Table of Contents

[Handling and Preprocessing Data 2](#_Toc4573)

[Multivariate Linear Regression 4](#_Toc4574)

[Logistic Regression 5](#_Toc4575)

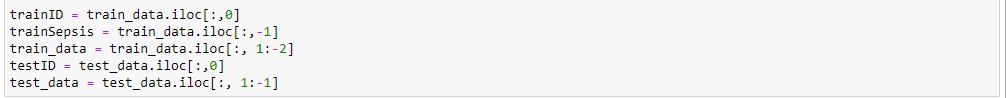
[Decision Tree 6](#_Toc4576)

[My Judgement 7](#_Toc4577)

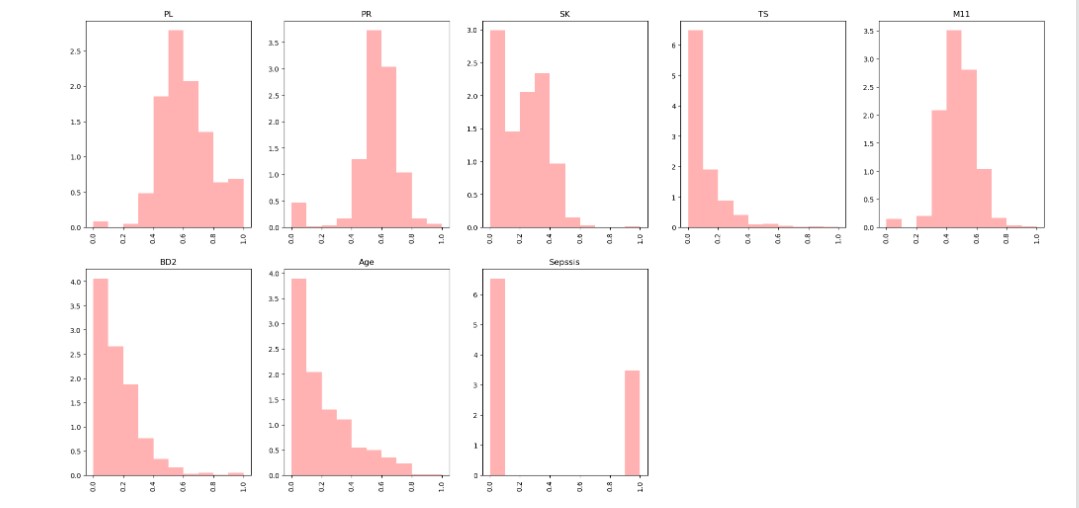
[Exporting Predicting Dataset to CSV 7](#_Toc4578)

# Handling and Preprocessing Data

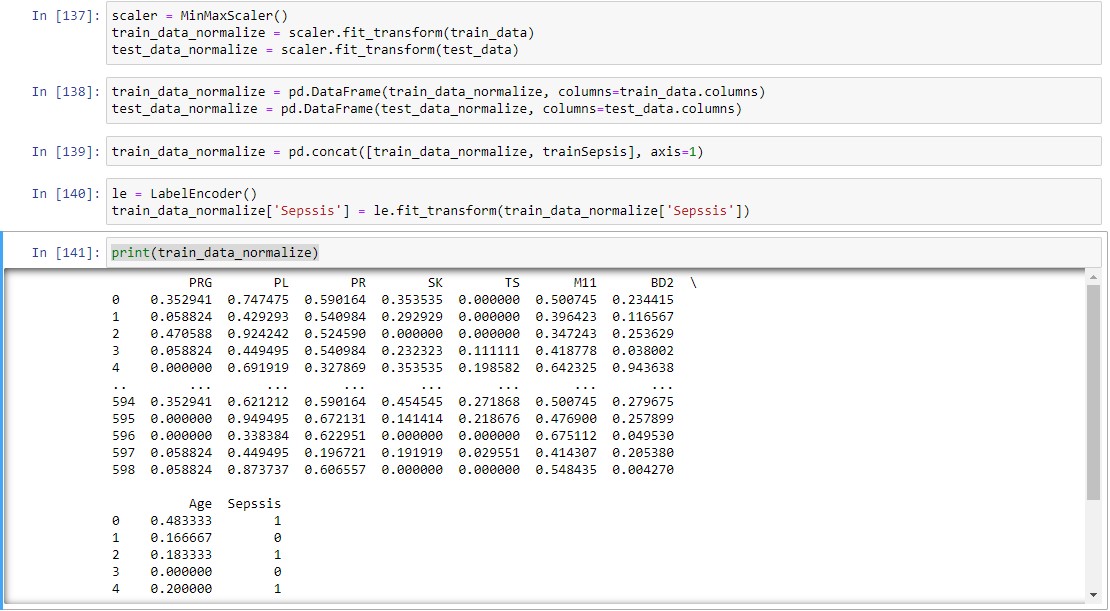
First, I will store the ID column of both train and test file for later use. After that, I need to take unnecessary columns (file ID and Insurance column) out of my data frame and take out Sepsis column for preprocessing step.



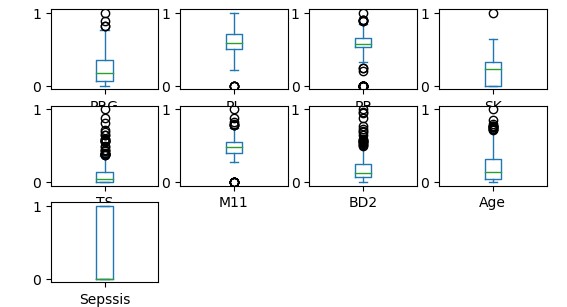
Plotting data, but none of them has Gaussian Curve, so I need to do normalization step.



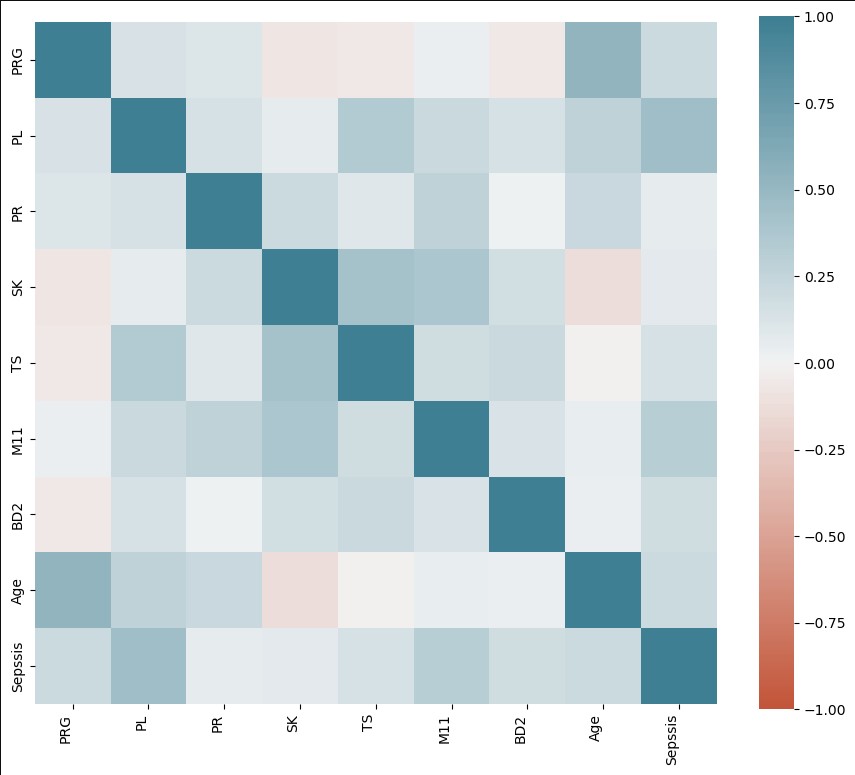
I use MinMaxScaler to normalize data and LabelEncoder to convert the value from Sepssis column to binary 1 and 0. After that, I combine them together. Here is the new result:



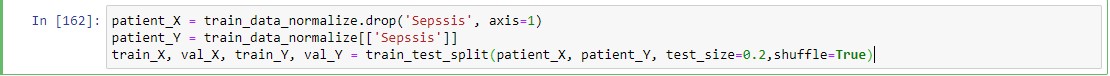
Next, I use box plot to plot outliers. The TS and BD2 seem has the most outliers, but I will not remove them as they may be some important data points:



I plot the correlation map to see the relationship between variables. As I can see the Sepsis column has a positive correlation with all columns, especially the PL column.

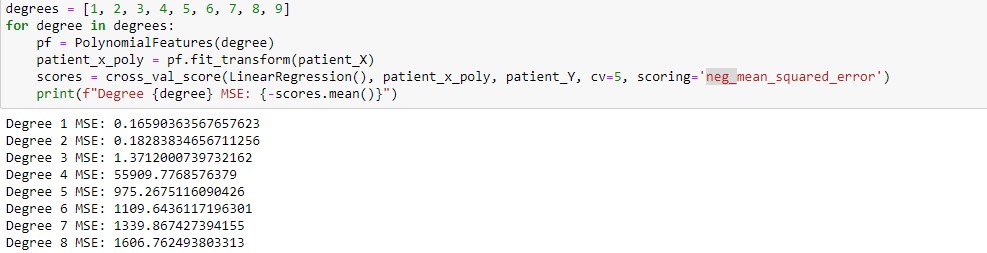


I use train\_test\_split to split data for training. I split data into train and validation:



# Multivariate Linear Regression

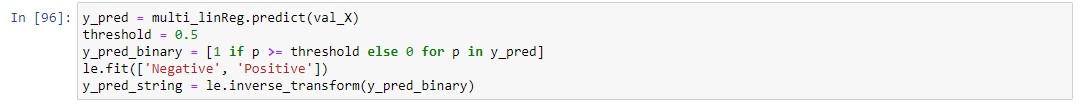
When all variables have positive correlation with Sepsis, Linear Regression is appropriate for training. However, I still want to check if any degree of Polynomial is better than Linear Regression. Therefore, I use cross validation to calculate the MSE from 1 to 9. It turns out the degree of 1, which means Linear Regression is still the best when compared to other degrees.



I want to use various independent variables to predict Sepsis, so I choose Multivariate Linear Regression:



The Linear regression cannot predict binary, so I need to do further steps such as set a threshold = 0.5, and the value will be automatically set to 0 and 1:



The reason I chose F1 score metric is that it is a combination of precision and recall. This is a medical situation, so we need to predict the positive cases as many correct cases as possible while we can reduce the number of false negative predictions.

After that, I calculate the train and validate F1 score. The gap between them is quite acceptable:



Here is the classification report:



# Logistic Regression

Logistic Regression is one of the most suitable models for predicting Sepsis.

Logistic regression needs parameter to train. First, I use GridSearch to search for parameters.

Then, I train a model with that:



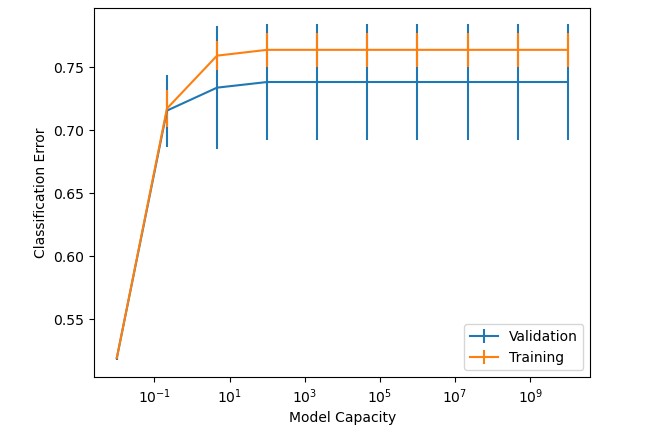
These are the results of the train and validate F1 score of Logistic regression model. This is the best result so far:



Here is the classification report:



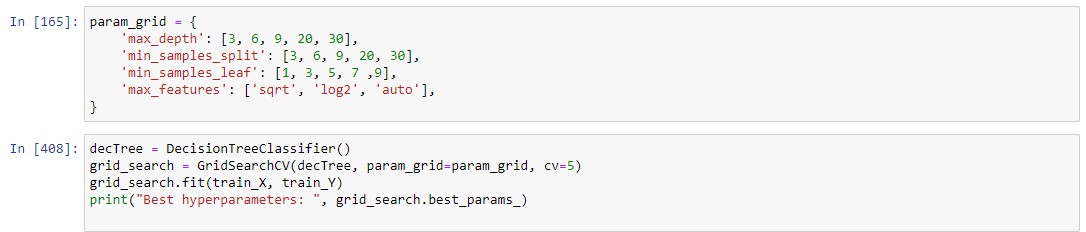
I plot a graph to see their performance. From my perspective, their performance is good:



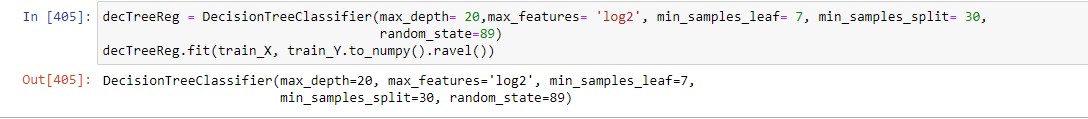
# Decision Tree

Decision Tree is good for handling complex data with various input, so it is appropriate for training predict sepsis model.

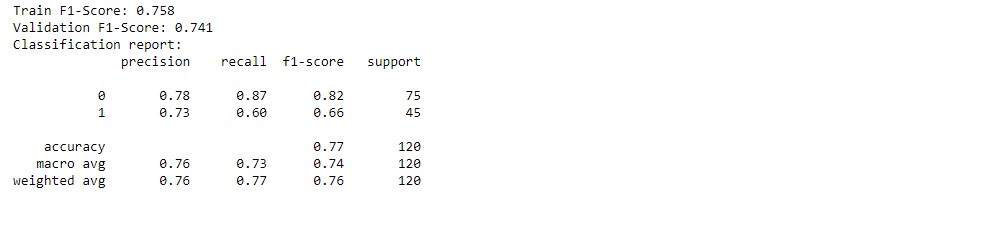
I once again use GridSearch to find the suitable parameter. After that, I use those parameters to train Decision Tree model



Training Decision Tree model:



Here is the F1 score result and classification report:

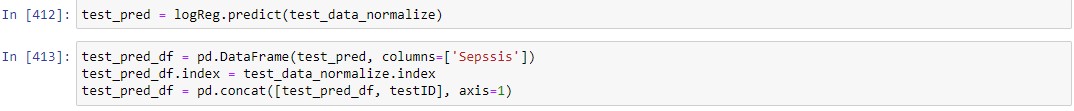


# My Judgement

I create a function called scoreF1 to calculate the mean of F1 of each model after 5 times run. The model which returns the highest score will be taken into consideration. Here is the result:



The Logistic Regression had the highest F1 score. Moreover, from the previous experiment, Logistic regression has the closest gap between validation and training, and their performance on the graph is good. Therefore, I will choose Logistic Regression for predicting test file.



# Exporting Predicting Dataset to CSV

I replace the predicted value 1 and 0 with Positive and Negative. And set the location of column to the correct location. Finally, I exported the Prediction file as CSV.

