Individual Report

In this report I will be talking about what I have contributed to the IN3062 coursework. The contributions I have made were towards the coding, testing and report. I will also use this report as a chance to write some feedback about the work I have completed. In addition, I will discuss the performance/teamwork of my team and me. In this coursework, as a team we decided to look at factors which affected stroke as our problem domain.

For the coding section everyone contributed a satisfactory amount that everyone in the group was happy with. Abarna: Decision Tree, Naive Bayes, Data Clean (drop unnecessary columns). Jeeves: Smote, Confusion Matrix, Logistic and Linear Regression, residence type and work type testing if it improves accuracy. My contribution to the code was Pandas Cleaning (BMI Missing, gender, removing unknown, smoking status, ever married), Train Test Split and Random Forest. Cleaning the data was important as we needed to remove unnecessary values, and change values to numerical values that can be used in the machine learning models. Given that the dataset was already small, I chose to replace missing BMI values with the median rather than removing the whole row completely because I had already removed rows of data which has smoking status as being 'unknown'. I chose to remove those because we couldn't take a guess on something like that because smoking does play a large factor in strokes. I also coded the train test split which was used to split out data frame into testing and training. The final part I did for the coding section was writing a function that took in the train test split data as parameters and trained the random forest model with it. Mine had the highest accuracy turnout but when it came to the confusion matrix it did not make many predictions which I was not happy with.

In addition, my contributions to the report were as follows: Working with Shajeevan to do classification or regression. We spoke about how our problem domain, predicting heart stroke, was a classification or regression problem. I also spoke about missing/misleading data and what I did to deal with it. Omitted data, I wrote about the types of data that I had removed from the dataset and what would happen to the results if I had left it in. For example, in some instances the smoking status was unknown, and I believed that would affect the models because smoking plays a large role in strokes. SMOTE, how Shajeevan worked on the code to fix the unbalanced dataset using a library called imbalaned_learn. It was used to improve the ratio of 95% non-stroke to 5% stroke data. We all worked together on conclusion because it was the final part used to summarise what we have learnt.

Finally, I will be talking about how we worked as a team together. As a team we worked very well. I believe this was largely down to the fact that we personally know each other which made communication between us a lot easier. In a group, doing any sort of work, it will be much easier if everyone is talking and getting along allowing us to provide feedback to one another and improve the tasks we were assigned. We also started the work early on which gave us time to get most of the work complete early and allow us extra time at the end to improve based off lab feedback and referring a lot to the PDF file provided on Moodle.

In conclusion, the coursework went very smoothly. We managed to complete the code in such a time that allowed us to write a good report which mostly covered what was required of us.