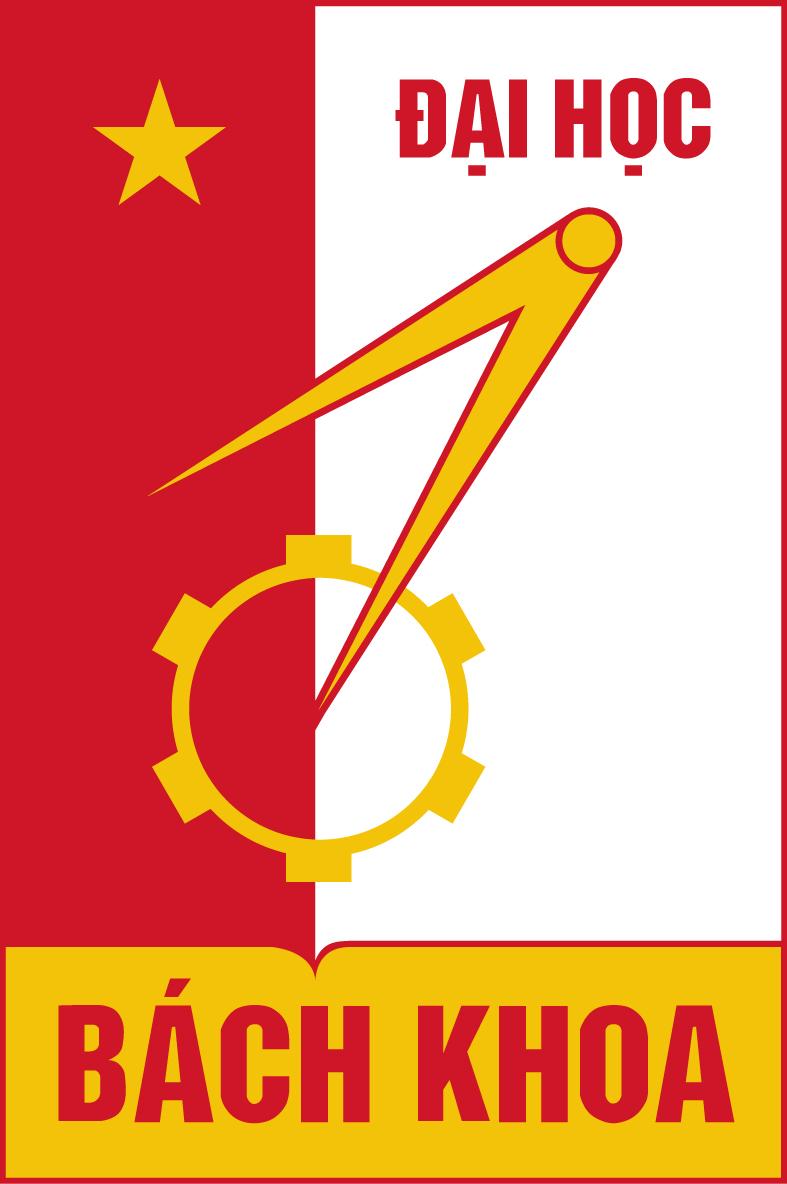
**Ha Noi University of Science and Technology**

**School of Information and Communication Technology**



**FINAL REPORT**

**"Prediction of patient's chance of intubation in ICU based on MIMIC-IV dataset"**

Instructor: Nguyen Hong Quang

Student:

Pham Anh Minh – 20194802

**Table of Contents**

Abstract…………………………………………………………………………..3

Dataset…………………………………………………………………………...3

##### EDA and Data Preprocessing………………………………………………… 3,4

Feature Selection Method………………………………………………………..4

##### Predictive Models…………………………………………………………… 4,5

Result Analysis…………………………………………………………………..5

##### Visualizations…………………………………………………………………….6

##### Tables…………………………………………………………………………….7

##### Conclusion……………………………………………………………………….8

##### Abstract

This project aims at using machine learning models to estimate the patient's chance of intubation in the Intensive Care Unit(ICU) based on the MIMIC IV(Medical Information Mart for Intensive Care IV) dataset. Before applying the machine learning models for the prediction, an exploratory data analysis is done to gain insights of the dataset. Then data preprocessing and feature selection methods are used to enhance the usability of the dataset. Three machine learning models are then trained for the prediction, predictive models include Logistic Regression, Decision Tree and Random Forest.

##### Dataset

The MIMIC(****M****edical ****I****nformation ****M****art for ****I****ntensive ****C****are IV) database is a vast repository of de-identified health-related data from more than 40,000 patients who were admitted to critical care units at the Beth Israel Deaconess Medical Center. This database is available free of charge and encompasses a broad range of information, including demographics, vital sign measurements taken at the bedside (approximately one data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality rates both inside and outside the hospital. For this project, we will use a pre-extracted dataset of the MIMIC-IV. It contains totally 36489 patient healthcare records with 60 indicators.

##### EDA and Data Preprocessing

For this part, I take several steps to look into the basic information of the dataset and applied several methods for data preprocessing.

These several steps include:

1. Check which features are include in the dataset.
2. Check the ratio of the True label to False label. (From this step, we noticed the dataset is **imbalanced**, therefore oversampling should be applied to the dataset to avoid overfitting)
3. Check the ratio of missing values in each column. (For column with over 80% missing values, we directly remove it from the dataset since it cannot provide much information for the models)
4. Calculate the correlation between each feature and draw the correlation heatmap. (Remove columns with over 95% correlations to reduce the number of features)
5. Use visualizations to check the distribution of each feature. (Most of the features are not in normal distribution, which means that it will be better to use median values to replace the missing values instead of using mean values)

##### Feature Selection Method

To further reduce the number of features used for the training of predictive models, I applied the **Genetic Algorithm** for the feature selection of each predictive model separately. Some important parameters set for the GA is: population\_size = 30, generations = 40. For each predictive model I recorded the 20 times accuracy score before and after feature selection to compare the performance. Besides, I also generated the confusion matrix of each model to make sure that the oversampling method works properly to help avoid overfitting.

##### Predictive Models

In this part, we train three predictive models for the prediction:

* Logistic Regression
* Decision Tree
* Random Forest

For each model, I replace the missing values with mean and median values respectively to create two dataset and train the model based on each dataset. Besides, since in the EDA part we noticed that the data is **imbalanced**, I also applied SMOTE(Synthetic Minority Oversampling Technique) to oversample the data to balance the number of True label and False label.

Since the dataset is imbalanced and I used SMOTE to oversample it, one potential problem is that some datapoints actually doesn’t exist in the originally dataset, which means that K-Fold validation is not applicable for the dataset because if the datapoints generated by SMOTE are split to testing dataset, then the testing dataset won’t be able to test the accuracy of the model. Therefore, in order to test the accuracy of the models, I trained and test each model for 20 times before and after feature selection and calculate the mean accuracy of the model to represent the overall accuracy of each model. Besides, I also calculated the standard deviation of the 20 times accuracy of each model to compare the stability of each model before and after feature selection.

##### Result Analysis

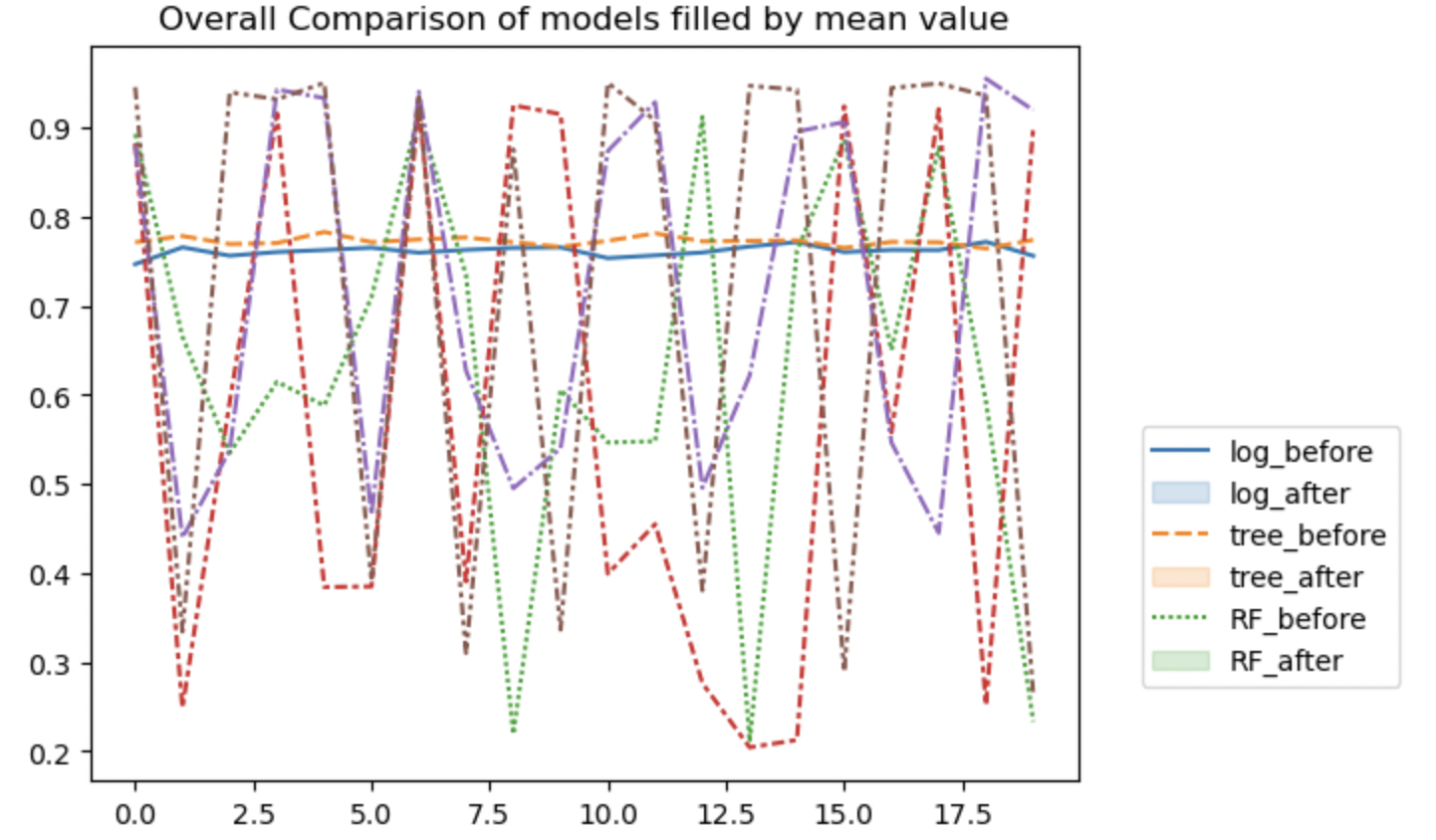
After training all the models separately and record the accuracy score of each model for 20 times, I draw visualizations of the to compare the performance of all the predictive models.

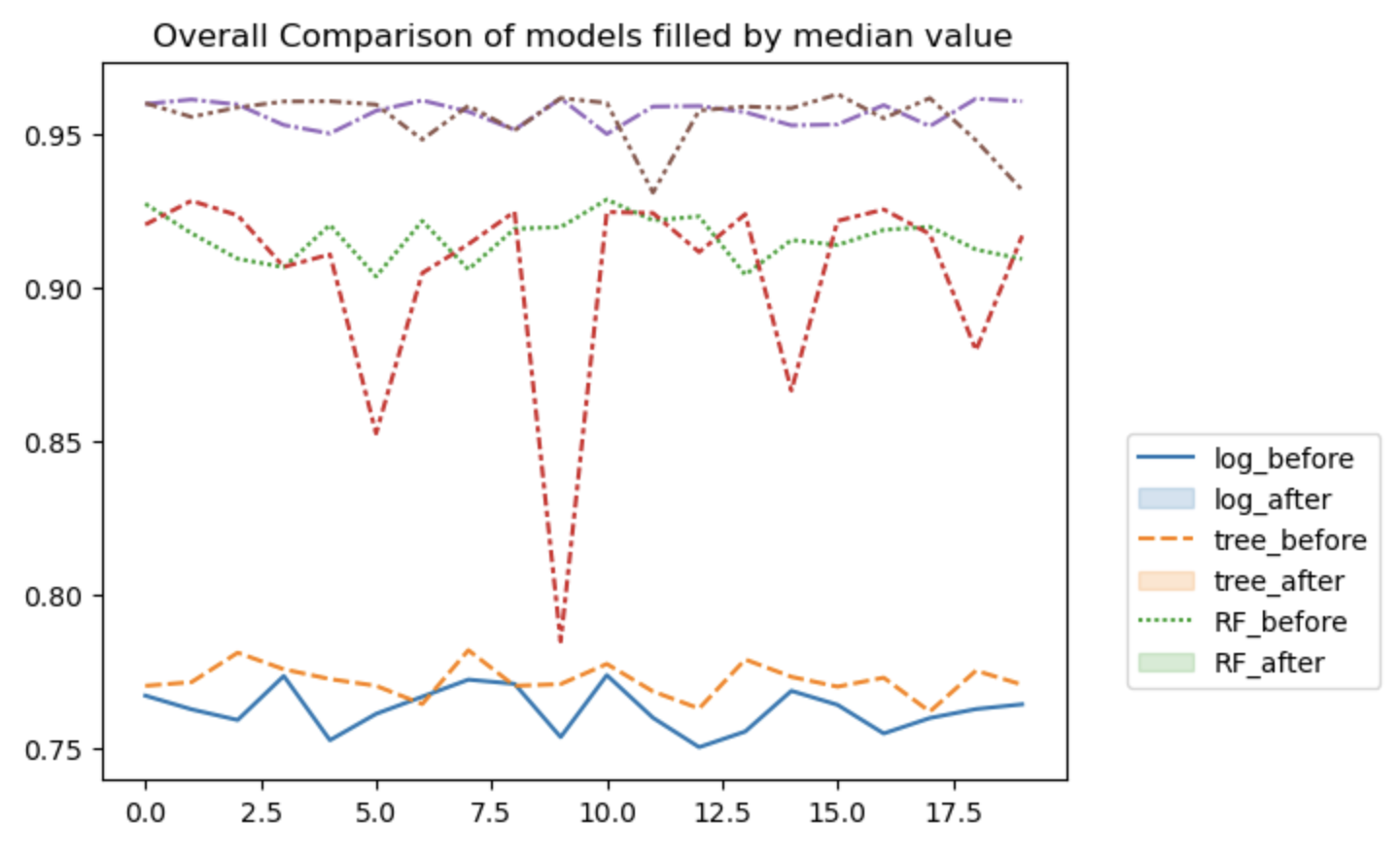
The visualizations includes:

1. The overall accuracy score of all the models.
2. The mean accuracy score comparison of all the models
3. The standard deviation of accuracy score of all the models.

Base on the overall accuracy score and mean accuracy score of all the models, we can come up with which is the best model for predicting the patient's chance of intubation in ICU. With the standard deviation of the accuracy score all the models, we can analyze which model has a better stablity when dealing with different data.

##### Visualizations





##### Tables

**Mean accuracy of the models filled the missing value with *mean* value:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | **Model** | **Mean Accuracy Before Feature Selection** | **Mean Accuracy After Feature Selection** | | Logistic Regression | 0.76 | 0.77 | | Decision Tree | 0.64 | 0.58 | | Random Forest | 0.72 | 0.72 | |  |

**Mean accuracy of the models filled the missing value with *median* value:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  |  |  | | --- | --- | --- | | **Model** | **Mean Accuracy Before Feature Selection** | **Mean Accuracy After Feature Selection** | | Logistic Regression | 0.76 | 0.77 | | Decision Tree | 0.92 | 0.90 | | Random Forest | 0.96 | 0.96 | |
|  |  |

##### Conclusion

Based on the mean accuracy and standard deviation of the predictive models, the best model for predicting the patient’s performance of intubation in ICU is the random forest model which filled the missing value with median value. The mean accuracy of its 20 times test is 0.96 and the standard deviation of it is 0.01. It achieves the best performance by only ask for 25 features after applying the genetic algorithm for feature selection.

By comparing the accuracy of all the models, it’s clear that the random forest can always achieve the best performance, followed by decision tree and then logistic regression. The main reason why logistic regression cannot achieve an accuracy more than 80% might be that this problem is not a linear separable problem, which means that it’s difficult to find a good decision boundary to separate the datapoints. Therefore, tree-based model will be more suitable for this problem because the models use information entropy for making the decisions. Besides, with ensemble learning, the random forest model can achieve a better stability than the decision tree model.

One phenomenon we noticed is that the accuracy of tree-based models filled missing value with mean value is extremely unstable. This is a not surprising phenomenon since in the EDA part, we notice that most of the features are not in normal distribution which means that the mean value cannot represent the population very well in this case. Therefore, there is no doubt that the models filled missing value with median value can achieve a better stability.