Sepsis Survival Minimal Clinical Records

Prepared By: INFO\_3142\_Group\_9\_Project\_1 (Nick Goudsbloem, Noor Al-Najar)

2023-11-03

<https://www.kaggle.com/datasets/joebeachcapital/sepsis-survival-minimal-clinical-records>

Problem Description

Sepsis is a type of infection that occurs when the body fails to treat infection properly. On top of this sepsis left unchecked it can and will be lethal. While sepsis can be cured it has to be done in quick time frame. The issue is that there are numerous factors that can decrease the chance of survival and increase the chances of survival.

Objective of Dataset

The objective of this dataset is to quantify the death toll after 9 days with a basic amount of data. The purpose is to predict the survival of patients within minutes. One of the restrictions put in place is the limited data which is done via the patients age, sex, and the number of previous sepsis episodes. The reason for this is that most hospitals won’t have time to get all patient documents. By having this restriction most hospitals would be able to calculate the priority of patients with minimal data.

Dataset Values

|  |  |  |
| --- | --- | --- |
| Attribute Name | Type | Description |
| age\_years | Integer | This value represents the age of the patient. |
| episode\_number | Integer | This value represents the number of times a patient has contracted sepsis. |
| sex\_0male\_1female | Boolean (0, 1) | 0 represents a male while 1 represents a female. |
| hospital\_outcome\_1alive\_0dead | Boolean (0, 1) | 0 represents a death while 1 represents alive. |

<Selected Algorithms (required 3)>

<opt 1>

**DecisionTree Regression**

<Describe training method>

The selected forecasting problem in this script revolves around predicting hospital outcomes based on patient attributes, such as age, sex, and episode number, with the goal of assessing the likelihood of an individual being alive or deceased immediately and after 9 days, given specific inputs. The dataset used in this context is named "sepsis\_survival\_primary\_cohort.csv." It includes attributes like age in years, gender (0 for male, 1 for female), episode numbers, and the binary target variable representing hospital outcomes (1 for alive, 0 for dead).

The selected algorithms for this forecasting problem involve a Decision Tree Regressor, which is a supervised machine learning technique used to make predictions. The algorithm uses a decision tree structure to recursively split the data into subsets based on feature values, making it suitable for regression tasks.

The Decision Tree Regressor in the script is trained on the training data using the `fit` method, and its performance is evaluated using Mean Squared Error (MSE) and R-squared on the test data. The predictions are made both immediately and after 9 days based on user input, and the results are interpreted by classifying individuals as "alive" or "dead" based on a threshold of 0.5. Additionally, the script calculates odds of dying under certain conditions and provides insights into the relationship between age and the number of deaths using visualizations. The decision tree model's structure is visualized to gain insights into the decision-making process. The evaluation procedure includes assessing the model's predictive performance, counting the number of deaths by episode, and visualizing age-related death counts.

<opt 2>

**RandomForestRegressor**

<Describe training method>

The script utilizes a Random Forest Regressor to address a forecasting problem associated with hospital outcomes. The goal is to predict the likelihood of an individual's survival (alive) or death within a medical context based on specific attributes, including age, gender, and episode number. The dataset, sourced from "sepsis\_survival\_primary\_cohort.csv," contains valuable attributes essential for making predictions. The selected features encompass age in years, gender (encoded as 0 for male and 1 for female), and episode numbers. These attributes play a crucial role in determining the ultimate hospital outcome, categorized as 1 for alive and 0 for deceased.

The chosen algorithm for this task is the Random Forest Regressor, a powerful ensemble learning method that leverages a collection of decision trees. In this script, a forest of 100 decision trees is employed, and the model is trained on a split of the dataset into training and testing sets. The user's input, specifying age, gender, episode number, and untreated days, is used to make immediate and 9-day predictions regarding hospital outcomes. The script interprets the predictions based on a threshold of 0.5, classifying individuals as "alive" or "dead." Additionally, it calculates the odds of death for both immediate and 9-day scenarios, considering an average survival rate of 30%.

To evaluate the model's performance, the script calculates Mean Squared Error (MSE) and R-squared on the test data, offering insights into the model's accuracy and predictive capabilities. It also quantifies and reports the number of individuals predicted as alive or dead within the test dataset. Furthermore, the script provides a detailed analysis of the distribution of deaths based on the episode count, both for the test and training data. It visualizes the relationship between age and the number of deaths, offering a comprehensive overview of the forecasting problem and model performance.

<opt 3>

<Describe training method>

<Accuracy comparison>