**Case Study Assignment – Data Mining – GR 30**

**Solution**

**Problem statement**

To build a classification model from the data collected by a healthcare organization together with a couple of government hospitals about the vitals that would reveal if the person might have a coronary heart disease in the next ten years or not. This model will be helpful in early identification of disease , improving the health conditions and the economy.

**Perform exploratory data analysis**

1. Getting to know about the number of samples and variables in the data
2. Checking first order statistics of the variables
3. Data types of the variables – object/floats/int
4. Segregating data into discrete and continuous
5. Checking the imbalance in the target variable (32.9%)
6. Missing values identification (only A2)

**Pre-process the data**

1. Missing value imputation (Medians)
2. Outlier identification (Boxplots)
3. Outlier removal
4. Removal of highly correlated features (>0.9 Pearson’s correlation coefficient)
5. Normalization (Standard scaler)

**Select Training data, test data**

1. A train test split of 80-20. Following is the shape of the data that is obtained. ((27424, 23), (6857, 23))
2. Random seed as 0 is taken to ensure reproducibility of split

**Train the model**

Using random forest as a classifier to train on a dataset of shape: 27,424 X 23

**Test the model (Predictions and reporting)**

Predicting on a test set of shape: 6,857 X 24

**Evaluate the model performance**

We have used the following metrics to evaluate the model performance

1. Accuracy i.e. (TP+TN)/(TP+FP+TN+FN) of 88.26 %
2. Other metrics are Precision, Recall and F-Score

**precision recall f1-score support**

**0 0.90 0.92 0.91 4598**

**1 0.84 0.80 0.82 2259**

**Suggest ways of improving the model**

An intuition about the attributes and their descriptions could help us to identify which attributes will have direct correlation with the target variable. We could also derive more features from the given variables if we know the descriptions about the provided variables.

**Any interesting observations**

1. Only A2 has 5.08 percent of NAs in the data. Rest of the variables do not have any missing values. The missing values are treated by median imputation.
2. 6 variables were highly correlated.
3. Outlier removal - After looking at the diagnostic plots in the EDA section , the variables ['A1', 'A3', 'A4', 'A5', 'A6', 'A7', 'A8', 'A9', 'A10', 'A12', 'A14'] are treated for outliers by the IQR method . The outliers are removed on the right tail. The variables ['IV', 'A2', 'A15', 'A16', 'A21'] are treated for outliers on both the tails .

**Challenges faced and how you mitigated the challenges**

We found out that certain variables are highly correlated and are not improving the model performance by a significant amount. We removed those variables - ['A6', 'A7', 'A9', 'A10', 'A12', 'A16']

**Assumptions if any**

We have assumed that variables with less than 10 cardinality are discrete variables and above 10 are continuous variables since the variables are masked and they don’t have any business meaning here.