Individual Report

Course Title: MSc in Data Analytics

Lecturer Name: Courtney Ford

Module Title: Machine Learning and Pattern Recognition

Assignment Title: Supervised Machine Learning – Regression

Submitted by:

10634205 - Sandeep Gopu

This was the first machine learning group assignment given to us in the first semester, so we were sure that we had to put in our 100 %. The assignment was to finalize a dataset from kaggle and using either linear or logistic regression model we have to train our model to predict the accuracy of it. So, we finalized a group of three members and sat together to discuss what all needs to be done. Firstly, we noted down the different steps involved in our assignments. Then decided to divide the tasks and assign it to the different group members. But we made sure to always be there for our team members whenever they faced any difficulty.

So, my part was to finalize a dataset from kaggle as per the instructions given in the assignment. I went through many datasets like Employee Attrition in Companies, Various Health Issues in US, Crypto currency trends, Football Soccer club but found that due to recent pandemic situation, the cases of heart attack got increased, which is becoming an area of concern as most of the cases are of young adults who were depressed in this pandemic situation and they are unable to share their emotions with others. Due to the surprising increase in heart attack, the health sectors are facing challenges in predicting the chances of attack as there are many factors/categories and sub factors/categories due to which it is not an easy task to create a plan manually for identifying which gender, the age bracket and the physical conditions/mental state of the individual is vital reason of heart attack. So, I come up with heart dataset from Kaggle which after analysing I found was fitting in the requirement given to us in the assignment.

So, I discussed with my group members and explain them why I come up with this dataset as best fit for our assignment. During the group discussion there was on point which was not align as per the assignment which was the dataset should be of minimum 15 columns, but the heart dataset is having only 14 columns. Then me and my group members decided to clarify this doubt with our professor and after discussing with our professor we finalized the heart dataset as it is one of the major sectors which plays a critical role in heart sector of every nation and if we train our machine learning code to predict an accurate result of hear attack cases then that model can play a huge role in upcoming days where sometimes even manual expertise is not enough to predict the chances.

Also, before finalizing the dataset I also deeply went into all the 14 attributes of this dataset and found that it has almost all the main categories which are generally necessary in training a model. Moreover, as the target column is “output” column and it has value either 0 or 1 so my group members and I decided to proceed with Logistic regression model instead of Linear regression model because as per the definition if a model has categorical value, then it satisfies the logistic regression and in our case the output has categorical value i.e either 0 or 1.