table 2: Weight Balancing Methods for Different Models

Model	Weight Balancing Method
LogisticRegression	class_weight=balanced
RandomForestClassifier	class_weight=balanced
CatBoostClassifier	auto_class_weights=Balanced
XGBoostClassifier	scale_pos_weight=ratio
LGBMClassifier	class_weight=balanced

My classification models can be categorized into three groups, with some models belonging to multiple groups: tree-based models, gradient boosting models, and linear models. CatBoost, XGBoost, and LGBM are all gradient boosting models. They work by combining multiple weak decision tree learners into a single strong learner to improve predictive performance. RandomForest is a tree-based model that constructs multiple decision trees by fitting them on random subsets of features and data points. This approach enhances prediction accuracy and reduces overfitting. Logistic Regression, on the other hand, is a simple linear model that uses a sigmoid function to estimate the probability of a prediction belonging to a specific class.

Table 3: Classification Accuracy of Different Models

Model	Classification Accuracy
CatBoostClassifier	0.6364
RandomForestClassifier	0.6545
XGBoostClassifier	0.6818
LGBMClassifier	0.7000
LogisticRegression	0.7181

The best performing model was LogisticRegression.

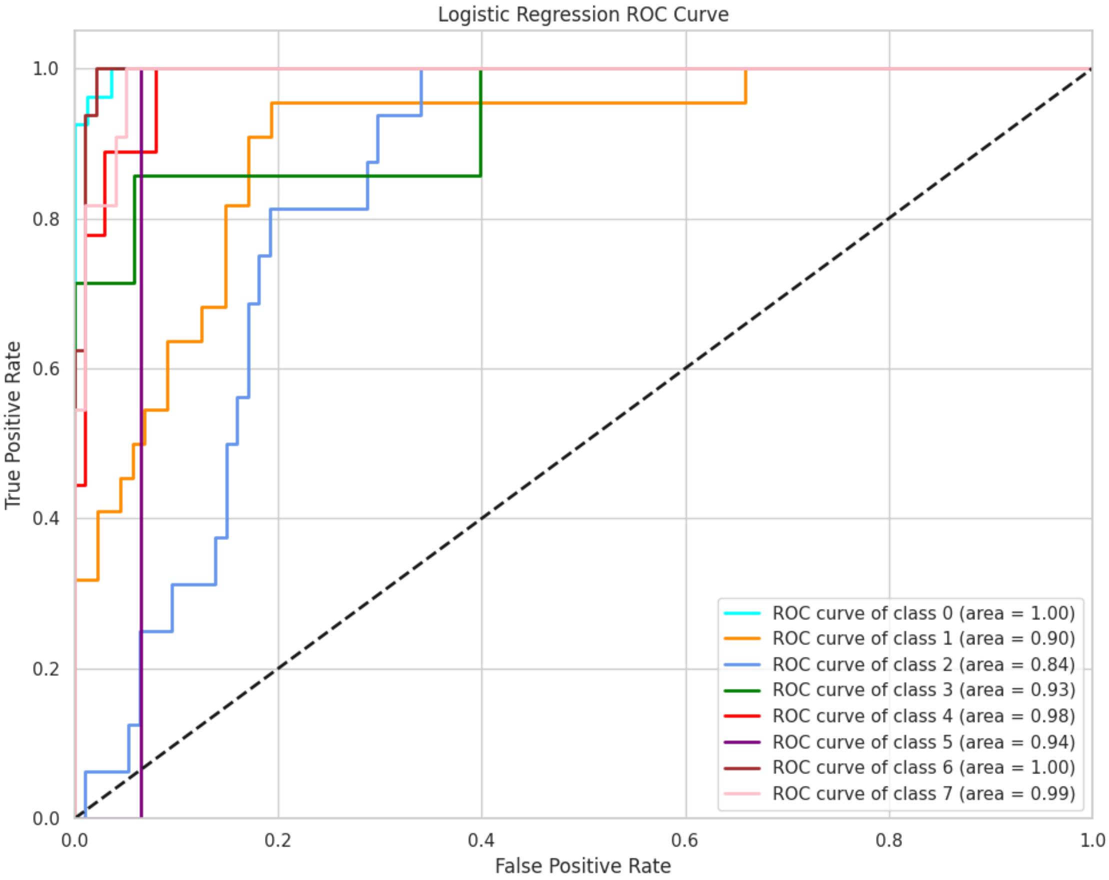

Figure 2: Logistic Regression ROC curve

The ROC curves for the logistic regression model show high AUC values for most classes, indicating strong performance in distinguishing between different classes. This demonstrates the model's good discriminatory ability. However, there is variation in AUC values among classes, suggesting that the model performs better for some classes than others. This variation could be due to factors like class imbalance and differences in feature importance.

Despite the high AUC values, the overall accuracy of the model is 71%, which does not fully align with the superior performance indicated by the ROC curves. This discrepancy can be attributed to several factors, including class imbalance, where the model might perform well on majority classes but struggle with minority classes, impacting overall accuracy. Additionally, the ROC curve evaluates performance across all thresholds, while accuracy depends on a specific threshold (usually 0.5), which might not be optimal for all classes. Furthermore, in multi-class classification, AUC values are often calculated using a one-vs-rest approach, potentially providing a more optimistic view of performance for individual classes.

Overall, while the ROC curves indicate that the logistic regression model has strong discriminatory power, addressing issues related to class imbalance and threshold optimization can help improve the overall accuracy and better reflect the model's performance.

5 Regression

With the goal of predicting a patient's future functional disability level, I set my target variable to be 'sfdm2', which consisted of values that measure a patient's functional disability level 2 months into the study, incorporating information from questionnaires and correlations with other features. Due to it being a continuous variable, I employed 5 different regression models to see which of them performed best when trying to predict a patient's functional disability level on a 1-4 scale. The values on the 1-4 scale help characterize the severity of functional disability, with 1 representing the lowest severity and 4 representing the highest:

Table 4: Functional Disability Indicators

Code	**Description**
1	No signs of moderate to severe functional disability from the interview.
2	Patient was unable to do 4 or more activities of daily living.
3	Sickness Impact Profile total score at 2 months is greater or equal to 30.
4	Patient intubated or in coma.

My regression models can also be categorized into three groups, with some models belonging to multiple groups: tree-based models, gradient boosting models, and linear models. CatBoost, XGBoost, and LGBM are all gradient boosting, tree-based models. They work by combining multiple weak decision tree learners into a single strong learner to improve predictive performance. RandomForest is a tree-based model that constructs multiple decision trees by fitting them on random subsets of features and data points. This approach enhances prediction accuracy and reduces overfitting. Linear Regression, on the other hand, is a simple linear model that estimates the relationship between the dependent variable and one or more independent variables by fitting a linear equation to observed data.

Table 5: Regression MSE of Different Models

Model	Regression MSE
RandomForestRegressor	0.6508
CatBoostRegressor	0.6526
LinearRegression	0.6847
LGBMRegressor	0.6968
XGBRegressor	0.7637

The best performing model was RandomForestRegressor.

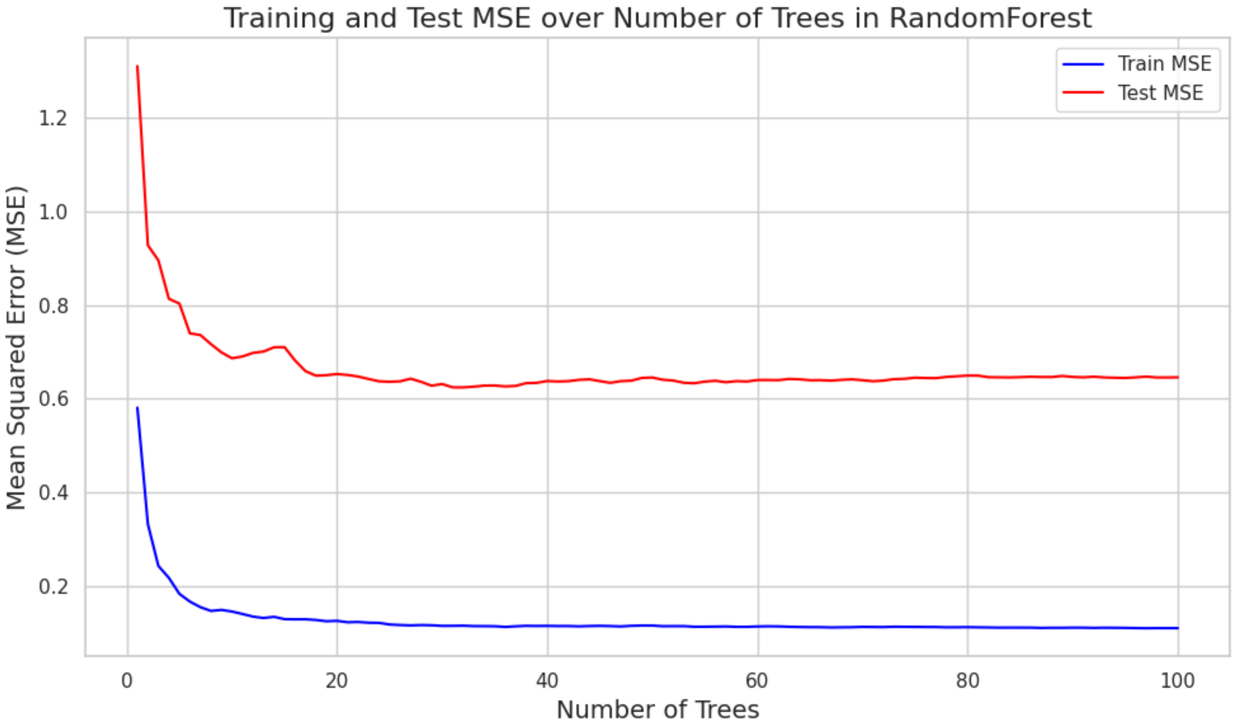

Figure 3: RandomForestRegressor Train/Test MSE

Both the training MSE (blue line) and test MSE (red line) decrease rapidly as the number of trees increases, indicating that the model is learning and improving its performance on both the training and test data. After around 20-30 trees, both the training and test MSE stabilize. This suggests that adding more trees beyond this point does not significantly enhance the model's performance on either the training or the test data.The test MSE is consistently higher than the training MSE, which is expected as the model typically performs better on the training data it has seen compared to the unseen test data. This gap indicates some degree of overfitting. Further tuning of other hyperparameters (e.g., max depth, min samples split) might help reduce the test MSE and mitigate overfitting. Applying regularization techniques like pruning or setting a maximum depth for the trees can help reduce overfitting.

6 Discussion

The primary challenge encountered during this project was the significant reduction in dataset size by 93.97%, resulting from the preprocessing steps. This drastic decrease in data points raised several concerns about the representativeness of the dataset and the potential biases it introduced. Notably, the reduction completely eliminated a fifth level for the regression problem, meaning that the scale should have been 1-5 instead of 1-4. During preprocessing, all rows with a value of 5 for their 'sfdm2' entry were deleted due to null values elsewhere in those rows. This is likely a major reason why the regression models exhibited high (poor) MSE values.

To address the issue of data reduction, several strategies should be considered in future work to preserve as much of the original dataset as possible. Although specific imputation values were applied for missing physiological data, future studies could benefit from advanced imputation methods, such as multiple imputation or model-based imputation, to better handle missing values without significantly reducing the dataset size. Other techniques like mean imputation or substitution could also be employed to retain more instances.

During this project, manual hyperparameter tuning was conducted for model training. However, due to the critical issue of dataset reduction, comprehensive hyperparameter tuning using grid search or random search was not pursued. Addressing the dataset size issue was prioritized over fine-tuning model parameters, as hyperparameter optimization alone would not resolve the underlying data insufficiency.

Future work should involve a more systematic approach to hyperparameter tuning once the data-related issues are mitigated. This could include automated optimization techniques, such as grid search or random search, to efficiently explore a wide range of parameter values and identify the optimal settings for each model.

7 Conclusion

Overall, the findings of this project underscore the critical need for effective data preprocessing and handling techniques in machine learning. By systematically addressing the challenges of missing data and dataset reduction, the predictive modeling of medical prognostics and diagnostics can be significantly improved. Enhancing the accuracy of disease classification and functional disability predictions enables healthcare providers to make earlier, more informed decisions, ultimately leading to better patient outcomes and potentially reducing the incidence of prolonged, painful end-of-life processes.

In conclusion, while the study successfully applied various classification and regression models to the SUPPORT2 dataset, the limitations posed by data reduction highlight areas for improvement in future research. By adopting more advanced data handling and model tuning techniques, the robustness and reliability of predictive models in medical applications can be further enhanced.