Heart Disease Prediction

By Bashar Eskandar

The dataset[1] that will be used combines data from 5 datasets, forming 918 unique data points with 11 attributes. Each record has a Boolean "HeartDisease" label, where '1' indicates the presence of heart disease and '0', its absence. This will be the prediction target.

The dataset does not have any empty or null values. Most attributes have float and int types which can be used directly, while others like "ChestPainType" and "ST_Slope" are String values that will be encoded according to a one-hot encoding schema. The number of data points with a "HeartDisease" of 1 is 508, while the count for a "HeartDisease" of 0 is 410. The dataset is then balanced.

The goal of this project is to create a machine-learning model that will predict the presence or absence of heart disease given a set of heart health indicators for any patient. This will be achieved through a K-Nearest Neighbors model. Given the relatively low number of data points, memory usage will not be an issue. The dataset will be split into a training set and a validation set by randomly choosing 80% of the data points and 20%, respectively. Considering the high dimensionality of the attributes vector relative to the number of data points, different approaches will be tested to attain the highest possible prediction accuracy, in addition to optimizing the *k* hyperparameter.

Two distance calculation methods will be tried: the Manhattan and Euclidian distance methods. The Manhattan method is often preferred for high dimensionality, which should work better for this dataset. The normalization of the attributes will be attempted to get better integration between binary and continuous attributes. Finally, k-fold cross validation will be performed to determine the best performing model.

The models will be evaluated through their confusion matrix. The most important model performance indicators in such a health-related setting would be the Misclassification, True Positive, False Positive, and True Negative rates.

The model will be made easy to use on a one-page website that will allow first-line health care providers to estimate the heart disease risk of patients. After inputting the required heart health indicators for a patient, the user can obtain a rapid High/Low risk assessment with the proper accuracy measures mentioned above. An interesting upgrade of the platform would be to allow only certified users ,such as cardiologists, to submit new labeled data points that would then be collected and added to the existing training data. By periodically retraining the model on an increasing number of high-quality data points, the accuracy and real-world impact of the model can be greatly improved.

[1] https://www.kaggle.com/fedesoriano/heart-failure-prediction