Heart Disease Prediction (Deliverable 2)

By Bashar Eskandar

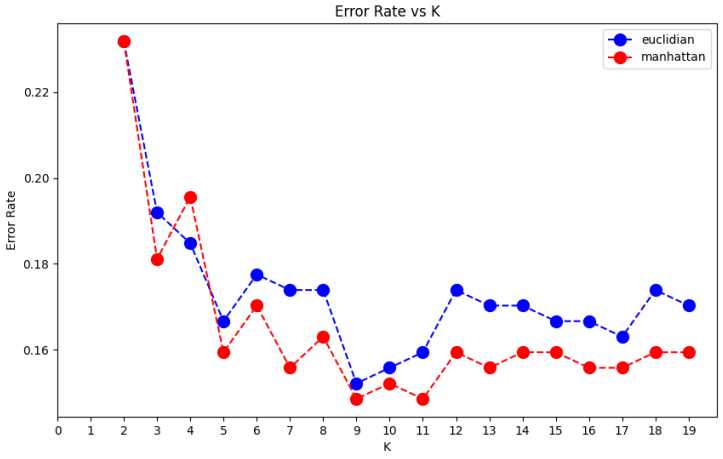
The goal of this project is to create a machine-learning model that will predict the presence or absence of heart disease given a set of heart health indicators for any patient.

Data Preprocessing

The proposed dataset[1] was used. This dataset contains 918 unique data points, each with a target label 'HeartDisease' that has value '1' when heart disease is present, and value '0' when no heart disease is present. The features vector dimension is 11, and features are health indicators such as Age and Fasting Blood Sugar. The data set had no duplicates or null values, and it was balanced: the 'HeartDisease' label distribution has a mean of 0.55, a std. of 0.50, and a skewness of -0.21. Labels "Sex", "ChestPainType", "RestingECG", "ExerciseAngina", and "ST_Slope" had to be encoded. Feature scaling was performed through mean normalization.

Machine Learning Model

The proposed KNN model was applied to the dataset. The sklearn library's preprocessing module was used to encode String attributes into positive integers and to apply mean normalization to the features to get better integration between continuous and discrete attributes. The data set was then randomly split by a ratio of 70%/30% into a training subset and a test subset. The 'KNeighborsClassifier' Class from the sklearn library was used to create the KNN model. For each Hyperparameter K in the interval (2,20), the error rate of the model with test data was calculated. This was done with both Euclidian and Manhattan distance calculation methods. The range starts with 2 to avoid overfitting and ends with 20 to avoid underfitting and increasingly long prediction time. Manhattan distance calculation method was tried as it usually performs the best with small data sets with high dimensionality, as is the case here. The obtained model uses a K value of 9 with Manhattan distance calculation to obtain a minimized error rate of 0.15. The following graph illustrates the error minimization process.



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

Model Performance and Project Feasibility

In such a health-related classification problem, it is important to minimize the false positive and false negative rates, in addition to minimizing the misclassification or error rate. The obtained model has a confusion matrix of [[104 22], [19 131]]. This results in a false positive rate of 0.17, a false negative rate of 0.13 and a misclassification rate of 0.15. These error rates are comparable to those of already existing medical tests. For a first-line rapid disease detection tool, these results are good enough to proceed with the deployment of the model as an easy-to-use web app that can be used by first-line health care providers.