# Performance Evaluation of Machine Learning Techniques for Prediction of Heart Disease

Dhruv Kaushik MT18037 K. Srivatsava MT18054 Gurpreet Singh MT18098

Sk. Subhani MT18117 P. Akhil Kumar MT18130

## 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.

## 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).

🔗 Dataset

## Features

13 distinct parameters have been taken into account such as:

- Age
- Sex
- Chest pain type (four values)
- Resting blood pressure
- Serum cholesterol in mg/dl
- Fasting blood sugar > 120 mg/dl
- Resting electrocardiographic results (values 0, 1 and 2)
- Maximum heart rate achieved
- Angina induced by Exercise
- Peak Old = ST depression tempted by workout comparative to rest
- Slant of the peak exercise ST segment
- Numeral of major vessels (0–3) coloured by fluoroscopy
- Thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

## Methods Used

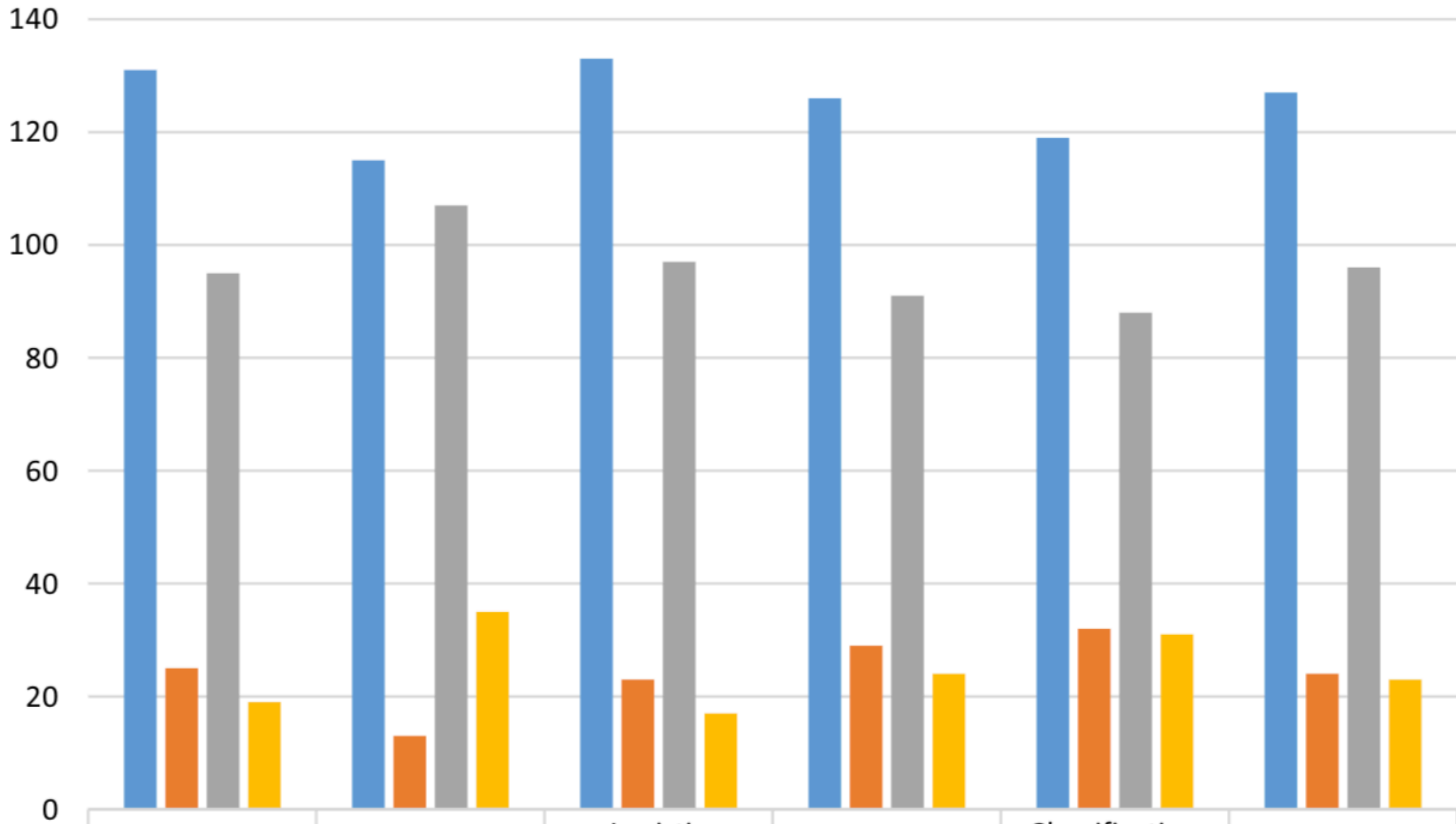
- Artificial Neural Network - A Multi Layer Perceptron is used in this case. An MLP, as it is called is a class of neural networks which is organised in layers like a multistage directed graph. Every node at every layer acquires an input from the linked node of preceding layer, and then it calculates value of a function and delivers input to the associated node in the following layer.
- Support Vector Machines - It is is a technique of supervised machine learning used for classification and regression. Here, SVM was used for binary classification having two categories absence and presence of heart disease, for $y_i = +1, -1$ correspondingly. The SVM classifiers examine for the optimum extrication hyperplane which in between from the two classes.
- Naive Bayes - One of the best operative classifiers is Bayesian network. The BN construction is directed acyclic graph wherein nodes resemble to province variables and edges among nodes resemble to dependences between variables. Naive Bayes classifiers are techniques of finding the suitable classification.
- Logistic Regression - Logistic regression is a discriminative classification technique that works on real-valued input vector. The measurement of input vector, to be categorised, is known as features or predictors.
- K-Nearest Neighbours - It is an instance-based learning technique that does not frame a comprehensive theoretical model commencing the training examples. As an alternative, it uses the somewhat simpler conception and the instance nearby in the input space is probably to fit in the similar class.
- Decision Tree - It is a method that characterises the data in the form of n-ary hierarchy through every node, and branch has a definite accompanying outcome, probability and weights. Root of the tree is designated normally by computing the entropy or by its converse, the information gained.

## Classification Performance Measurements

- True Positive (TP) - no. of samples with absence of heart disease predicted as absence of heart disease.
- False Positive (FP) - no. of samples with presence of heart disease predicted as absence of heart disease.
- True Negative (TN) - no. of samples with presence of heart disease predicted as presence of heart disease.
- False Negative (FN) - no. of samples with absence of heart disease predicted as presence of heart disease.
- Accuracy - $$\frac{TP+TN}{(TP+FP+FN+TN)}$$
- Sensitivity - $$\frac{TP}{(TP+FN)}$$
- Specificity - $$\frac{TP}{(TN+FP)}$$
- Precision - $$\frac{TP}{(TP+FP)}$$
- Recall - $$\frac{TP}{(TP+FN)}$$
- $F_1$ Measure - $$\frac{2*Precision*Recall}{(Precision+Recall)}$$

Data samples were partitioned in tenfold, every fold was applied in testing, and remaining folds were used as training during cross validation.

Classification Results using Machine Learning




| | ANN | SVM | Logistic Regression | kNN | Classification Tree | Naïve Bayes |
|---|---|---|---|---|---|---|
| TP | 131 | 115 | 133 | 126 | 119 | 127 |
| FP | 25 | 13 | 23 | 29 | 32 | 24 |
| TN | 95 | 107 | 97 | 91 | 88 | 96 |
| FN | 19 | 35 | 17 | 24 | 31 | 23 |

ANN

| | Predicted class | | |
|---|---|---|---|
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 131 (84.0 %) | 19 (16.7 %) | 150 |
| Presence | 25 (16.0 %) | 95 (83.3 %) | 120 |
| Total predicted | 156 | 114 | 270 |

Hidden layer neurons: 10, regularization factor: 1.5, max iterations: 2000

**ANN**

kNN

| | Predicted class | | |
|---|---|---|---|
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 126 (81.3 %) | 24 (20.9 %) | 150 |
| Presence | 29 (18.7 %) | 91 (79.1 %) | 120 |
| Total predicted | 155 | 115 | 270 |

Learning parameters—metrics: euclidean, continuous attributes: normalized, unknown values ignored: no, number of neighbors: 5, weighting: by distances

**K-NEAREST NEIGHBOURS**

SVM

| | Predicted class | | |
|---|---|---|---|
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 115 (89.8 %) | 35 (24.6 %) | 150 |
| Presence | 13 (14.7 %) | 107 (85.1 %) | 120 |
| Total predicted | 128 | 142 | 270 |

Learning parameters—kernel: RBF, $e-^{0.0050*(x-y).\ (x-y)}$, cost (C): 0.1, numeric precision: 0.0005, estimation of class probabilities: yes, normalization of data: yes

**SUPPORT VECTOR MACHINES**

Naïve Bayes

| | Predicted class | | |
|---|---|---|---|
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 127 (84.1 %) | 23 (19.1 %) | 150 |
| Presence | 24 (15.9 %) | 96 (80.7 %) | 120 |
| Total predicted | 151 | 119 | 270 |

Learning parameters—probability estimation: relative frequency, LOESS window size: 0.5, number of points in LOESS: 100, adjust classification threshold: no

**NAIVE BAYES**

| Logistic regression | | | |
|---|---|---|---|
| | Predicted class | | |
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 133 (85.3 %) | 17 (14.9 %) | 150 |
| Presence | 23 (14.7 %) | 97 (85.1 %) | 120 |
| Total predicted | 156 | 114 | 270 |

Learning parameters—training error cost (C): 1.0, regularization-type: L1 (absolute weights), normalization: yes

**LOGISTIC REGRESSION**

| Classification trees | | | |
|---|---|---|---|
| | Predicted class | | |
| | Absence | Presence | Actual total |
| Actual class | | | |
| Absence | 119 (78.8 %) | 31 (26.1 %) | 150 |
| Presence | 32 (21.2 %) | 88 (73.9 %) | 120 |
| Total predicted | 151 | 119 | 270 |

Learning parameters—attribute selection: information gain, binarization: no binarization, pruning: two instances in leaves, recursively merge leaves with same majority class: yes, pruning with m-estimate: $m = 2$

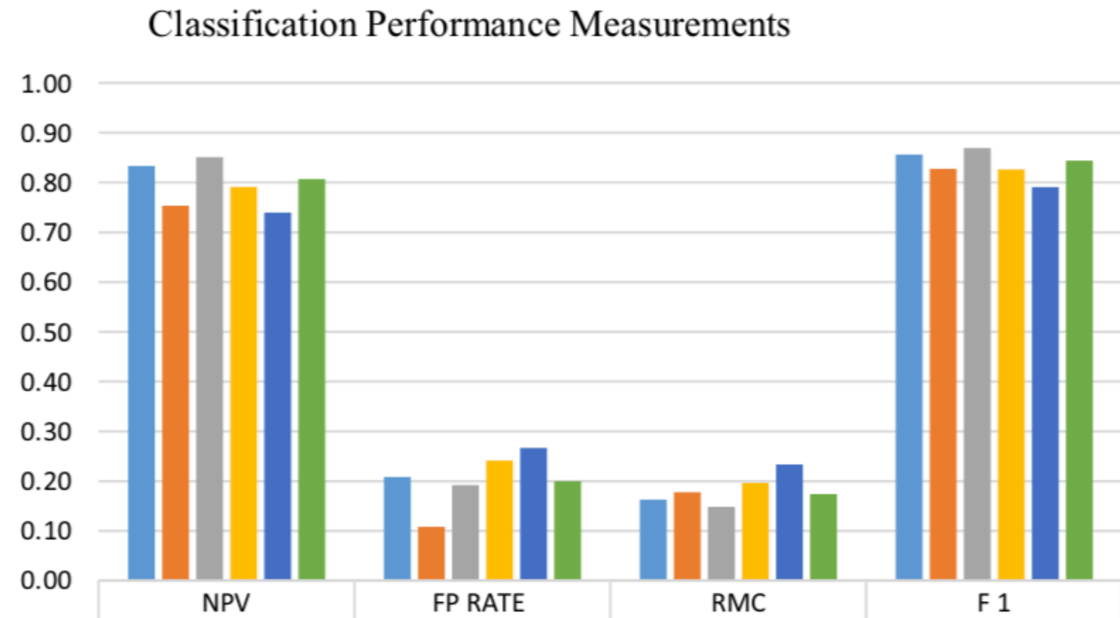
**DECISION TREES**

**Table 7** Classification performance measure indices for tenfold cross validation using machine learning techniques

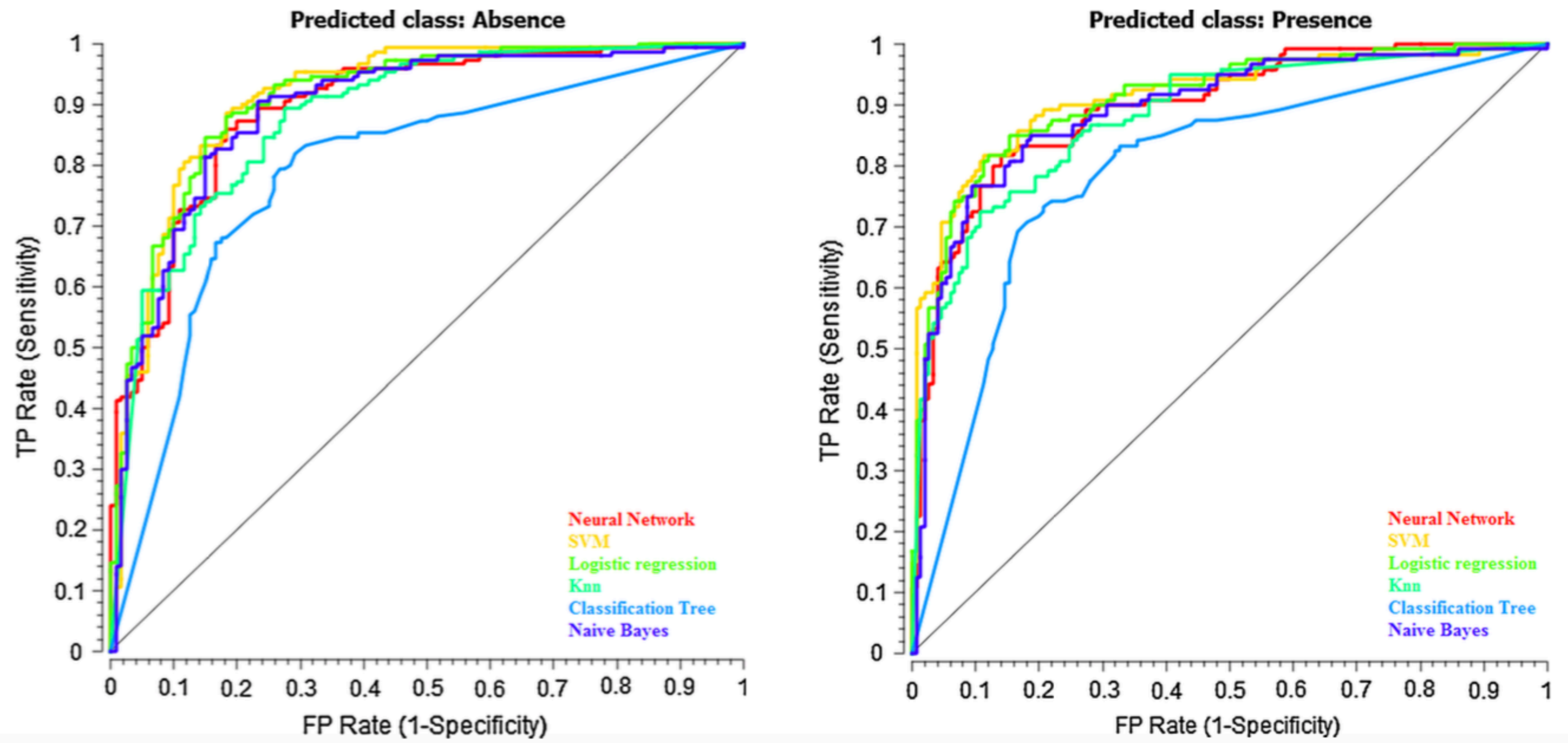
| | CA | Sens. | Spec. | Pre. | NPV | FPR | RMC | F1 |
|---|---|---|---|---|---|---|---|---|
| ANN | 0.84 | 0.87 | 0.79 | 0.84 | 0.83 | 0.21 | 0.16 | 0.86 |
| SVM | 0.82 | 0.77 | 0.89 | 0.90 | 0.75 | 0.11 | 0.18 | 0.83 |
| Logistic regression | 0.85 | 0.89 | 0.81 | 0.85 | 0.85 | 0.19 | 0.15 | 0.87 |
| kNN | 0.80 | 0.84 | 0.76 | 0.81 | 0.79 | 0.24 | 0.20 | 0.83 |
| Classification tree | 0.77 | 0.79 | 0.73 | 0.79 | 0.74 | 0.27 | 0.23 | 0.79 |
| Naive Bayes | 0.83 | 0.85 | 0.80 | 0.84 | 0.81 | 0.20 | 0.17 | 0.84 |

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

**Fig. 5** Classification performance measurements of machine learning algorithm




| | NPV | FP RATE | RMC | F 1 |
|---|---|---|---|---|
| ANN | 0.83 | 0.21 | 0.16 | 0.86 |
| SVM | 0.75 | 0.11 | 0.18 | 0.83 |
| kNN | 0.85 | 0.19 | 0.15 | 0.87 |
| Naïve Bayes | 0.79 | 0.24 | 0.20 | 0.83 |
| Logistic Regression | 0.74 | 0.27 | 0.23 | 0.79 |
| Classification Tree | 0.81 | 0.20 | 0.17 | 0.84 |



## Conclusion

- The **highest classification accurac**y of 85% was reported using Logistic Regression 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.
- 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%.