Prediction of Survival After Thoracic Surgery in 1 year based on real life data

Final Project Report

PROGRAMMING IN PYTHON [B]

Submitted by:

NAME	ID
NABIL, ABIDUR RAHMAN	19-41607-3
DATTA, ARPITA	19-41608-3
PRIOTY, SAZIN ISRAK	19-41635-3
SANAD, ZARIF AMIR	19-41742-3
LEO, NAFINUR	20-42195-1
OMI, IFAJ HOSSAN	20-42387-1

Submitted to:

DR. AKINUL ISLAM JONY
Associate Professor
Faculty of Science and Technology
American International University-Bangladesh (AIUB)

PROJECT OBJECTIVES

The main purpose of doing this project is to get a proper idea about the machine learning and the model creation in machine learning. This project mainly focuses on the supervised machine learning so one of its purposes can be defined by it. Moreover, different types of Python libraries are used here which also serve a purpose to get a good grip of them. One of the major reasons to work with this project is to work with the real-life data and predict something realistic. This project mainly focuses on predicting the survival which can be defined as a classification problem. With completing this project, it will increase the knowledge about these type problems. This model can be an important asset because it will help people with the info of survival which is one thing people thrive for. Lastly, it can be said that this project's main outcome is to increase the knowledge about the machine learning and different algorithms.

FLOWCHART

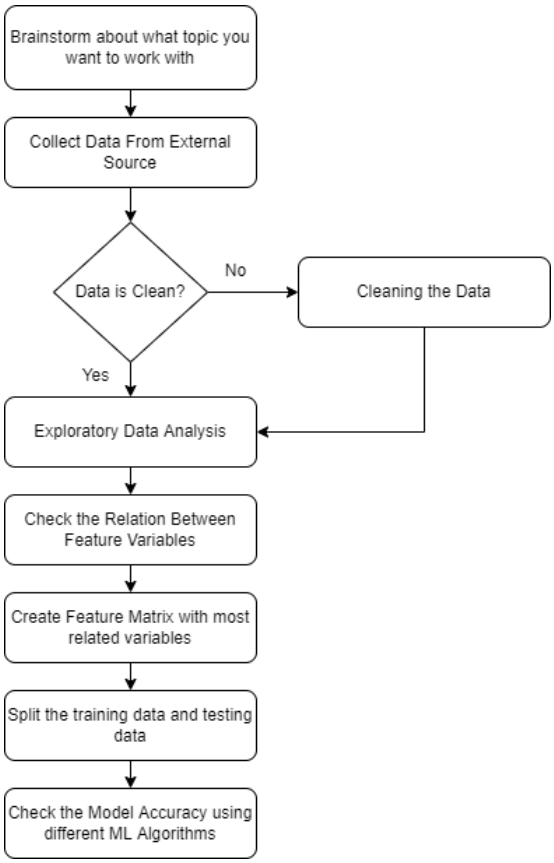

Figure 1: Flowchart for generating the machine learning Model

DATA PREPROCESSING

Data preprocessing is one of the most important tasks to make a machine learning model. Because with dirty data, it is almost impossible to generate an accurate machine learning model. There are several techniques of data cleaning as checking the null value, checking if one of the columns of data set has most empty values, replacing null data with mean, median and mode value. In our project, the dataset was completely clean with all the rows filled up with real data and no null value. Still for surety, the dataset info was checked using a command and it returned all the columns as not null and for this reason this project dataset did not need any cleaning and we could proceed.

DATA ANALYSIS

Data Analysis is another important part of machine learning model creation. Because with data analysis, we can initially get an idea of the prediction most of the time. We can also get the idea of which feature variables are important because those variables will be useful to create the feature matrix. In our project there were several visualizations. There are two or three columns named FVC – Forced Vital Capacity, Age, If the patient had cough before surgery or not etc. They were among with other visualized to recognize a pattern so that from those graphs we can understand if that survival problem is related to age or some disease before the surgery. Unfortunately, all the visualizations did not give any clear indication to understand the survival dependency. One of the major reasons for this can be declared as the dataset contains mostly categorical data. Overall, the data analysis part helped us a little bit to choose the feature matrix. Below two graphs are given.

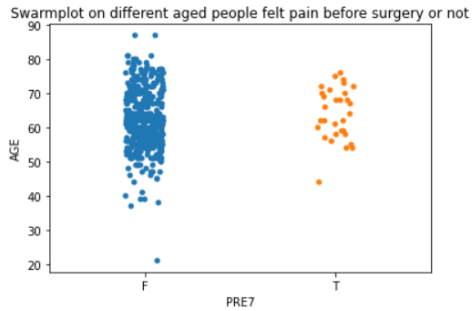

Figure 2: Plotting of comparing age with people feeling pain before surgery

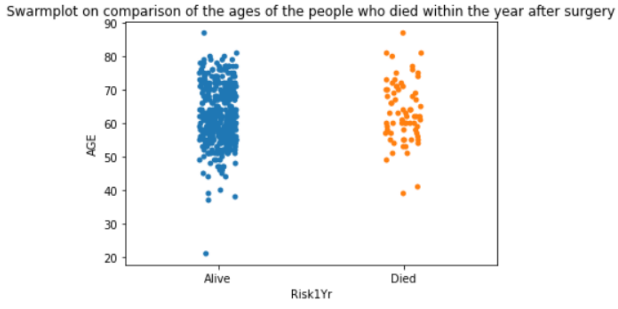

Figure 3: Plotting of comparing age with people died after surgery.

DATA SPLIT

Data splitting is very important because without it the model cannot be generated. Data needs to be spitted into two parts one is for training the dataset and another part is for testing the dataset. In our project there were total of 470 instances of data. Where 329 of 470 is used for training the model. The rest 141 data is used to test the model's accuracy. The spitting mostly depends on user as both training and testing is equally important for the model to be accurate.

MODEL CREATION

After the initial data preprocessing and analysis also with the creation of feature matrix and separating the target variable it is time for creating the machine learning model. Total 5 types of classification models were used. The SVM, DT, KNN, LR and NB. All of them were used to test the model. After the overall testing, the SVM, LR and NB gave the highest accuracy which is 89.36%. Which is not the best but not the worst as well.

DISCUSSION AND CONCLUSION

This model is created based on a thoracic surgery dataset which predicts after this surgery in 1 year the person will survive or not. This model has an accuracy of 89.36%. Which is not the best but not the worst as well. This accuracy was generated based on all the best possible decisions such as choosing the important column to test the data. Still as the accuracy is not around 95% which is thought to be ideal it might be the problem of the dataset. For example, it could have more columns which had more correlation with other variables to predict better. Also, it could have more rows of data to train the model better. This model has increased the knowledge about machine learning specifically the supervised machine learning. As, it was based on real life dataset, the prediction can also be called realistic if it predicts the correct result.