Group 9: Emilien Daulhiac, Lanxi (Nancy) Du, Yuhao (Bruce) Wang

SEP 787 Machine Learning: Classification Models
Project Proposal

1. **Problem Statement:**

Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death. In the U.S., 1.7 million of population develop sepsis and 270000 people die from it each year (Reyna et al., 2019). According to the statistic record, one third of patients who die in hospital have sepsis (Reyna et al., 2019). The early prediction of sepsis against severe injury in hospital can be potentially lifesaving and predicting sepsis for non-sepsis patients and early sepsis patients consume limited resource. Due to the several advantages of early sepsis treatment, the topic of this project is to predict if the patient has sepsis according to their physiological data.

2. **Solution Approach:**

In this analysis, we have chosen the Random Forest classifier as our model. Random Forest Algorithms is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes for classification or mean prediction of the individual trees for regression.

There are many reasons why we choose to use Random Forest method to deal with this problem instead of other methods such as linear regression.

1. Handling of Large Data: Randon Forest is capable of efficiently handling large datasets with high dimensionality. Our dataset is medical data which contains a substantial number of rows and several features, falls into this category.
2. Robustness of Overfitting: Random Forest reduce the risk of overfitting which is a frequent problem with decision trees by creating a number of decision trees and making final decision based on the majority vote of individual trees.
3. Handling Mixed Data Type: Our dataset is clinical data which contains both numerical and categorical features. Random Forest is an ideal choice for this problem to process data smoothly.
4. Handling Non-Linearity: Medical data often contains non-linear relationships. Random Forest can capture the non-linear relationships effectively due to its own non-linearity property.

It is worthy noticing that while random forest is a strong candidate for this problem because of the listed advantages, other model may perform better on this task. As the result, other algorithms such as Ada Boost and Gradient Boost will also be considered to implement if they provide better results.

3. **Methodology:**

In this project aimed at predicting sepsis based on physiological data, various methods will be employed to enhance the accuracy and efficiency of the prediction model. Given the complex nature of sepsis and its life-threatening implications, decision trees emerge as a particularly suitable approach. Decision trees are versatile tools that can effectively handle classification tasks, making them ideal for predicting whether a patient is at risk of developing sepsis based on their physiological parameters. The inherent interpretability of decision trees also allows medical professionals to comprehend and trust the decision-making process, crucial for a healthcare context. Among the decision tree algorithms, the Random Forest algorithm stands out as a promising choice for this specific problem. Random Forest

combines multiple decision trees, which helps mitigate overfitting and enhances the robustness of the model.

Recognizing sepsis as a critical, life-threatening condition, the choice of Random Forest over boosting techniques like XGBoost and AdaBoost is rooted in several advantageous features. Firstly, Random Forest demonstrates a heightened robustness to overfitting, a crucial characteristic in healthcare applications where model generalizability is paramount. The ensemble nature of Random Forest, which aggregates multiple decision trees, aids in mitigating overfitting and fortifying model stability. Additionally, Random Forest excels in capturing complex non-linear relationships inherent in physiological data, making it well-suited for this predictive modeling task.

The following steps will be followed to prepare our model:
Exploratory Data Analysis: Apply descriptive statistics, complete data visualizations, conduct correlation analysis for potential multicollinearity, and identify outliers.
Data Preprocessing: Clean the data, handle missing values, perform feature scaling, complete undersampling to address class imbalance, and splitting the dataset into training and testing sets.
Feature Engineering: Encoding categorical variables and creating new features that could enhance the model's predictive power (e.g., calculating the change in vital signs over a period of time),

4. **References:**

Deng, H. F., Sun, M. W., Wang, Y., Zeng, J., Yuan, T., Li, T., Li, D. H., Chen, W., Zhou, P., Wang, Q., & Jiang, H. (2021). Evaluating machine learning models for sepsis prediction: A systematic review of methodologies. iScience, 25(1), 103651. https://doi.org/10.1016/j.isci.2021.103651

He, Z., Du, L., Zhang, P., Zhao, R., Chen, X., & Fang, Z. (2020). Early Sepsis Prediction Using Ensemble Learning With Deep Features and Artificial Features Extracted From Clinical Electronic Health Records. *Critical Care Medicine*, *48*(12). https://journals.lww.com/ccmjournal/fulltext/2020/12000/early_sepsis_prediction_using_ensemble_learning.68.aspx

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in python. Springer Texts in Statistics. Retrieved November 30, 2023, from https://www.statlearning.com/

Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. D. (2019). Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019 (version 1.0.0). PhysioNet. https://doi.org/10.13026/v64v-d857.

Scikit-learn. (n.d.). Ensembles: Gradient boosting, random forests, bagging, voting, stacking. Retrieved November 30, 2023, from https://scikit-learn.org/stable/modules/ensemble.html#ensembles-gradient-boosting-random-forests-bagging-voting-stacking

Tang, T. (2023). Class Imbalance Strategies — A Visual Guide with Code. Towards Data Science. Retrieved November 30, 2023, from https://towardsdatascience.com/class-imbalance-strategies-a-visual-guide-with-code-8bc8fae71e1a

Zabihi, M., Kiranyaz, S., & Gabbouj, M. (2019). Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models. *2019 Computing in Cardiology (CinC)*, Page 1-Page 4. https://doi.org/10.22489/CinC.2019.238

Group 9: Emilien Daulhiac, Lanxi (Nancy) Du, Yuhao (Bruce) Wang

5. **Dataset Explanation:**

Dataset Source: Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Moody, B., Westover, M. B., Sharma, A., Nemati, S., & Clifford, G. D. (2019). Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019 (version 1.0.0). PhysioNet. https://doi.org/10.13026/v64v-d857

The dataset comes from the PhysioNet/Computing in Cardiology Challenge 2019, an international competition around developing open-source machine learning solutions for complex physiologic signal processing and medical classification problems (Reyna et al., 2019).

Data was sourced from the electronic medical records (EMRs) of two U.S. hospital systems: Beth Israel Deaconess Medical Center and Emory University Hospital. The dataset includes observations from **40,336** ICU patients collected over the period of 2009 to 2019. The following features are included:

- static patient demographic variables (6),
- hourly vital sign summaries (8) and
- laboratory values (26) for each day (24-hours) of the patient's ICU stay.

All patient features were condensed into hourly bins (e.g., multiple heart rate measurements in a 1-hour time window were summarized as the median heart rate measurement of the hour). The following criteria were used to exclude patients from the dataset:

- Patients with less than 8 hours of data in the ICU
- Patients who developed sepsis less than 4 hours after ICU admission
- Patient records after discharge from the ICU
- Patient records after 2 weeks from admission to ICU (i.e., patient records were limited to 2 weeks of hourly data)

Only **2,932 (7.3%)** patients from the dataset are labeled as having developed sepsis, indicating an imbalanced dataset. Our group will complete the following pre-processing steps to prepare the data for Decision-Tree based modelling.

1. **Build Dataset:** Combine individual patient record files into one table, aligning based on hourly data. A pre-combined dataset may be used: https://www.kaggle.com/datasets/salikhussaini49/prediction-of-sepsis/data
2. **Feature Exploration:** Use data visualization and preliminary Ada boost to identify most important features.
3. **Data Transformation:**
    a. **Addressing Missing Values:** The dataset contains a large number of missing values to reflect the nature of data available in hospital EMRs. Different types of features will have different approaches:
        i. Patient Demographics: patients with missing values may be dropped
        ii. Redundant vital sign and laboratory features may be dropped
        iii. Important vital sign and laboratory value features will be imputed on a patient-by-patient basis. Globally applying the median value or using similar imputation techniques may result in inaccuracies as patients may have different baseline data.
    b. Features Scaling: Continuous variables will be scaled.
    c. Encoding: Categorical variables will be encoded using one-hot encoding.
    d. Addressing Class Imbalance: Undersampling will be used to balance the class label.