Project Description

Data preparation:

Even though we changed the dataset from the previous part of the project, we decided to work on a classification problem again. Also, we did not change the theme, and it’s about again predicting a heart attack, assessing risks of heart attacks. Since the dataset is raw, it contains values which should be turned into numerical values or be dropped. We decided to drop “Patient ID”, “Country”, “Continent” and “Hemisphere” as these columns likely have nothing to do with the prediction. Likewise, we will turn “Blood Pressure” to two columns as an upper blood pressure and the lower blood pressure, and append them to the values. We also modified “Sex” column by mapping into values of 1s and 0s. Additionally, we filled out missing values in the “Exercise” column by inserting the median value of the column.

Algorithms:

To predict a possible heart attack, we used almost all distinct machine learning algorithms that we have covered during the course. We even used some algorithms that we did not explicitly cover but are easily available in the python library; for example, “RandomForestClassifier”. Alongside “RandomForestClassifier”, we employed “LogisticRegression”, “SVM clasiffier”, “K means clustering”, “K nearest neighbors ” and deep artificial neural networks. Since the data preparation part was absolutely the same for each of the methods, only few changes were necessary to move to another algorithm, and each successive algorithm took less and less time. Moreover, it turned out that the first successful, debugged algorithm took more time than the rest of algorithms’ consumed time combined, since modifying a few lines was usually enough for another algorithm and data preparation codes were exactly the same. Here, “time consumed” refers to the time we spent on the project, and not “time” from computational complexity perspective. “K means clustering” did not directly show accuracies but did demonstrate existence of two distinct groups. Of course, we could add additional features to the “K means clustering”, so that it would show accuracy scores, but we considered that would be beyond our course requirements. Overall, we decided to restrain ourselves by having “pure mathematical analysis” so that we don’t go far beyond our course requirements.

Fine Tuning, Parameters Tuning:

To test different algorithms for various parameters so that we pick the most effective value for the parameters, we decided to simply loop over the parameters to get the best possible values. Even though this method may take an extended period of time, it guarantees that we find the best possible values approximately. We did parameters tuning and testing using loops for all algorithms except” SVM classifier”, “deep neural networks”, ”Random Forest Classifier” and ” k means clustering” because “SVM classifier” and “random forest Classifier” took an impractical amount of time and “k means clustering” just shows two clusters. As for ANN, we decided not to test parameters because we would waste enormous amount of time without guarantees that we would have the best value for parameters.

Analysis:

The overall project is not short of deficiencies for many reasons; however, we believe the two most important contributors to the project’s shortcomings are our lack of deep mathematical knowledge of the used machine learning algorithms and boundaries set by our limited introductory course on machine learning, its mathematical understanding. Also, we decided to split the dataset into two rather than three parts for simplicity reasons even though splitting dataset into three parts- training, evaluation and testing-is much more appropriate. Even with all the shortcomings, we still managed, though in limited quality, to predict heart attack risks with reasonable accuracy.