Heart Disease Classification Report

Sanjana Farial (14.02.04.100)
Sharmin Sultana (13.02.04.006)
Group ID (B1)
Lab Group No (6)

1 Problem Description

As we know, the presence or absence of heart disease in a person depends on several properties. We have worked with a dataset containing various features of male and female of different ages to classify if they have heart disease or not.

2 Dataset Description

Our dataset contains 9 columns in total including the target column. 8 of the columns display different features and condition of patients and the last column which is the target column shows if the existence of heart disease in the relative patient is positive or negative. There are information of 274 patients in total.

3 Description of the Models

In our project, we have applied 5 different models in order to find out which one performs the best. The dataset contains string values that are converted into numerical values and split into two parts, 80% data is for training and the rest 20% is for testing the models. To avoid unwanted deviation of results over non categorized data, we have scaled and standardized our dataset with **standardscaler** method. We have performed k-fold method to divide our dataset into 5 parts for further operation training and testing operation, all of the 5-folds were used for training and testing.

3.1 Individual Methods

We have used 3 individual models in our project:

3.1.1 SGD Classifier

Stochastic Gradient Descent (SGD) is a linear classifier and works with large scale data represented as dense or sparse arrays of floating point values for the features. The gradient of the loss is estimated and the model is updated along the way with a decreasing strength schedule.
Some of the other **advantages** of the SGD classifier are, it is efficient and simple to implement.

One of the many **disadvantages** of SGD is, it is sensitive to feature scaling or non-standardized data.

3.1.2 SVM Classifier

Support Vector Machine (SVM) classifier is a non probabilistic binary linear classifier but it is different to linear classifier as it creates an area with boundary between the classes in contrast to linear classifier.
Advantages of SVM classifiers are, it is memory efficient and versatile. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.
Disdvantages of SVM classifiers are, limitation in speed and size, both in training and testing. Working on discreet and non-scaled data leads to a very poor performance.

3.1.3 KNN Classifier

K-Nearest Neighbour (KNN) classifier finds a predefined number of training samples closest in distance to the new point, and predicts the label from these. The number of samples is a user-defined constant in K-Nearest Neighbour learning. Neighbors-based methods are known as non-generalizing machine learning methods, since they simply remember all of its training data. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
Advantages of KNN classifiers are, it is robust with noisy data and very effective if the training split is large.
One of the **disadvantages** of KNN classifiers is, its computation cost is quite high as we need to compute the distance for each training sample.

3.2 Ensemble Models

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. We have used two ensemble methods in our project:

3.2.1 Random Forest Classifier

In **Random Forest Classifier**, each tree in the ensemble is built from a sample drawn with replacement from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.
Some significant **advantages** of Random Forest classifiers are : It can work with large dataset. Random Forest classifier is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

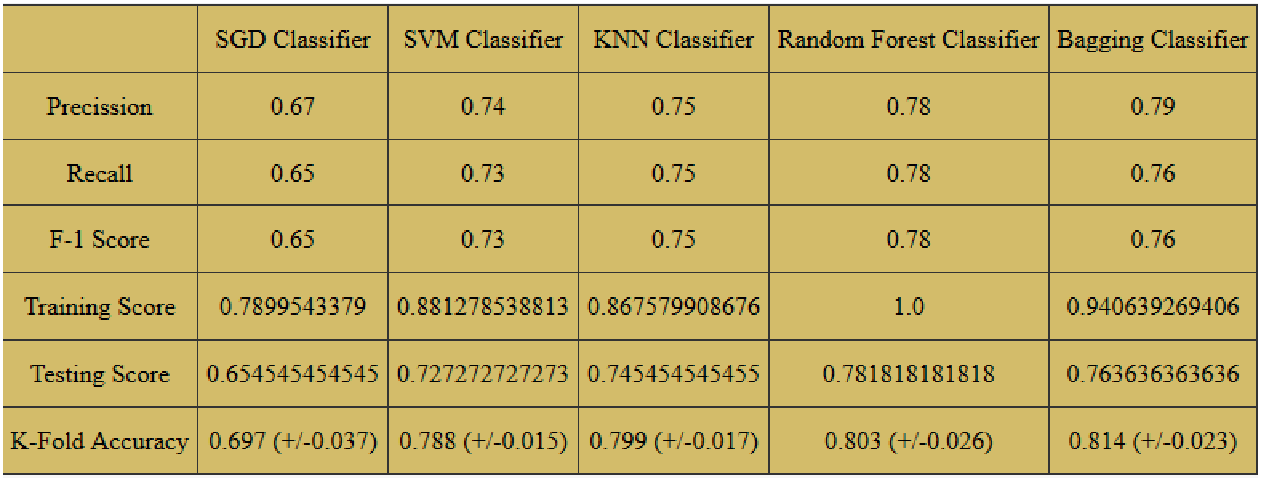
One **disadvantage** of Random Forest classifier is, Random forests have been observed to overfit for some datasets with noisy classification task.

3.2.2 Bagging Classifier

Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method. It creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set.
In our model, the **Decision Tree Classifier** is ensembled with **Bagging** classifier. Decision Tree classifier breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision.

Random Forests improve variance by reducing correlation between trees, this is accomplished by random selection of feature-subset for split at each node whereas, Bagging improves variance by averaging, majority selection of outcome from multiple fully grown trees on variants of training set. It uses Bootstrap with replacement to generate multiple training sets.

4 Comparison of the Performance of the Models



	SGD Classifier	SVM Classifier	KNN Classifier	Random Forest Classifier	Bagging Classifier
Precission	0.67	0.74	0.75	0.78	0.79
Recall	0.65	0.73	0.75	0.78	0.76
F-1 Score	0.65	0.73	0.75	0.78	0.76
Training Score	0.7899543379	0.881278538813	0.867579908676	1.0	0.940639269406
Testing Score	0.654545454545	0.727272727273	0.745454545455	0.781818181818	0.763636363636
K-Fold Accuracy	0.697 (+/-0.037)	0.788 (+/-0.015)	0.799 (+/-0.017)	0.803 (+/-0.026)	0.814 (+/-0.023)

Figure 1: Comparison of Models

5 Discussion

Both of the ensemble methods, Bagging and Random forest has the best performance all over the other models, but Random Forest classifier is overfitting in a sense that the training accuracy is 100%. The individual model that produces a result closest to them is the K-nearest neighbours classifier. Scaling and standardizing the data doesn't affect the performance of the ensemble methods much whereas, a distinctive improvement is seen in the individual methods when data are standardized.