Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques

Group: 22

- Aditya Mittal
- Dikscha Sapra
- Deepti Batra
- Meenakshi Maindola
- Sharat Agarwal

Introduction and Motivation

- Lot of data in healthcare industry which has not been yet utilised to the full extent.
- Heart disease prediction is a very important topic in this field and can be used to correctly predict whether a person will or will not have a certain heart disease.
- The dataset used for this is UCI Machine Learning Repository : Heart Disease Database and have combined two sub- databases to create and access the accuracy.
- Three techniques were applied in the paper : Decision Trees, Naive Bayes and Neural Networks. We applied two more techniques namely, SVM and Logistic Regression.

Information Gain in Decision Tree Induction

- Assume that using attribute A a set S will be partitioned into sets $\{S_1, S_2, \ldots, S_v\}$
  - If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtrees S_i is

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

- The encoding information that would be gained by branching on A $\quad Gain(A) = I(p,n) - E(A)$

ID3 algorithm

- Split (node, {examples}):
  1. A ← the best attribute for splitting the {examples}
  2. Decision attribute for this node ← A
  3. For each value of A, create new child node
  4. Split training {examples} to child nodes
  5. For each child node / subset:

     if subset is pure: STOP

     else: Split (child_node, {subset})

Dataset Description

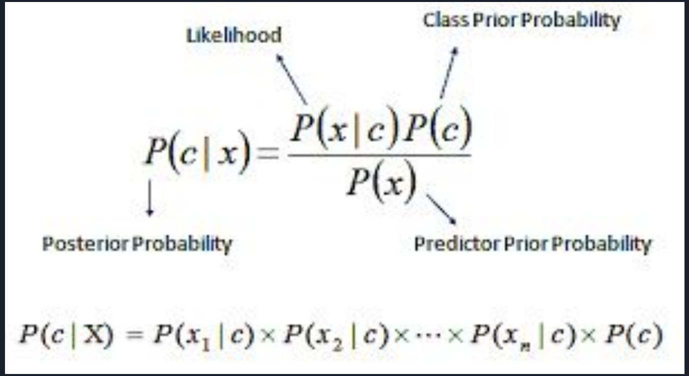
- Cleveland Heart Disease database.
- Statlog Heart Disease database.
- Each record has values for 13 different input attributes.

Sr. no	Attribute	Description
1	age	Age in years
2	sex	Male or female
3	cp	Chest pain type
4	thestbps	Resting blood pressure

5	chol	Serum cholesterol
6	restecg	Resting electrographic results (ECG)
7	fbs	Fasting blood sugar
8	thalach	Maximum heart rate achieved
9	exang	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	Slope of the peak exercise ST segment
12	ca	Number of major vessels colored by fluoroscopy
13	thal	Defect type

Naive Bayes

- Supervised Machine Learning Algorithm.
- $P(c|x)$ is Posterior probability of a class given feature.
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is likelihood i.e. probability of feature given class.
- $P(x)$ is evidence i.e. probability of feature.
- $P(c_1|x) > P(c_2|x)$ (Class 1, Heart Disease)
- $P(c_1|x) < P(c_2|x)$ (Class 1, No Heart Disease)

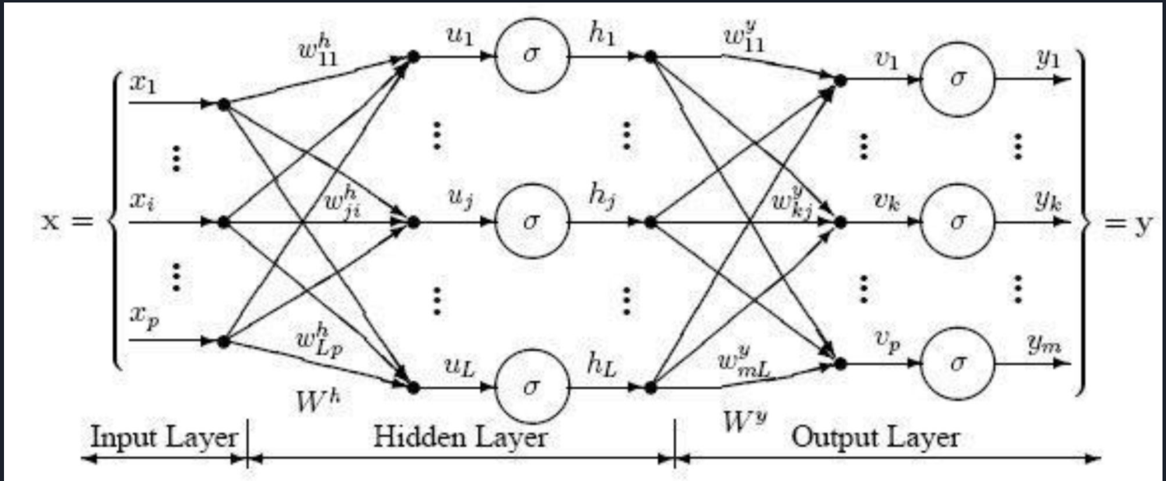


Decision Trees:

- Supervised machine learning Algorithm
- Types of nodes: Decision nodes, leaves
- Best Algorithm for Decision Tree: Iterative Dichotomiser 3 (ID3)
- J48 is implementation of ID3 developed by weka.

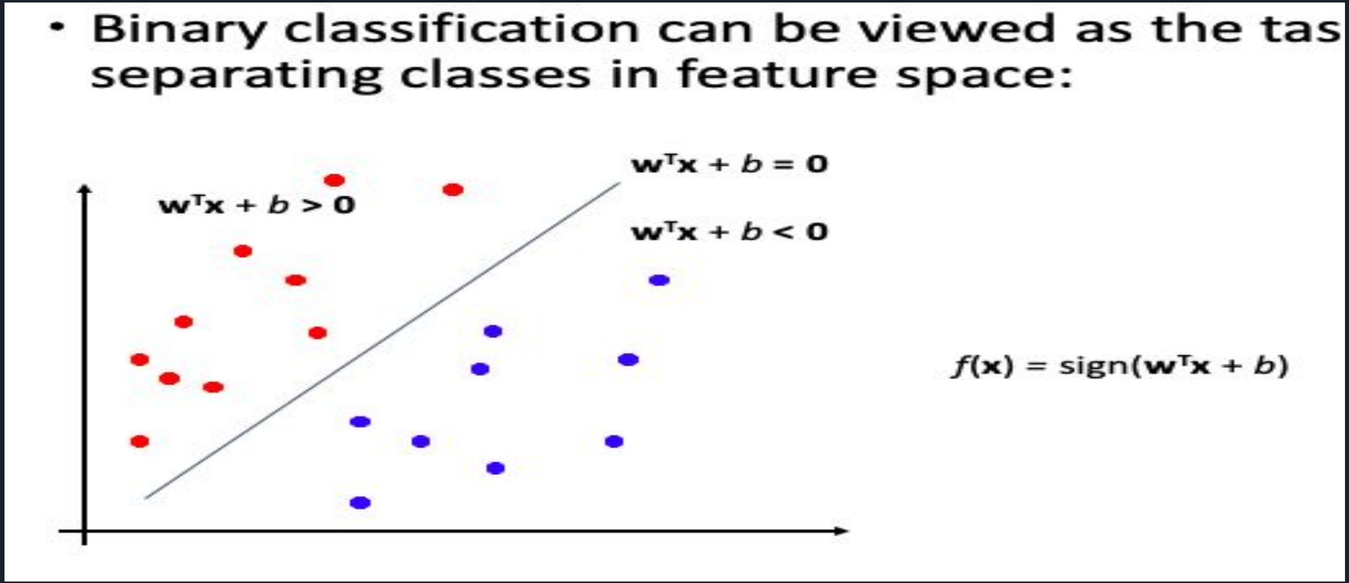
Neural Network

- Multilayer Perceptron neural network is used.
- Network structure:
  - 3 hidden layers, one input and output layer.
  - Sigmoid function on output layer.
- Y > 0.5 = 1 (Positive class, heart disease)
- Y < 0.5 = 0 (Negative class, no heart disease)



SVM

Idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable



Logistic Regression

The logistic distribution constrains the estimated probabilities to lie between 0 and 1.

*The estimated probability is:

$$p = 1/[1 + \exp(-a - b\,X)]$$

*if you let $a + b\,X = 0$, then $p = .50$

*as $a + b\,X$ gets really big, p approaches 1

*as $a + b\,X$ gets really small, p approaches 0

RESULTS for Naive Bayes:

Results produced by paper:

Confusion matrix for Naive Bayes:

	a	b
a	110	5
b	10	145

Results produced by us:

	a	b
a	119	13
b	10	128

RESULTS for Decision Trees :

Results produced by paper:

Confusion matrix for Decision Trees:

	a	b
a	123	4
b	5	138

Results produced by us:

	a	b
a	122	8
b	10	130

RESULTS for Neural Networks:

Results produced by paper:

Confusion matrix for Neural Networks:

	a	b
a	117	0
b	2	151

Results produced by us:

	a	b
a	120	5
b	6	139

RESULTS for SVM :

Results produced by us:

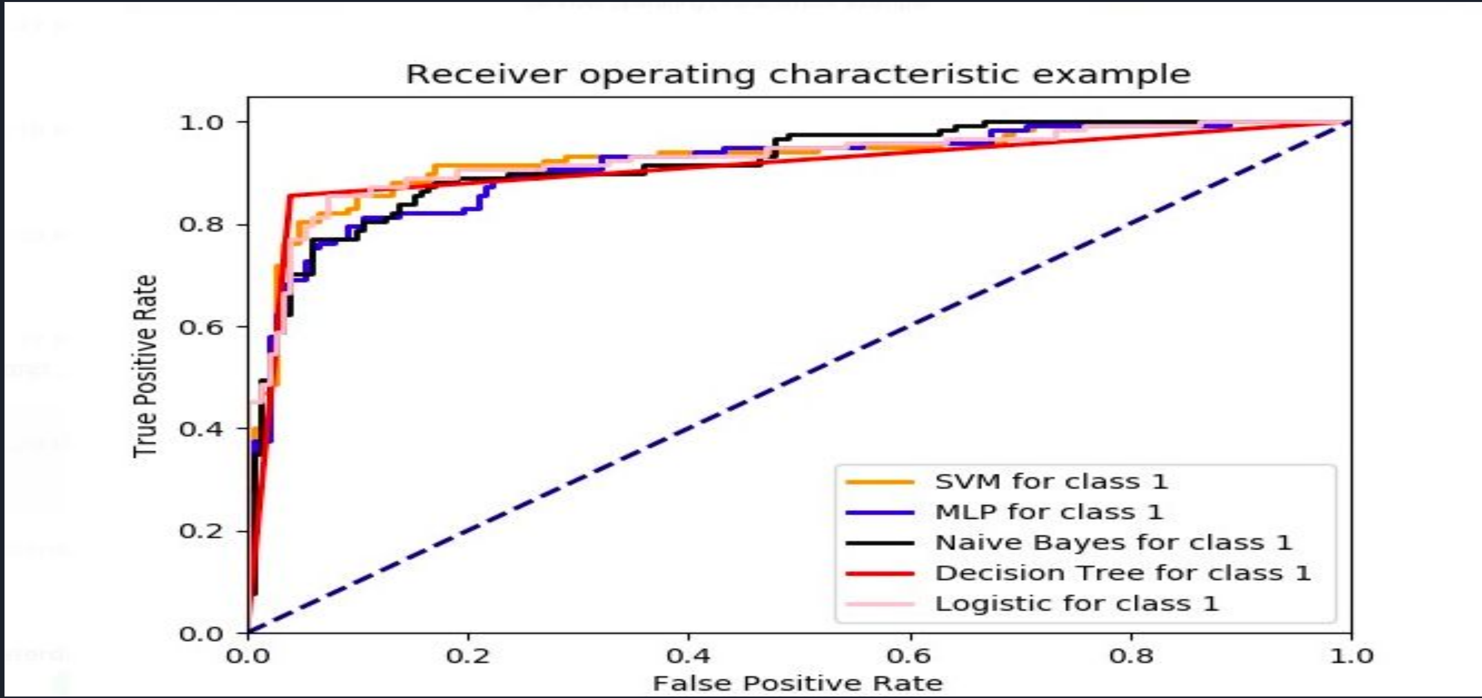
	a	b
a	125	7
b	12	126

RESULTS for Logistic Regression:

Results produced by us:

	a	b
a	123	4
b	18	125

AUC-ROC Curve:



Performance Comparison:

Metric	Our Approach	Paper's Approach
Naive Bayes	91.4%	94.44
Decision Trees	93%	96.66
Neural Networks	95.9%	99.25
SVM	92.9%	-
Logistic Regression	91.8%	-

Conclusion:

- Accuracy achieved in case of Neural Networks was maximum i.e. 95.9%.
- Apart from replicating paper, we tried using two other techniques: SVM and Logistic Regression.
- Also, 20 folds cross-validation was performed .