Christopher Cui

Deliverable 1

For my final project, I will be analyzing a dataset of heart disease cases with the goal of building a model to predict heart disease in patients. This dataset is taken from the UCI Heart Disease Data set, which is a subset of 4 databases from Cleveland, Hungary, Switzerland and the VA Long Beach.

I chose this dataset because it has an established reputation in the machine learning field. Up to this date, the Cleveland database is the only one that has been used to predict heart disease using ML by researchers.

Furthermore, this dataset contains 14 features that are used to predict whether or not an individual is expected to have heart disease. The prediction is one of classification, where the predicted output (14th feature) has two values of either 1 (heart disease) or 0 (no heart disease).

I plan to use multiple machine learning models and compare their accuracy, as well as find the most accurate model for this particular dataset. There are 303 data points in this dataset. I plan on doing a 80/20 split between training and testing data points, and then further split the training data points to a 80/20 split between those and validation points.

Various ML models were compared by their performance and usage for disease risk prediction in a study published in the journal of BMC Medical Informatics and Decision Making written by Shahadat Uddin et al.,

The model used by this project will be heavily influenced by the results of this study. Although SVM is intrinsically suited for 2 class problems and Random Forest is for multiclass problems, this study found that Random forest had the highest accuracy out of all the models.

In short, the study found that Random Forest, which is a collection of decision trees, showed superior accuracy (53%) in 9 out of the 17 studies it was used in (Uddin 1). This was followed by SVM (support vector machine) which had an accuracy of 41%. Due to this significant difference in accuracy between the models, I expect Random Forest to be the most effective model for this heart disease dataset. However, I will also be implementing SVM in this dataset to compare the effectiveness of the models.

I plan to create a web app for this project using React.js, HTML and python. The layout of the webpage will be simple. There will be only 2 pages, the homepage and the about page. On the homepage, the user will have 14 text boxes to enter the 14 features for a patient. They will then click “Predict heart disease” which will run the model and output the prediction based on the 14 features entered by the user.

Works Cited:

1. <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8>