Heart Disease Prediction using Machine Learning Techniques

PART-I

I. Introduction

Conferring to the report of World Health Organisation, heart diseases cause millions global deaths per year. Heart disease diagnosis is complicated nonetheless critical task that is essentially being accomplished precisely and proficiently. This task is frequently made on the understanding and acquaintance of doctor. This causes excessive time and cost. It is essential to have a framework that can effectually recognise the prevalence of heart disease in thousands of samples instantaneously. It is critical to recognise contrivances proficient of producing high accuracy of prediction in heart diseases. State-of-the-art data mining approaches are applied to discern knowledge from clinical data for research in medical informatics, essentially in heart disease prophecy.

II. Dataset

StatLog Heart Disease dataset available at UCI Machine Learning Laboratory was used in this study. This dataset consists of 270 samples with 150 samples without heart disease (absence) and 120 samples with heart disease (presence).

A. Features

13 distinct parameters have been taken into account such as:

1. Age.

2. Sex.

3. Chest pain type (four values).

4. Resting blood pressure.

5. Serum cholesterol in mg/dl.

6. Fasting blood sugar [120 mg/dl

7. Resting electrocardiographic results (values 0, 1 and 2).

8. Maximum heart rate achieved.

9. Angina induced by Exercise.

10. Peak Old = ST depression tempted by workout comparative to rest.

11. Slant of the peak exercise ST segment.
12. Numeral of major vessels(0–3) coloured by fluoroscopy.
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect.

III. Techniques

A. Artificial Neural Network(ANN)- MLP

An MLP is a class of feed-forward artificial neural network. Multi-layer means here it should contains atleast three layers. In each layer, we can have any number of neurons and each neuron is a fully connected to previous layer neurons. MLP can learn any non-linear function by choosing appropriate activation function.

B. Support Vector Machines (SVM)

It is a discriminative and supervised model which tries to increase the classification confidence by maximizing the margin. It also has a special ability to classify non-linearly separable data by using kernel functions.

C. Naive Bayes (NB)

Naïve Bayes is a generative and supervised model. It uses joint probability to classify the data points. It calculates the prior and likelihood values of the dataset and classifies the samples according to posterior value with respect to class (category).

D. Logistic Regression

Logistic Regression is a supervised classification technique which learns the hyper plane to separate classes. For this the data should be linearly separable and it uses squishes the objective function using sigmoid function. So that the values will be between ‘0’ to ‘1’. We

generally observe this output as probability of the class.

E. K-Nearest Neighbours (KNN)

It is neighbourhood based supervised technique. In this, each test point is classified based on the similarity of the k- nearest points using majority votes. For odd number of classes use k=even number and for even number of classes use k=odd number to avoid ambiguity.

F. Decision Tree

It is a decision based classifier which uses tree like model. Each node of the tree has chosen based on the 'information gain' or 'gini' impurity.

IV. Results

Authors have done the classification by selecting the features and done the normalization and compared with all other models.

Paper results:

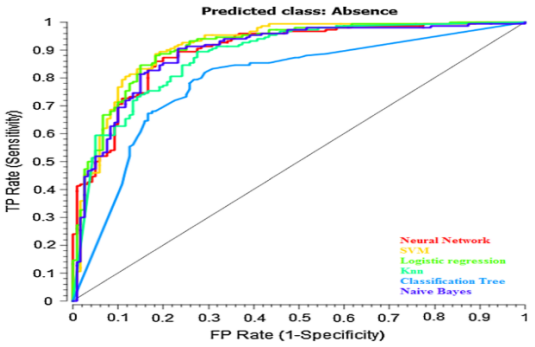
	CA	Sens.	Spec.	Pre.	NPV	FPR	RMC
ANN	0.84	0.87	0.79	0.84	0.83	0.21	0.16
SVM	0.82	0.77	0.89	0.90	0.75	0.11	0.18
Logistic regression	0.85	0.89	0.81	0.85	0.85	0.19	0.15
kNN	0.80	0.84	0.76	0.81	0.79	0.24	0.20
Classification tree	0.77	0.79	0.73	0.79	0.74	0.27	0.23
Naive Bayes	0.83	0.85	0.80	0.84	0.81	0.20	0.17

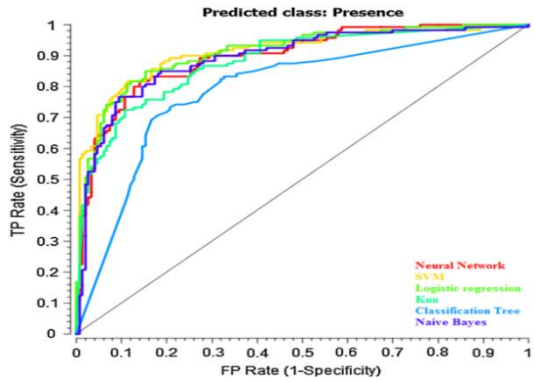
Classification accuracy (*CA*), sensitivity (*Sens.*), specificity (*Spec.*), precision (*Pre.*), negative predi value (*NPV*), false positive rate (*FPR*), rate of misclassification (*RMC*), F1 measure (*F1*)

Our results:

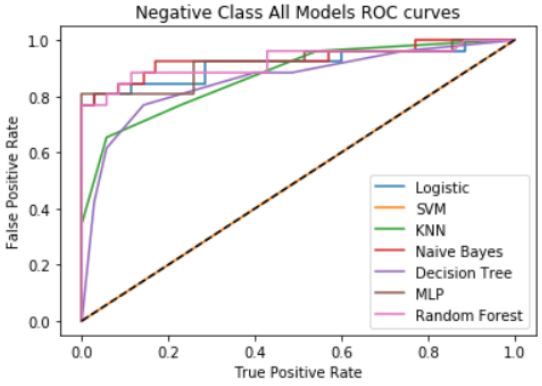
Metrics ↓	Log regr	SVM	KNN	NB	Dec Tree	ANN	Rand Forest
Accuracy	0.9016	0.918	0.8689	0.8689	0.8197	0.8525	0.8852
Specificity	0.8077	0.8077	0.8077	0.8846	0.7692	0.8077	0.8462
Sensitivity	0.9714	1.0	0.9143	0.8571	0.8571	0.8857	0.9143
Precision	0.8718	0.875	0.8649	0.9091	0.8333	0.8611	0.8889
NPV	0.9545	1.0	0.875	0.8214	0.8	0.84	0.88
FPR	0.1923	0.1923	0.1923	0.1154	0.2308	0.1923	0.1538
RMC	0.0984	0.082	0.1311	0.1311	0.1803	0.1475	0.1148
F1 score	0.9189	0.9333	0.8889	0.8824	0.8451	0.8732	0.9014

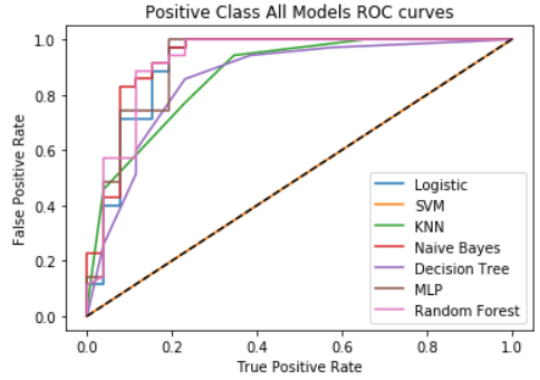
ROC plots:





Our plots:





Logistic Regression:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	21	5
No-Disease	1	34

ANN:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	21	5
No-Disease	4	31

SVM:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	21	5
No-Disease	0	35

Random Forest:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	21	5
No-Disease	4	31

Naïve Bayes:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	23	3
No-Disease	5	30

KNN:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	21	5
No-Disease	3	32

Decision Tree:

Confusion matrix :

Actual labels ↓	Predicted labels → Disease	No-Disease
Disease	20	6
No-Disease	5	30

V. Conclusion

The highest classification accuracy was given using Logistic Regression which was of 85% with sensitivity and specificity of 89 and 81%, respectively. The second highest classification accuracy is achieved by ANN (84%). Additionally, these two methods have shown utmost sensitivity of 89 and 87% correspondingly in the paper.

SVM achieves highest specificity of 89% which indicates that this classifier is most suitable for identification of patients with heart disease (presence class). Moreover, SVM also has the highest precision of 90%.

PART-II

I. Objective

Heart diseases are major health issues worldwide. No of patients with heart diseases are growing due to bad lifestyle and lack of health awareness. So it is important to have a model which can effectively predict the occurrence of heart diseases. The major machine learning techniques were evaluated for heart disease prediction problem. To analyse these methods we used various classification evaluation metrics and in addition we also assessed on receiver operative characteristics curve.

II. Analysis & Improvements

Authors of the paper used a train test split of 60% train and 40% to evaluate the machine learning techniques. There are only around 300 data points and taking the split of 60% for training with 13 features will lead to curse of dimensionality. So we changed the train and test split to 80% and 20% which resulted in a better generalization accuracy and better F1 score. We also better hyper-parameter tuned each model which resulted in an improvement of each model when compared to the base paper results.

III. Conclusion

The highest classification accuracy of 91% was reported using SVM with sensitivity and specificity of 100% and 81%, respectively. The second highest classification accuracy is achieved by Naïve Bayes (90%).

Naive Bayes achieves highest specificity of 88% which indicates that this classifier is most suitable for identification of patients with heart disease (presence class). Moreover, Naive Bayes also has the highest precision of 91%.

Analyzing the ROC Curves the MLP and Random Forest are the best and most generalized models. The other metrics also supports the fact that these can be used effectively.

Decision Tree turns out to the worst of all from the ROC.

PART-III

Docker instructions for our Project

Commands:

#Login to docker:

>docker login

#to pull docker

>docker pull dhruv035/dhruv_mt18035_assgn_2

#to run docker

>docker run --name=ubuntu -itd ubuntu:18.04

to create local container for docker image

>docker exec -it ubuntu /bin/bash

to run our project

>python3 BDMH_project.py