**Case Study Assignment – Data Mining**

**Team Members:**

1. Tarang Karoli (2019AD04052)
2. Piyush Khanna (2019AD04035)
3. Ajay Kumar Yadav (2019AD04010)
4. Rakesh Yadav (2019AD04111)

**For the given case study, the following steps were performed**:

1. The dataset was imported and split into sets of independent variable and the target variable(s) based on feature extraction.
2. An exploratory data analysis was performed on the dataset to check if there were any missing values.
3. The missing values were replaced with the mean of the respective columns.
4. As the data range of the different columns were all over the place we needed to standardise the values, thus the data was pre-processed to be normalised.
5. Next step included the splitting of the data in to training and test data. This was performed using Stratified K-fold cross validation, so that the mean of the different splits remain even during building the model.
6. As to classify the data and train the model several classifiers were available, few of them along with their accuracy metrics are:
7. Naïve-Bayes Classification (Accuracy: 79%)
8. Decision Tree Classification (Accuracy: 88%)
9. Random Forest Classification (Accuracy: 91%)

So of the three classifiers tested we went along with **Random Forest** classifier.

1. Next step was to fit the model over different iterations and after that we predicted the data alongside the expected target data.
2. To evaluate the model performance the different measure were used: **Accuracy**, **F1-Score**, **Precision** and **Recall**, and **Confusion Matrix** was also produced.

**Ways to improve the model**:

Rather than performing classification using only one set of classifier we can perform classifications using different classifiers and then ensemble them and stack them to beget a more accurate classification model.

Challenges faced during the task:

1. One of the challenges faced was feature selection in the initial phase of exploratory analysis and what columns to keep for building the model. As observed that the ID column has not much contribution during the classification.
2. The train and test split of the data, as it can simply be done by assigning the percentage of the data assigned for train and vice-versa. But that may lead to under fitting of the model and thus we required a method of split which will consider each tuple of data and evaluate them equally, and this was achieved through Stratified K-Fold Cross Validation.

**Observations**:

From the produced result we can observe that the cases of the positive cases of coronary heart disease is quite few as compared to the negative cases of the heart disease and thus it indicates the better health performance of the citizens and in association a well economic performance of the given country.