**Report**

The provided dataset contains information about patients who had heart attack. It has some attributes that are related to the heart attack patients and some other information that are not related to it. There are some issues in this dataset:

* Does not have column headers
* Has small number of records
* Has missing data
* Has irrelevant information, such as Name, Group, etc.
* Does not have some other info that might be important to analyse data and predict target such as patients’ gender and patients’ weight.

First thing this dataset needed was data preparation and data cleaning. Actions needed were:

* Add column name to the dataset
* Fill missing values, there are 3 approaches were available
  + Delete missing value records, this option is not the best as there is small number of records in the dataset and the percentage of missing value records was high compare to the total number of records.
  + Use aggregated values such as mean, median, etc., this is the option that was used, and mean was the function that was applied to fill missing values
  + Build predictive analysis model for each column to predict its value based on other provided attributes, this option will be time consuming and will not be used due to the limited test time.
* Exclude columns that are not related to the target such as name and group, and exclude columns that have high correlation to other attributes such as alive-at-1 (Redundant)
* Create new column to mark records of patients who survived 2 years (field called “survived-2”), this new column could have three values
  + Survived 24 months, either alive now or not. These records will be marked as 1
  + Did not survive 24 months, not alive. These records will be marked as 0
  + Alive but it has not been 24 months yet. These records will be deleted as it is could be either, the patient could live 24 months or more, or could die before the 24 months.

Next stage is Feature Selection, Model Selection and Hyper-Parameter Optimization. In this stage there are 2 approaches.

* To predict the value of survive-2 (if the patient survived 2 years or not), the target could be either 0 or 1, which means it is classification model or logistic regression mode. In this case the classification model was used.
* The second approach is the find the value of number of months survived (or survived so far) from this number each patient can be classified if survived 2 years or not.

To decide the features that are going to give the best prediction, Exhaustive Search method was used to try all possible combinations of features in term to find the ones the gives the best prediction results. Prediction models may have parameters such as the K in Knn model or the number of trees in the Random Forest model, etc. there for Random Search and Grid Search methods were used to find the best hyper-parameters. The Cross Validation 5 folds was used to ensure that model does not overfit and test results represent the prediction model actual prediction ability. Information of the best model and its parameters, and features were kept being used in the actual test dataset.