INT305 2020 Coursework 3

Project Proposal

Name: Yuxuan Wu

Student ID: 1716309

Approval Signature: Kaiyshi Wu

Date: 2020.10.27

Background information:

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.
People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Dataset:

Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.
Samples: 299 samples
Features: 'age', 'anaemia', 'creatinine_phosphokinase', 'diabetes', 'ejection_fraction', 'high_blood_pressure', 'platelets', 'serum_creatinine', 'serum_sodium', 'sex', 'smoking'

Labels: 'DEATH_EVENT

Method:

1. Logistic Regression 2. K Nearest Neighbor 3. Decision Tree Classifier 4. Random Forest Classifier 5. SVM 6. XG Boost 7. Cat Boost

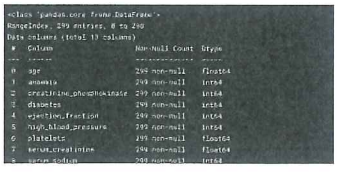
Feature selection

Prior research:

Previous research made some visualization in each feature. They explore the data through some EDA (exploratory data analysis). They also try to detect and extract relevant feature in order to build a prediction model.

Initial exploration:

I downloaded and played with the data. I found that our data are all numerical and do not have missing values which will make our work easier



Target contribution:

I would expect to spend most of my time in data visualization in this work. Since the data sample is a standard classification problem, accompanied by clear feature. The machine learning model itself is not difficult, and the algorithms are not difficult. Therefore, I plan to focus on the dataset feature, to explore the potential patterns and possible correlations. Since I was new to data analysis, the priority was to get familiar with the data and make some clear plots to better demonstrate the ideas.