Comparison of Supervised Learning Algorithms on Prediction of Heart Disease

Chithra Bhat, Keerthana Jayaprakash, Sneha Bezawada, Ujjwala Yarra, Vimohitha Sridhar

CMPT 726 Machine Learning, Fall 2017, Simon Fraser University

Abstract

With increased focus on preventing medical errors and incorrect diagnosis, Clinical Decision Support Systems have become prevalent in the medical industry. This tool allows decision-makers to apply evidence-based methods to make objective clinical decisions based on existing patient data. The main objective of the project is to develop such a system to classify heart disease in a patient using Supervised Machine learning models. We considered Naive Bayes, Logistic Regression , Support Vector Machines and ensemble models like Random Forest , Gradient Boosting and XG Boosting to evaluate the performance.

1 Dataset Information

The dataset used in this project is taken from UCI Machine Learning Repository. It has 920 records from 4 medical institutions across USA and Europe.

The Dataset consist of patients medical information which is classified into 76 attributes. Due to the number of attributes in the datasets, we will refrain from explaining each of them in detail, as this information can be found in the link http://archive.ics.uci.edu/ml/datasets/heart+Disease. The dataset has both continuous and categorical values. The missing values are denoted as -9 in the dataset.The class variable (num) refers to the presence of heart disease in the patient.

2 Data Preprocessing

The original 4 datasets have been merged into 2 datasets. One representing hospitals in USA and the other in Europe.In the USA dataset out of 76 attributes, 37 attributes have been dropped due to the presence dummy information in majority of the rows. Similarly 32 attributes have been dropped from Europe dataset. An additional column loc representing location of hospital has been added to the dataset.

There were missing values in the remaining attributes. For continuous values, we imputed the missing values with their mean. For categorical attributes with very few missing values we randomly assigned a category to them.The discrete valued columns have been split into dummy columns using pandas getdummies() function. Since different features have different scale, the entire data has been standardized to 0 mean and unit variance.

We also performed correlation analysis between all the attributes. Only one attribute has been taken and rest are dropped from such highly correlated columns. For instance smoke, cigs attributes had very high correlation and so the smoke attribute has been dropped. We also found some interesting results from our analysis. Cp , exang have the highest correlation with the predictor attribute which means patients who exhibit chest pain or exercise induced angina are highly susceptible to heart disease. Apart from these age, sex, ekgyr, proto, thal, slope, old peak, cyr are highly correlated to the target.

3 Supervised Learning Algorithms

3.1 Naive Bayes Classifier

3.1.1 Introduction:

Naive Bayes is a classification algorithm for binary and multi-class classification problems. The most distinctive feature of Naive Bayes classifier is that, the hypothesis is formed by counting the frequency of various data combinations within the training examples.

3.1.2 Hyperparameter Tuning:

The best parameters occurred at : 'alpha': 2.0 for both the datasets.

> 'alpha': alpha is a additive smoothing parameter and the default value is 1.0.If no smoothing is required ,the value is 0.0.

3.1.3 Evaluation

Table 1: Model Evaluation Metrics

Dataset	Training Accuracy	Testing Accuracy	Sensitivity	Specificity
USA	0.78947	0.80991	0.81818	0.80519
Euro	0.85576	0.86666	0.80392	0.92592

The model is evaluated based on confusion matrix metrics, which shows the number of correct and incorrect predictions made by the classification model compared to the target value in the data.As classification accuracy is the percentage of number of correct predictions made divided by the total number of the predictions, it is alone not sufficient to evaluate the model. Hence we consider the Sensitivity and Specificity metrics as well. The Sensitivity of both datasets is almost similar whereas the Classification accuracy and specificity for the Euro Dataset is higher, proving that the Naive Bayes Classifier performs better on Euro Dataset as compared to USA dataset.

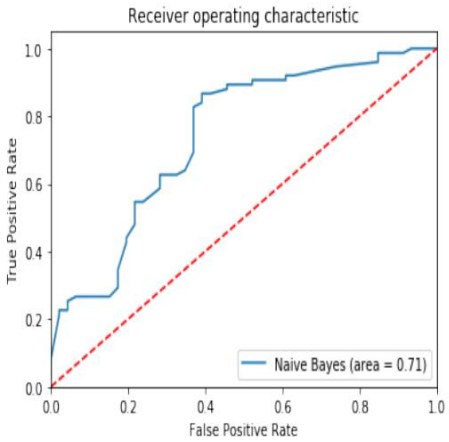

Figure 1: ROC Curve on USA Dataset

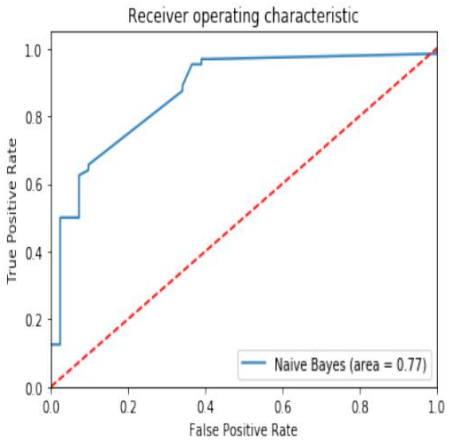

Figure 2: ROC Curve on EURO Dataset

The Receiver Operating Characteristic(ROC Curve)(Fig 1, Fig 2) which plots the Sensitivity metric on Y-axis versus the Specificity metric(also known as False Positive Rate) on the X-axis, measures the usefulness of a diagnostic test in general by also evaluating its quality.

3.2 Logistic Regression

3.2.1 Introduction:

Logistic Regression is best used when it comes to predicting an outcome target variable from one or more independent continuous and/or categorical variables. Logistic Regression qualifies as a better classifier for the medical field as it aims at building a biologically acceptable model that can identify the relationship between dependent and independent variables in a way that will have the best suitability with use of the least variable. As it does not have any assumption limitations and is mathematically flexible, it can be easily interpreted and leads to meaningful implementations.

3.2.2 Hyperparameter Tuning:

For the USA dataset, the best hyperparameters are : 'C': 1, 'penalty': 'l1' and for the Euro dataset, the best hyperparameters are: 'C': 100, 'penalty': 'l2'

> 'C': The parameter C is the the inverse of regularization strength in Logistic Regression.The smaller values of C constraint the model more as inverse to the larger values of C which give more freedom to the model.

'penality': It is used to specify the norm used in the penalization.l1 penalty generates sparser solutions while L2 doesn't. Sparse model is a great property to have when dealing with high-dimensional data, for at least 2 reasons : Model compression and Feature selection.

3.2.3 Evaluation

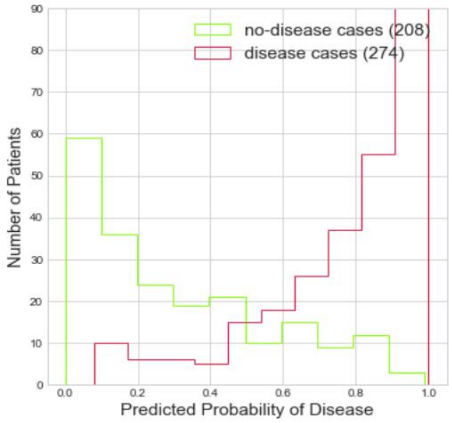

Figure 3: Prediction of Heart Disease Rateon USA Dataset

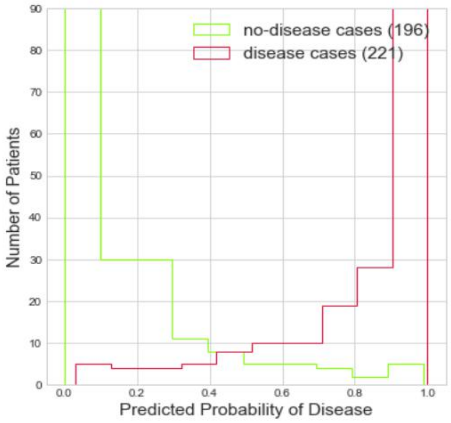

Figure 4: Prediction of Heart Disease Rateon EURO Dataset

In a pharmaceutical company, in any medical related industry, they will be more concerned with accurately detecting the disease in a patient that is High Sensitivity and specificity. Our evaluation shows high specificity than sensitivity values which means the the model is not as good at detecting the presence of the disease

Table 2: Model Evaluation Metrics - GridSearchCV

Dataset	Training Accuracy	Testing Accuracy	Sensitivity	Specificity
USA	0.8310	0.7685	0.7346	0.7916
Euro	0.91025	0.8761	0.8076	0.9433

From the Model Evaluation Metrics table here, the Euro dataset has higher specificity compared to USA dataset, which means that the EURO dataset is more accurate and performs better in heart disease prediction.

3.3 Support Vector Machine - SVM

3.3.1 Introduction:

In this project we have used support vector classification model to train and test the dataset. The support vector machine algorithm first plots the data in n-dimensional space, and then finds a hyper-plane which will differentiate the two classes.

The main advantages of using SVM in our project is, it has a regularization and parameter to avoid data from overfitting.

3.3.2 Hyperparameter Tuning:

To find the best parameter values, we built a parameter grid and evaluated each parameter value. The major parameters that impacts SVM performance are:

C: is a Penalty parameter of the error term. It also controls the trade-off between smooth decision boundary and classifying the training points correctly. The best C value is found using 10 cross validation on the training datasets. USA dataset C= 5 and Euro dataset C= 25

Kernel: Linear kernel is found to be the best parameter value on both the dataset. This uses a function to transform the data into higher dimensional feature space which separates the classes linearly.

3.3.3 Recursive Feature Elimination:

To optimize the model, we have used a feature selection algorithm called Recursive Feature Elimination. This algorithm will eliminate irrelevant features and reduce the data dimension to increase the learning efficiency. RFECV from scikit

learn performs RFE in a cross-validation loop to find the optimal number of features. Following graphs illustrates the cross-validation score and the optimal number of features used to achieve the score.

Number of optimal features - USA Dataset:14, Euro Dataset:26

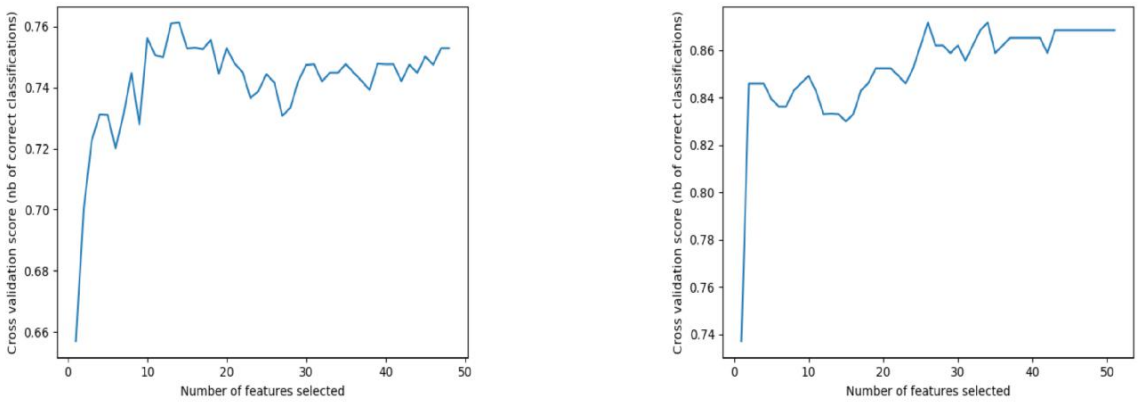

Figure 5: Cross-Validation Score Vs No of Optimal Features for USA and EURO Dataset

3.3.4 Evaluation:

The model is evaluated on metrics like classification accuracy and confusion matrix.

Table 3: Model Evaluation Metrics

Dataset	Training Accuracy	Testing Accuracy	Sensitivity	Specificity
USA	0.8310	0.7685	0.7254	0.8000
Euro	0.9294	0.8857	0.8235	0.9444

Euro dataset has better accuracy than USA dataset. Because Euro dataset has attributes which are highly informative to predict the presence of heart disease. The sensitivity and the test accuracy is comparatively low for USA dataset which implies the model is not a best suit to identify the presence of heart diseases. Also from the above table it is evident that the model works well on Euro dataset.

3.4 Random Forest Classifier

3.4.1 Introduction:

Another method that works effectively with classifying heart disease is Random Forest. It is an ensemble learning classifier which aggregates the votes from different decision tree built from randomly selected subset of training set, to decide the final class. It combines bagging and randomized selection of features. Some of the main features of Random Forest clas-sifier that makes it suitable for our data are: It works well with high dimensional data,handles missing and unbalanced data well,generates relative importance of features which comes in handy for feature selection.The scikit learn RandomForestClas-sifier is implemented to train our dataset.

3.4.2 Hyperparameter-Tuning:

The following parameters are tuned for our algorithm using scikit learn GridSearchCV with 10 fold cross validation.

Criterion: gini is the default splitting criterion but after performing grid search entropy came out to be better. USA dataset : criterion = entropy EURO dataset: criterion = entropy

n_estimators: This parameter refers to number of trees to built before averaging the predictions. USA dataset : n estimators = 51 EURO dataset: n estimators = 152

max_depth: Lesser the value of max depth, less is the overfitting of data. USA dataset : max depth = 3 EURO dataset: max depth = 2

min_samples split: For small dataset with more features, min samples split reduces the overfitting of data. USA dataset : min samples split = 53 EURO dataset: min samples split = 61

3.4.3 Feature Selection:

10 features with highest importance and their correlation with the target attribute are shown in Fig:6 and Fig: 7.It can be seen that more than half of them are highly correlated with the target variable.

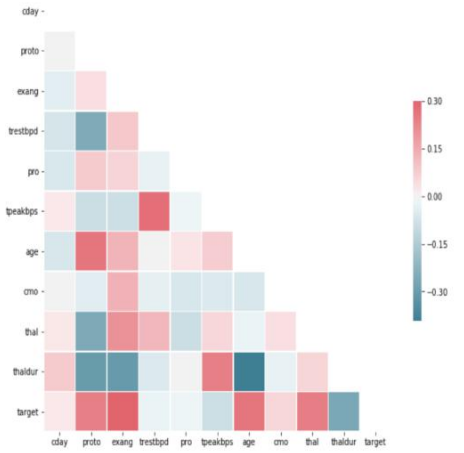

Figure 6: Correlation matrix for USA dataset

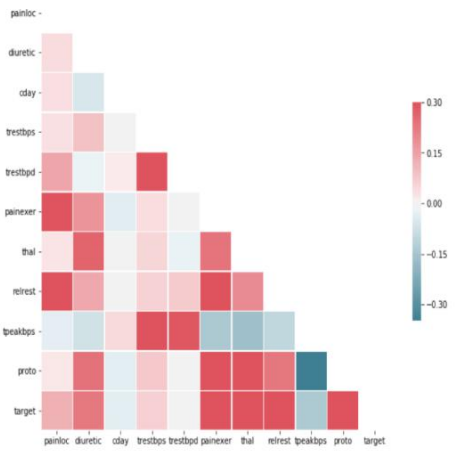

Figure 7: Correlation matrix for EURO dataset

3.4.4 Evaluation:

Table 4: Model Evaluation Metrics

Dataset	Training Accuracy	Testing Accuracy	Sensitivity	Specificity
USA	0.7922	0.77685	0.9	0.60784
Euro	0.8942	0.8285	0.8688	0.76666

The EURO dataset has higher accuracy compared USA. This is because some attributes like painloc, painexer which are highly informative for our classification are not present in the USA dataset. The sensitivity results are high for both the datasets implying the algorithm is very good at correctly identifying the presence of heart disease in a patient.

3.5 Gradient Boosting Classifier - GBT:

3.5.1 Introduction:

In this section, a few alternative and enhanced machine learning approaches are proposed for coronary heart disease prediction utilizing Boosting algorithms. Gradient boosting builds an additive model in a forward stage-wise fashion; allowing for the optimization of arbitrary differentiable loss functions. At each iteration, the pseudo-residuals/remaining errors are computed and a weak learner is fitted to these pseudo-residuals. Then, the contribution of the weak learner to the strong one is computed using a gradient descent optimization process. The computed contribution is the one minimizing the overall error of the strong learner.A decision tree model was constructed in order to serve as a baseline comparison to the ensemble models. Data was evaluated by using 10-fold cross validation and performance of the system is evaluated by classifiers accuracy, sensitivity and specificity to check the feasibility of our system.

3.5.2 Hyperparameter-Tuning:

The following parameters are tuned for our algorithm using scikit learn GridSearchCV.

Criterion: The default splitting criterion is to estimate gini index. However, entropy turned out to be better after performing grid search .
USA dataset : criterion = entropy EURO dataset: criterion = entropy

n_estimators: This parameter refers to number of trees at which boosting is terminated. USA dataset : n estimators = 30 EURO dataset: n estimators = 100

max_depth: This parameter minimizes the chances of overfitting the data. Hence, a smaller value is preferred. USA dataset : max depth = 3 EURO dataset: max depth = 3

min samples leaf: This parameter reduces overfitting of the data.
USA dataset : min samples leaf = 5 EURO dataset: min samples leaf = 5

subsample: This indicates the fraction of observations to be selected for each tree. Values slightly less than 1 makes the model robust by reducing the variance.
USA dataset : subsample = 0.9 EURO dataset: subsample = 0.9

3.5.3 Extreme Gradient Boosting - XGBoost:

XGBoost is an implementation of gradient boosted trees.The accuracy of XGBoost and GBT are almost same for this dataset, but XGBoost is a lot faster and memory-efficient. Hyper parameters values were set same as GBT.

3.5.4 Feature Selection:

Features with highest importance are shown in Fig:8 and Fig: 9.

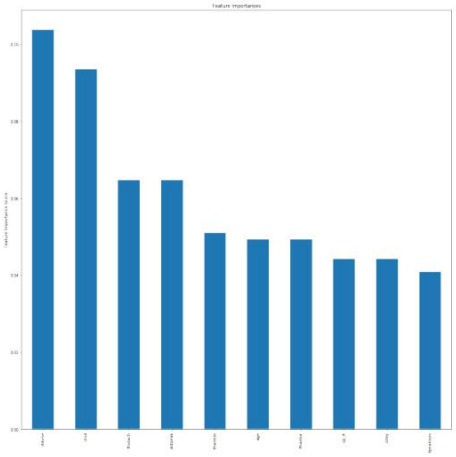

Figure 8: Most informative features for US dataset

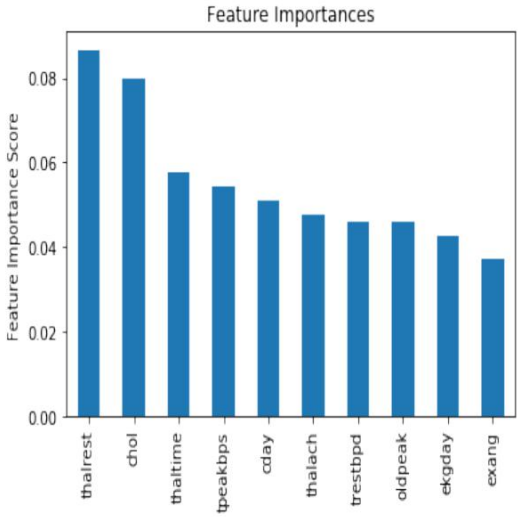

Figure 9: Most informative features for EURO dataset

3.5.5 Evaluation:

Table 5: Model Evaluation Metrics

Dataset	Decision Trees	Gradient Boosting	XGBoost
USA Training Accuracy	0.7182	0.9940	0.9881
USA Testing Accuracy	0.7310	0.8268	0.8268
Euro Training Accuracy	0.8041	0.9960	0.9862
Euro Testing Accuracy	0.7857	0.8888	0.8888
Sensitivity for US	0.7650	0.8162	0.8275
Specificity for US	0.6563	0.7758	0.7931
Sensitivity for Euro	0.7671	0.8939	0.8787
Specificity for Euro	0.8113	0.8833	0.8666

It is clear from model evaluation metrics that Gradient Boosting and XGBoost algorithms can outperform Decision Tree. Hence, these two classifiers are best suited for heart disease prediction as they produce larger gains in accuracy.

4 Conclusion:

In this project, we have reviewed the performance of 6 supervised learning models on heart disease dataset, a binary classification problem. We have calculated accuracy, sensitivity and specificity as evaluation metrics for all the classifiers to check the feasibility of our system. Based on the evaluation metrics, it is evident that ensemble methods outperform individual learning algorithms. Our results also show that boosting algorithms as the best choice in prediction of coronary heart

disease. The Feature selection methods we implemented such as RFE and feature importances has resulted in similar set of features as important for our prediction.Some of them are age,sex,cp(chest pain),slope(heart rate slope),exang(excercise induced angina),loc(location),proto(excercise protocol).

Table 6: Comparision of all models on USA dataset

Model	Test Accuracy	Sensitivity	Specificity
Naive Bayes	0.80991	0.81818	0.80519
Logistic Regression	0.7685	0.7346	0.7916
SVM	0.7685	0.7254	0.8000
Decision Tree	0.7310	0.7650	0.6563
Random Forest	0.77685	0.9	0.60784
Gradient Boosting	0.8268	0.8162	0.7758
XG Boost	0.8268	0.8275	0.7931

Table 7: Comparision of all models on EURO dataset

Model	Test Accuracy	Sensitivity	Specificity
Naive Bayes	0.8666	0.80392	0.92592
Logistic Regression	0.8761	0.8076	0.9433
SVM	0.8857	0.8235	0.9444
Decision Tree	0.7857	0.7671	0.8113
Random Forest	0.8285	0.8688	0.7666
Gradient Boosting	0.8888	0.8939	0.8833
XG Boost	0.8888	0.8787	0.8666

References

[1] Yeshvendra K. Singh,Nikhil Sinha,Sanjay K. Singh Heart Disease Prediction System Using Random Forest, Communi-cations in Computer and Information Science book series (CCIS, volume 721), 2017.

[2] Sumit Bhatia, Praveen Prakash, and G.N. Pillai SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features, Proceedings of the World Congress on Engineering and Computer Science 2008 WCECS 2008, October 22 - 24, 2008, San Francisco, USA