IDA PROJECT REPORT

Topic : Probabilistic classification using statistical Naïve Bayes classifier

Group 7

Group Members:

NESANURU PREM KUMAR REDDY	S20190010127
JUPURU HARSHITH	S20190010072
CHANDA VARUN REDDY	S20190010027
JEEPALYAM MOHITH	S20190010071
GUDURU CHETAN	S20190010057

Date of Submission: 30/11/2021
Dataset Used: Heart Disease Dataset

Problem Statement:

- Split the dataset into a training and testing set appropriately. Obtain the appropriate contingency table from a training data set comprising the prior and posterior probabilities.
- Test the classifiers using k-fold cross validation technique. Run with different values of k and then choose the optimum result.
- Furnish the accuracy using an appropriate confusion matrix and report the performance evaluation with a different matrix (e.g., Precision, Recall, F1 score, etc.).

I. Understanding the theory to solve the project problem:

A. Dataset: The given dataset is a dataset from 1988.The target field is an integer with values 0 or 1 which concludes a patient has any heart disease or not.

B. Definition: Naive Bayes classifier is a supervised classification technique. It works on the principle "If it walks like a duck, quacks like a duck, then it is probably a duck". It uses Bayes theorem of probability to perform probabilistic prediction.

C. Why choose Naive Bayes:

1. Heart disease classification involves identifying healthy or sick individuals. Linear Classifier such as Naive Bayes is relatively stable with respect to small variation or changes in training data.
2. Naive Bayes is fast and hence real time predictions can be made.

D. Assumptions: All variables in the dataset are not correlated to each other.

E. Bayes Theorem: Naive Bayes classifier uses Bayes Theorem.

Let E1, E2, ... , En be n mutually exclusive and exhaustive events associated with a random experiment. If A is any event which occurs with E1 or E2 or … or En, then

$$P(E_i|A) \ = \frac{P(E_i).P(A|E_i)}{\sum_{i=1}^{n} P(E_i).P(A|E_i)}$$

F. Prior and Posterior probabilities: If P(A) and P(B) are the prior probabilities, then posterior probabilities are calculated as P(A|B) and P(B|A).

G. Algorithm: Given a set of k mutually exclusive and exhaustive classes Y = $\{ y_1, \ y_2, \ ..., \ y_k\}$, which have prior probabilities P(Y_1), P(Y_2),..,P(Y_k).

There is an n-attribute set X = $\{X_1, X_2, ..., X_n\}$, which for a given instance have values $X_1 = x_1$, $X_2 = x_2$, $X_n = x_n$.

For each $y_i \in$ Y, calculate the posterior probabilities, i=1, 2, …, k.

$$P(Y = y_j \mid (X_1 = x_1)\ AND\ (X_2 = x_2)\ AND\ ...\ (X_n = x_n))$$

$$= P(Y_j)\ X \prod_{i=1}^{n} P(X_i = x_i \mid Y_j)$$

We classify the given test value with the maximum posterior probability from the above formula.

predicted class = argmax{ P(Y=yi) }

Y_j is the output.

H. Some examples where Naive Bayes is used:

1. Medical data classification
2. Real time predictions
3. Spam filtering
4. Sentimental analysis

II. Implementation of the project:

A. Importing libraries:

1. **caTools and ROCR:** It is a library used for ROC curves.
2. **Caret:** It is a library used for data splitting, preprocessing.
3. **E1071:** It is a library used for statistics and some probability formulae.
4. **ModelMetrics:** It is a library used for root mean square error.
5. **Gpairs and ggplot2:** It is a library used for plotting the data.

B. Loading the dataset

The dataset contains 1025 rows with 14 attributes where 13 are defining the data and the 14th attribute is target label “target”.

C. Data Cleaning: All the null values from the dataset are removed.

D. Categorical and continuous variables are separated into two lists

Categorical attributes are :

("sex","cp","fbs","restecg","exang","slope","ca","thal")

Sex : 0 or 1

cp (chest pain) : 0 or 1 or 2 or 3

fbs (fasting blood pressure) : 0 or 1

restecg (resting electrocardiographic results) : 0 or 1 or 2

exang (exercise induced angina) : 0 or 1

slope (the slope of the peak exercise ST segment) : 0 or 1 or 2

ca (number of major vessels colored by fluoroscopy) : 0 or 1 or 2 or 3

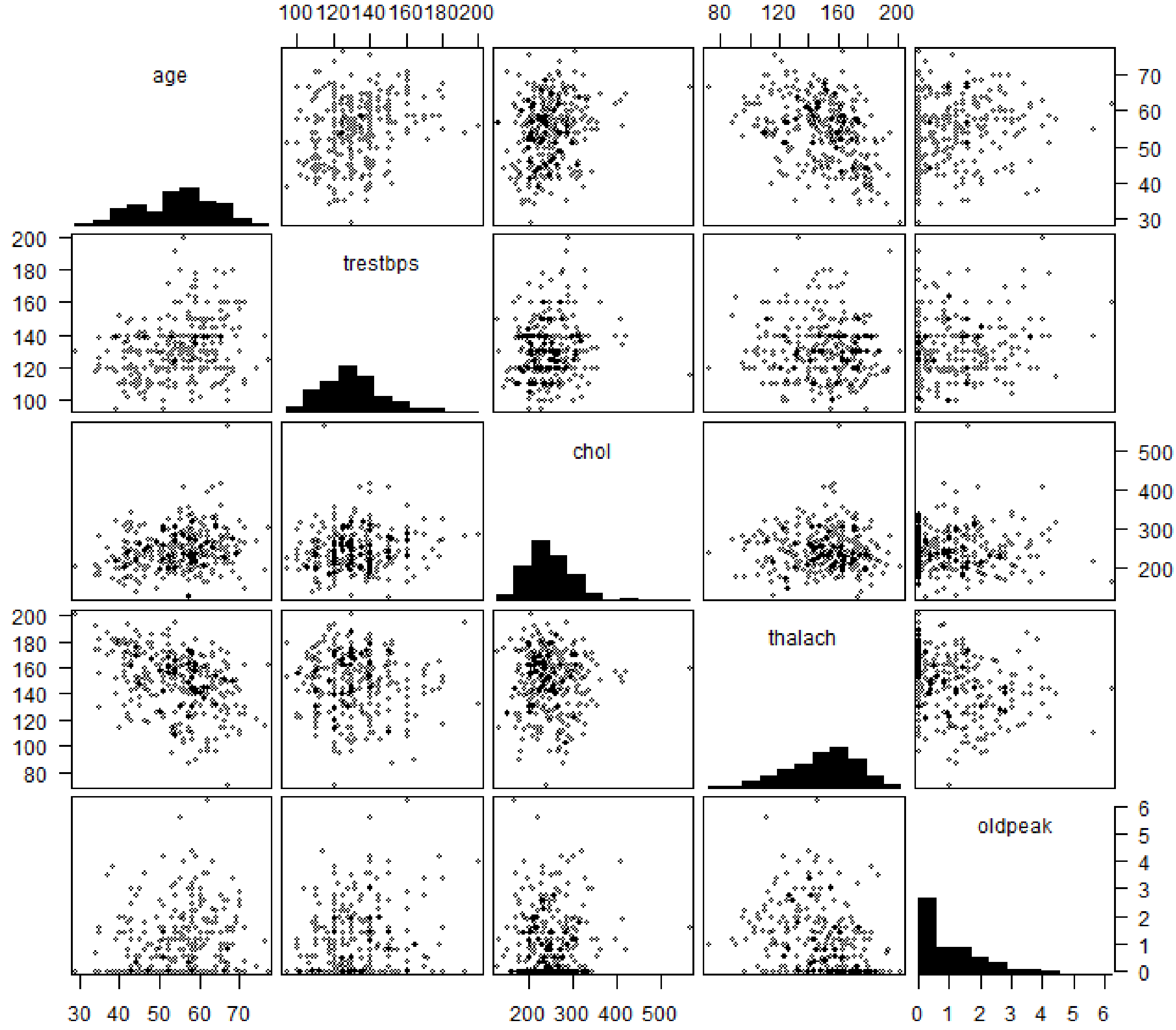
thal : 0 or 1 or 2 or 3

Continuous attributes are :

("age","trestbps","chol","thalach","oldpeak")

(age, resting blood pressure, serum cholesterol, maximum heart rate achieved, ST depression induced by exercise relative to rest)

E. Gpairs:



F. Kfold:

1. Here we assume minimum K as 3 and maximum K as 15.
2. In a for loop we send the dataset and values of K from 3 to 15 to the function kFoldCrossVal.
3. For each value of K, the kFoldCrossVal function splits the train and test data and returns the mean accuracy.
4. These mean accuracies are appended to the list avgAcc.
5. From the avgAcc list we will get the best k value.

G. Predicting with Naive Bayes classifier on total data:

1. First the train and test data are split in the ratio 3:1.

2. Then the train data is used to draw the contingency table comprising prior and posterior probabilities of categorical attributes and target attribute.
3. Train and test data are then sent into the function predict.

H. Naive bayes:

1. First we need to find the likelihood of categorical and continuous attributes using the train data.
2. We will find the likelihood of the categorical attributes using the formula:

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \cap Y = y_j)}{P(Y = y_j)}$$

Simply the number of rows having X = x_i and Y = y_j divided by the number of rows having Y = y_j.

3. We calculate the mean and standard deviation of the train data to find the likelihood.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma = \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right]^{\frac{1}{2}}$$

4. From the mean and standard deviation values obtained we use the normal distribution formula to find the likelihood of the test data.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

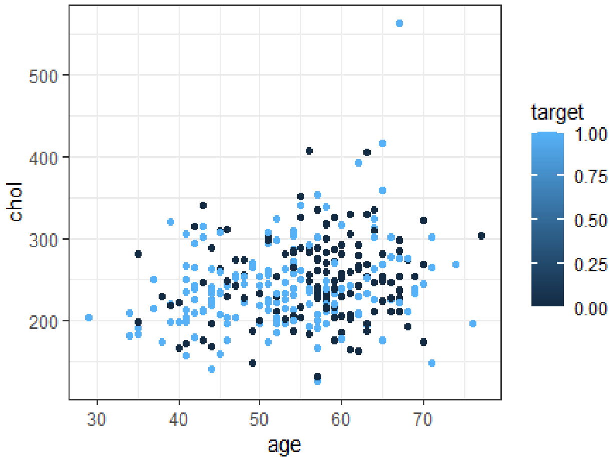
5. We go through each row of the test data and we predict the target value.

$$P(Y = y_j \mid (X_1 = x_1)\ AND\ (X_2 = x_2)\ AND \dots (X_n = x_n))$$

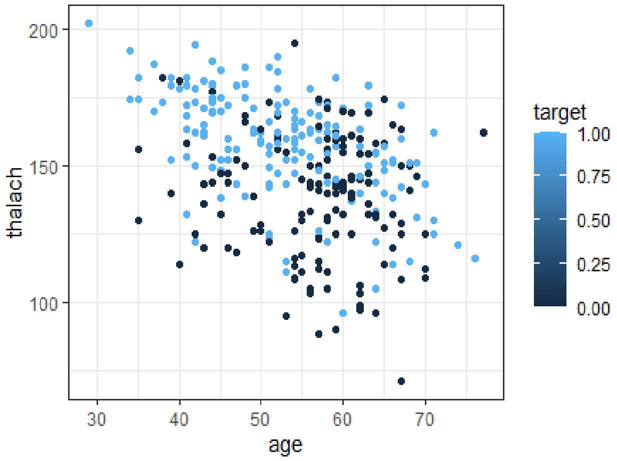
$$= P(Y_j)\ X \prod_{i=1}^{n} P(X_i = x_i \mid Y_j)$$

III. Data Visualization:

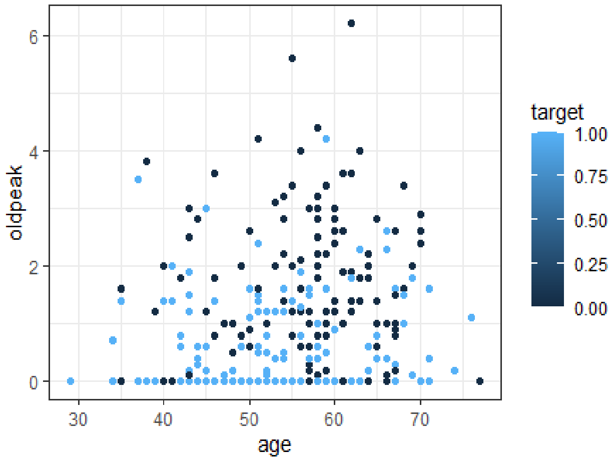
Scatterplot age V/S chol



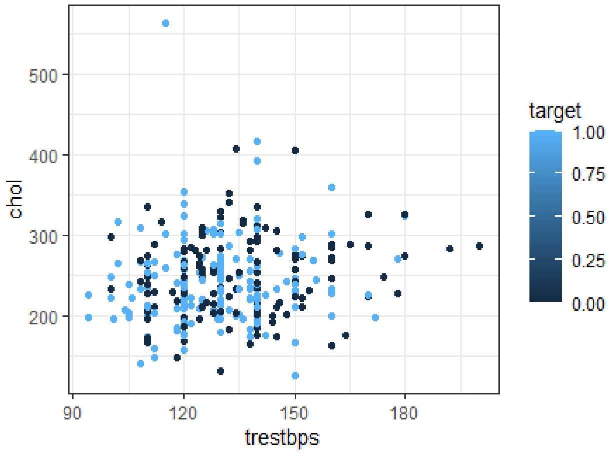
Scatterplot age V/S thalach



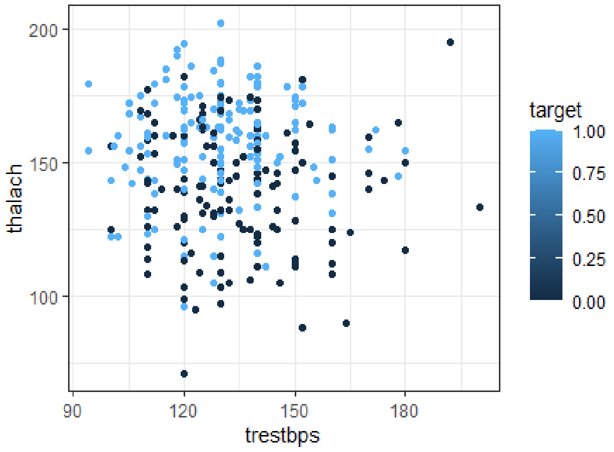
Scatterplot age V/S oldpeak



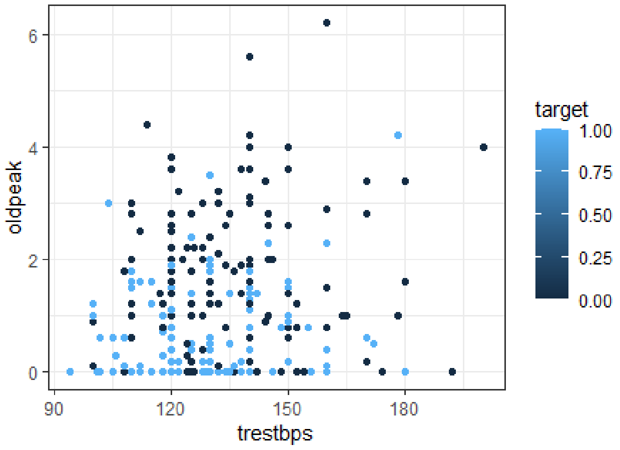
Scatterplot trestbps V/S chol

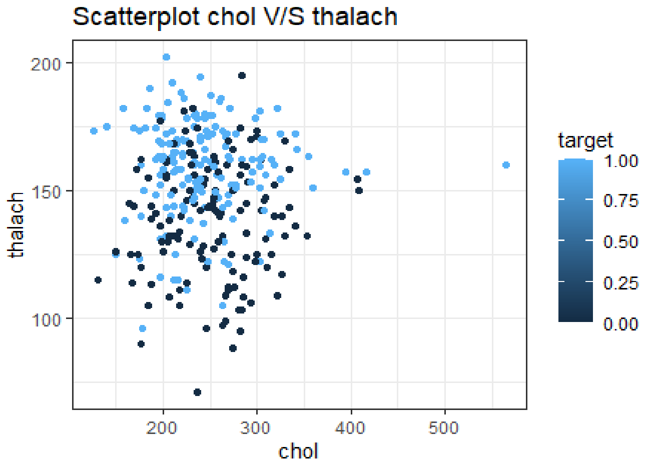


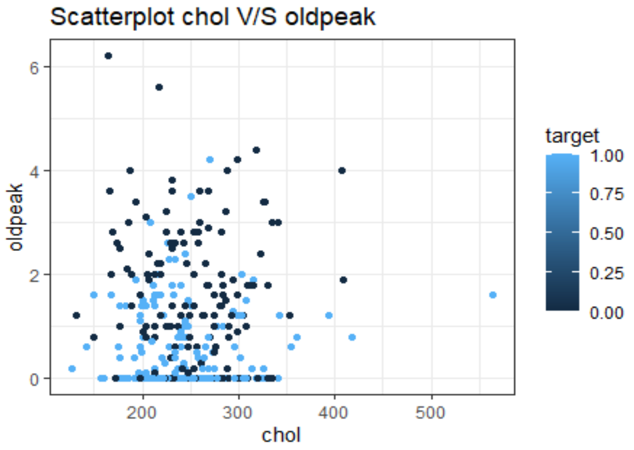
Scatterplot trestbps V/S thalach

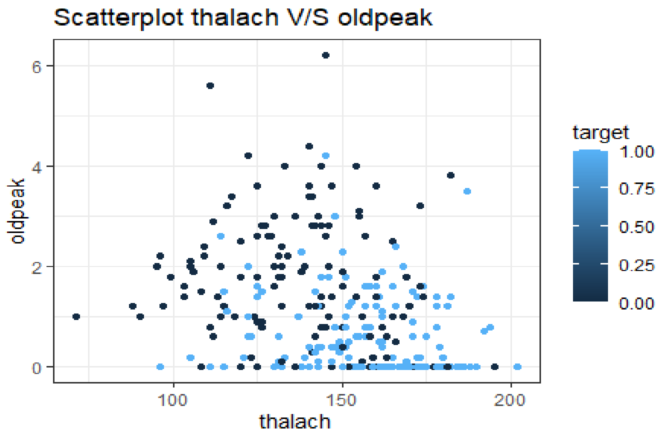


Scatterplot trestbps V/S oldpeak









IV. Experimental results:

A. K Fold Cross Validation:

KFold Mean Accuracies:

K	K-Fold Mean Accuracies
3	83.96872
4	83.39844
5	83.51220
6	83.72549

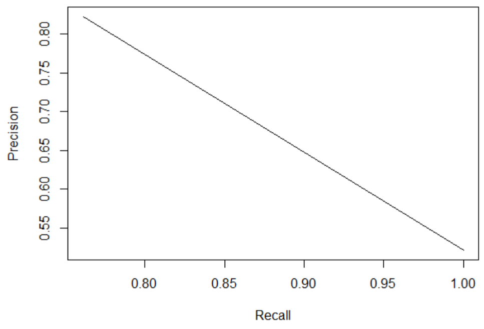
7	83.36595
8	83.49609
9	83.97247
10	83.33333
11	83.38221
12	83.52941
13	83.53057
14	83.36595
15	83.52941

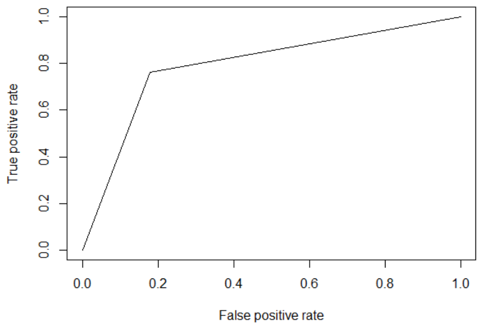
The maximum K Fold Mean Accuracy is 83.97247 at K = 9.
Hence the best value of K is 9.

B. Confusion matrix:

	predicted_target	
target	0	1
0	101	32
1	22	102

Precision = 0.859873
Recall = 0.822581
F1 score = 0.790698
Accuracy = 78.988327
Root Mean Square Error = 0.458385





Area Under Curve(AUC) value : 0.791166

C. Contingency Tables:

Prior Probabilties of target and sex

	sex	
target	0	1
0	0.08072917	0.39583333
1	0.23567708	0.28776042

Posterior Probabilties of target and sex

	sex	
target	0	1
0	0.2551440	0.5790476
1	0.7448560	0.4209524

Prior Probabilities of target and cp

	cp			
target	0	1	2	3
0	0.35416667	0.03385417	0.07161458	0.01692708
1	0.10546875	0.13932292	0.22786458	0.05078125

Posterior Probabilities of target and cp

	cp			
target	0	1	2	3
0	0.7705382	0.1954887	0.2391304	0.2500000
1	0.2294618	0.8045113	0.7608696	0.7500000

Prior Probabilities of target and fbs

	fbs	
target	0	1
0	0.39062500	0.08593750
1	0.45572917	0.06770833

Posterior Probabilities of target and fbs

	fbs	
target	0	1
0	0.4615385	0.5593220
1	0.5384615	0.4406780

Prior Probabilities of target and restecg

	restecg		
target	0	1	2
0	0.283854167	0.182291667	0.010416667
1	0.223958333	0.296875000	0.002604167

Posterior Probabilities of target and restecg

	restecg		
target	0	1	2
0	0.5589744	0.3804348	0.8000000
1	0.4410256	0.6195652	0.2000000

Prior Probabilities of target and exang

	exang	
target	0	1
0	0.20963542	0.26692708
1	0.45572917	0.06770833

Posterior Probabilities of target and exang

	exang	
target	0	1
0	0.3150685	0.7976654
1	0.6849315	0.2023346

Prior Probabilities of target and slope

	slope		
target	0	1	2
0	0.04296875	0.30729167	0.12630208
1	0.02734375	0.15364583	0.34244792

Posterior Probabilities of target and slope

	slope		
target	0	1	2
0	0.6111111	0.6666667	0.2694444
1	0.3888889	0.3333333	0.7305556

Prior Probabilities of target and ca

	ca				
target	0	1	2	3	4
0	0.156250000	0.144531250	0.109375000	0.063802083	0.002604167
1	0.410156250	0.070312500	0.019531250	0.006510417	0.016927083

Posterior Probabilities of target and ca

	ca				
target	0	1	2	3	4
0	0.27586207	0.67272727	0.84848485	0.90740741	0.13333333
1	0.72413793	0.32727273	0.15151515	0.09259259	0.86666667

Prior Probabilities of target and thal

	thal			
target	0	1	2	3
0	0.00390625	0.04427083	0.12500000	0.30338542
1	0.00390625	0.01822917	0.42187500	0.07942708

Posterior Probabilities of target and thal

	thal			
target	0	1	2	3
0	0.5000000	0.7083333	0.2285714	0.7925170
1	0.5000000	0.2916667	0.7714286	0.2074830