COSC2753 Machine Learning

Lecturer – Nguyen Thien Bao

RMIT UNIVERSITY

Assignment 1

ICU SEPSIS MONITOR ASSIGNMENT

Report by
Luong Nguyen | S3927460

Semester A - April 12th, 2023

Table of Contents

I. Problem Overview

The aim of this project is to develop a reliable model to aid the ICUs in their efford of monitoring patients during their stay to achieve the following:

- Early prediction of sepsis development in ICU patients can significantly reduce health complications and optimize ICU resources, including bed availability.
- The ability to predict sepsis in ICU patients has significant benefits for both patient outcomes and resource management.
- Sepsis monitoring is a crucial aspect of ICU care, and early intervention can save lives and improve outcomes.

II. Exploratory Data Analysys (EDA)

1. Introduction to our dataset.

The given dataset contains information of patients who have undergone treatment in ICU. There are 11 columns in the dataset, namely ID, PRG, PL, PR, SK, TS, M11, BD2, Age, Insurance, and Sepssis.

2. Data Cleaning and Preprocessing

Before conducting any analysis, it is essential to ensure that the data is clean and free from errors. The following steps have been taken to ensure data quality:

- Checked for missing values in each column.
- Encode our data value for **Possitive** as **1** and **Negative** as **0**
- Checked for invalid values or outliers.
- Checked for duplicated rows.

As per the requirement, the ID and Insurance columns would be eliminated as we are confident that those features have no bearing on whether a patient receives the Sepsis column.

3. Visualizing Data – Histogram (Appendix A)

Some of the features, including PR, PL, and M11, follow a normal distribution but exhibit unusual data with a value of 0. Other features such as PRG, SK, TS, BD2, and Age are skewed and may cause our model to underperform in scenarios with higher values of these features. Additionally, the Sepsis cases are imbalanced, with a 1:2 ratio of positive to negative cases. This may lead to a biased model, so data balancing techniques such as oversampling or undersampling may be necessary to ensure balance between the two cases

4. Visualizing Data – Scatter Plot (Appendix B)

From the out plotted data, there are no convincing signs of a link between the value of characteristics and whether a patient develops Sepsis.

Therefore, we must take into account the potential of outliers showing in the scatter plot. As a consequence, the Box plot was created, which is used to represent the value distribution of each characteristic.

5. Visualizing Data – Box Plot (Appendix C)

Based on the box plot analysis, there are noticeable differences in the median, first quartile, and third quartile values of various variables between Sepsis-positive (1) and Sepsis-negative (0) patients. Specifically, the box plots reveal that the majority values of PL for Sepsis-positive and Sepsis-negative patients do not overlap,

indicating that PL may play a significant role in predicting Sepsis. Similarly, the median and third quartile values of PRG, Age, and M11 are considerably different between the two groups, suggesting that these variables may also be useful in predicting Sepsis.

However, several variables such as BD2, TS, and Age have a high number of outliers that need to be taken into account when making predictions. Moreover, there is no clear decision boundary in any of the variables, indicating that using a single feature for prediction is unlikely to be effective.

6. Visualizing Data – Pair Plot (Appendix D)

Based on the pair plot analysis, it appears that there is no clear separation between patients who are positive and negative for Sepsis using any single or pair of features. However, there is less overlap between positive and negative cases in pairs such as (M11-PL), (M11-PRG), (M11-PR), (SK-PL), (PL-PR), and (PL-Age). This suggests that using multiple features may lead to better performance. Nonetheless, it is worth noting that some features such as BD2, TS, and Age have a high number of outliers that need to be addressed in the analysis. Overall, it is likely that predicting Sepsis using any single feature would not be effective and a combination of features needs to be considered.

7. Visualizing Data – Correlation Matriz (Appendix E)

The dataset shows low interdependence between its characteristics, as indicated by the correlation graph where the highest correlation is only 0.53 between Age and PRG. PL (1) is the most influential characteristic in predicting Sepsis. Other characteristics have a moderate association with Sepsis, except for PR and SK, which do not show any significant relationship.

8. Conclusion

In conclusion, our EDA indicates that predicting Sepsis using any single feature is unlikely to be effective. Instead, a combination of features needs to be considered. PL is the most influential characteristic in predicting Sepsis, while other characteristics have a moderate association, except for PR and SK, which do not show any significant relationship. Outliers in several variables should be addressed, and data balancing techniques may be necessary due to imbalanced Sepsis cases.

III. Approaches

1. Choosing a Machine Language Model

Our objective is to predict the occurrence of Sepsis in patients, which is a classification problem. To achieve this, we will evaluate five supervised machine learning models, including Linear Regression, Polynomial Regression, Logistic Regression, Decision Tree Classifier, and Random Forest.

Linear Regression uses a first-degree function for predicting continuous variables. On the other hand, Polynomial Regression uses an nth-degree function for the same purpose. Logistic Regression is used for predicting discrete variables, and it employs a sigmoid function to produce output values ranging from 0 to 1. Decision Tree Classifier employs a tree structure of multiple nodes to classify data into different branches, and Random Forest combines multiple trees for classification.

In our EDA, we concluded that none of the features have a clear decision boundary for predicting Sepsis. Even with pair plot analysis, there is no visible separation between positive and negative patients. Since most features do not have a clear separation, the performance of Logistic Regression and Decision Tree Classifier/Random

Forest is expected to be similar. Therefore, we will choose the better performing model between Logistic Regression and Random Forest as the final model..

2. Choosing Input Features

One crucial decision we need to make is whether to use a **univariate** or **multivariate** approach for the analysis. From our EDA in the previous section, we have found that PL has the highest correlation with the occurrence of Sepsis. Therefore, it will be included as one of the features for **Logistic Regression**, regardless of the approach we take. However, for **Random Forest**, due to the nature of a decision tree where multiple features are required to improve classification accuracy, we will not test the univariate approach.

Apart from that, we will also conduct a test with all 11 features to determine which features have the most significant impact on the classification, by analyzing the coefficients for **Logistic Regression** and feature importance for **Random Forest**. In conclusion, we will evaluate two different input scenarios to ensure that the model we select is the most effective:

- A Univariate model with only PL as the input attribute for Logistic Regression.
- A Multivariate model with all 11 features as input attributes for both Logistic Regression and Random Forest

3. Model evaluation methods

There are several methods to evaluate the performance of the models. For Logistic Regression, a loss function called Log loss is used to measure the accuracy of the model's prediction compared to the expected output. In addition, since this is a classification problem, other metrics such as Recall**,** Precision, F1 score**,** and ROC AUC can also be used to evaluate the model's performance from different perspectives (how well it classifies each class). A Confusion matrix is also a useful tool to visualize how the data is classified.

Given the issue statement's goal of saving costs and preventing health consequences, reducing the number of False Negatives (patients with Sepsis but categorized as Negative) is critical. As a result, Recall for Positive and Precision for Negative will be prioritized or maximized. The F1 score is also vital in ensuring that the two needs are balanced. Lastly, for model validation, K-fold cross-validation will be utilized, which entails using various sections of the data for validation in each fold.

4. Constrains

Logistic Regression

Logistic Regression tends to overfit by maximizing weights to get the output as close to 0 or 1 as possible, as discussed on Stack Exchange [1]. To prevent this, regularization methods are applied to restrict the weights and avoid overfitting. There are three regularization methods for Logistic Regression [2][3]:

- L1 (Lasso Regression) - some weights may become zero, resulting in feature selection. It adds the sum of absolute values of weights to the loss function and is robust against outliers[3].
- L2 (Ridge Regression) - all weights are non-zero, but some are very small. It adds the sum of squared values of weights to the loss function, with higher penalty for outliers[3].
- Early stopping, which stops the training process early to prevent overfitting.

However, early stopping regularization approach is not suitable for this project, as there is no clear indication of when to stop training, and it is a risky approach that can greatly affect the model output if the training is not enough.

Based on the "Choosing Input" section, we have decided to try out two different input features. Therefore, we will apply different regularization methods as follows:

- Univariate with PL: Since there is no need for feature selection, we will apply L2 to avoid excluding the only feature.

- Multivariate with all features: In this case, both L1 and L2 will be tested and compared since our dataset has features that are suitable for both regularization methods. Features with many outliers such as TS and BD2 might perform well with L1. However, since the dataset has few features, feature selection from L1 might affect the result.

Pruning Methods for Decision Tree Classifier/Random Forest

There are two main approaches to pruning decision trees: post-pruning and pre-pruning. Sklearn provides the Minimal Cost-Complexity Pruning method for post-pruning, which involves tuning the ccp_alpha parameter to choose the best level of pruning [4]. For pre-pruning, Sklearn offers several tunable parameters, including min_impurity_decrease, min_samples_split, and min_samples_leaf. These parameters determine the minimum impurity required for a node to be split [5].

IV. Hyperparameter Tuning

We decided in the previous part that Random Forest and L1, Multivariate, Logistic Regression produce the best results among the models evaluated, thus these are the models we would utilize and aim to enhance by tuning.

For the Logistic Regression model, we tuned the following parameters (definitions from sklearn documentation[8]):

1. C: Inverse of regularization strength, meaning that the smaller the value, the more penalized the model is.
2. max_iter: The maximum number of iterations taken for solvers to converge.

For the Random Forest model, we tuned the following parameters (definitions from sklearn documentation [9]):

1. ccp_alpha: Complexity parameter used for Minimal Cost-Complexity Pruning.
2. min_impurity_decrease: A node will be split if this split induces a decrease of the impurity greater than or equal to this value.
3. max_depth: Maximum depth of each tree.
4. min_samples_split: The minimum number of samples needed to split a node.
5. min_samples_leaf: The minimum number of samples required to be at a leaf node.

The tuning method employed in this case is Grid Search, which is an exhaustive algorithm. According to Gianluca Malato's paper, Grid Search tests every possible combination of parameter values and returns the best-performing combination [7]. If running Grid Search on all parameters simultaneously is computationally expensive, multiple Grid Searches will be performed on subsets of parameters. When the recommended values from Grid Search exhibit significant variations or do not enhance the model's performance compared to the default model, the range of tested parameters will be adjusted.

V. Ultimate Judgement and Analysis

1. Model Selection

The chart Below provide metric on both Final Models.

Metric	Logistic Regression	Random Forest
Avg Accuracy	75.32%	80.17%
Avg ROC AUC	0.7529	0.8026
Positive F1 Score	0.7475	0.8041
Negative F1 Score	0.7533	0.7952
Positive Recall	0.7392	0.8269
Negative Recall	0.7666	0.7782
Positive Precision	0.7604	0.7866
Negative Precision	0.7445	0.8168

Based on this comparison, several conclusions can be drawn:

1. Random Forest outperforms Logistic Regression on all metrics according to the cross validation results.
2. Random Forest achieves higher accuracy, especially in identifying positive cases as seen in the higher positive recall (sensitivity) and F1 score. This could minimize false negatives, though the higher false positive rate (lower precision) may result in some unnecessary use of resources.
3. While both models have reasonably good performance, there is still room for improvement by optimizing modeling techniques and hyperparameters to increase predictive power, especially for the negative class.
4. Additional metrics like log loss would also strengthen the comparison, providing insights into the accuracy of predicted probability estimates from each model. Lower log loss is preferred.

In summary, while both Logistic Regression and Random Forest are viable techniques for predicting Sepsis risk according to the cross validation results, Random Forest demonstrates some advantages in identifying positive cases with higher sensitivity. However, its higher false positive rate may lead to overestimation of risk for some patients. Further testing and optimization are needed to maximize the performance of the optimal model on new data for real-world application. This additional work would allow selection of the definitively superior approach between the models evaluated here.

2. Advantages

The final Random Forest model achieved a balance between Positive and Negative scores across all metrics. The precision for Negative and recall for Positive values, which were being maximized, were both over 80%, indicating that the model was not overfitting and could be used for unseen cases

The low False Negative values (high recall) and False Positive values (high precision) indicated that the Random Forest model was avoiding wrongly classifying a patient with Sepsis and not wasting resources, respectively.

3. Disadvantages

The Random Forest model may not predict well with cases with higher Age, Plasma Glucose (PRG), Blood Work Result-2 (SK), Blood Work Result-3 (TS), and Blood Work Result-4 (BD2) due to several positively skewed features.

4. Future Improvements

To rigorously determine the optimal model for clinical use, test set evaluation is essential, ideally using data from external healthcare organizations to assess generalizability. Oversampling methods could address class imbalance and improve specificity/PPV. Feature selection may reduce noise from less meaningful variables.

Advancing to neural networks or deep learning may also capture complex nonlinear relationships in the data that boost performance. However, these techniques require larger datasets for optimal results.

In summary, while Random Forest shows promise for predicting Sepsis risk on internal validation compared to Logistic Regression, test set evaluation using new data is still needed to conclusively select the single best method for implementation based on medical and operational priorities. Future work should thus focus on continued

optimization, testing and comparison between alternative models to develop the solution with the highest clinical utility for improving patient care and resource management.

References

[1] “Why is logistic regression particularly prone to overfitting in high dimensions?,” Cross Validated, Jun. 01, 2020. https://stats.stackexchange.com/questions/469799/why-is-logistic-regression-particularly-prone-to-overfitting-in-high-dimensions (accessed Apr. 12, 2023).

[2] K. Melcher, “Understanding Regularization for Logistic Regression | KNIME,” KNIME. https://www.knime.com/blog/regularization-for-logistic-regression-l1-l2-gauss-or-laplace (accessed Apr. 12, 2023).

[3] K. Pykes, “Fighting Overfitting With L1 or L2 Regularization: Which One Is Better? - neptune.ai,” neptune.ai, Jul. 22, 2022. https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization (accessed Apr. 12, 2023).

[4] “Post pruning decision trees with cost complexity pruning,” scikit-learn. https://scikit-learn/stable/auto_examples/tree/plot_cost_complexity_pruning.html (accessed Apr. 12, 2023).

[5] “Random Forest Hyperparameter Tuning in Python - GeeksforGeeks,” GeeksforGeeks, Dec. 28, 2022. https://www.geeksforgeeks.org/random-forest-hyperparameter-tuning-in-python/ (accessed Apr. 12, 2023).

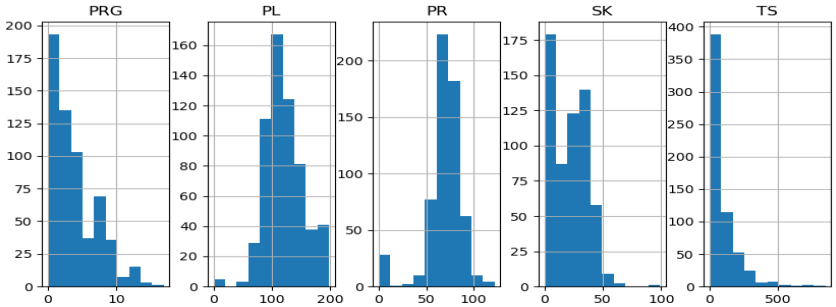
[6]“2. Over-sampling — Version 0.10.1,” 2. Over-sampling — Version 0.10.1. https://imbalanced-learn.org/stable/over_sampling.html# (accessed Apr. 12, 2023).

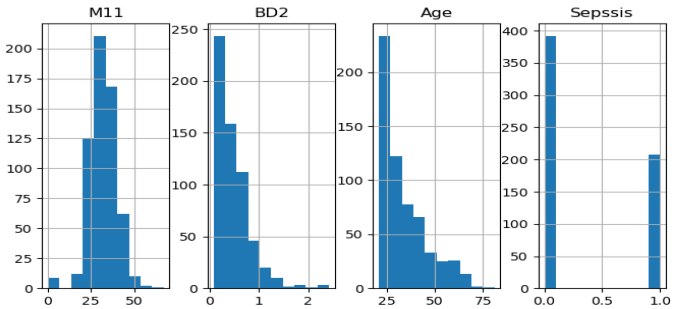
[7] G. Malato, “Hyperparameter tuning. Grid search and random search | Your Data Teacher,” Your Data Teacher, May 19, 2021. https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/ (accessed Apr. 11, 2023).

[8] “sklearn.linear_model.LogisticRegression,” scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Apr. 11, 2023).

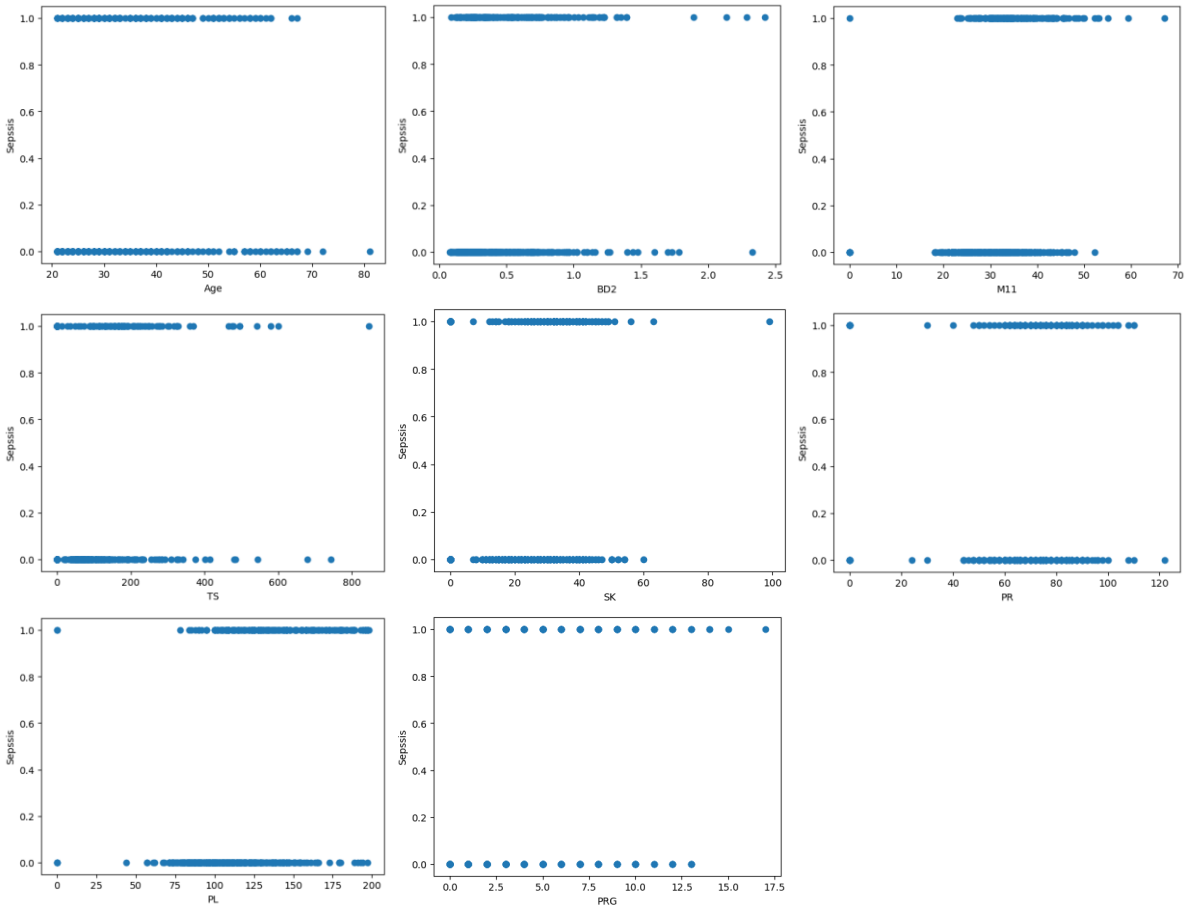
[9] “sklearn.ensemble.RandomForestClassifier,” scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Apr. 11, 2023).

Appendix A

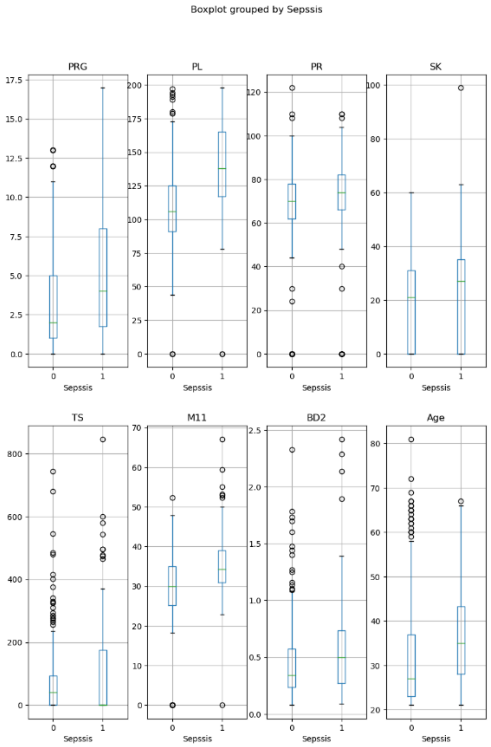




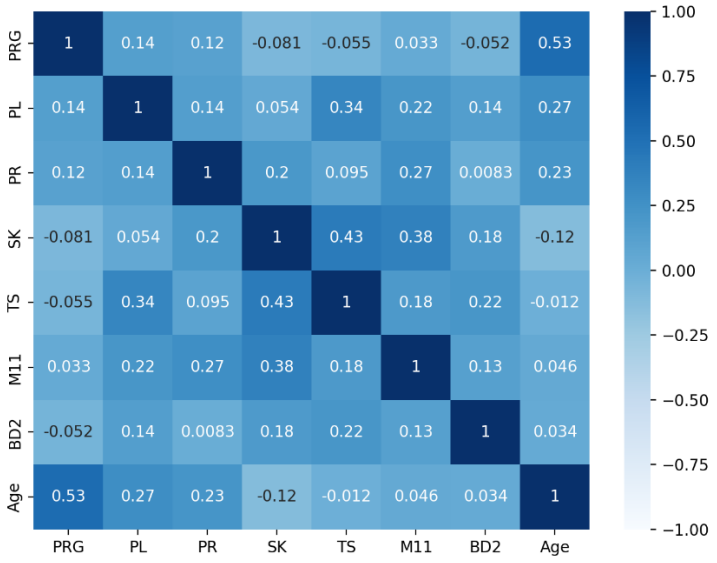
Appendix B



Appendix C



Appendix E



Appendix D