A Comparative Empirical Study of Classification Techniques on Early Stage Diabetes Risk Prediction

By: **Krishna Ravali J (19MBMB12)**

Date:15-Nov-2020

OBJECTIVES

- Problem Statement
- Introduction
- Dataset Description
- Data Mining Techniques
- Results and Discussion
- Conclusion

PROBLEM STATEMENT

Predicting the risk of diabetes at its early stage in the patients based on presence/absence of particular health conditions.

Approach

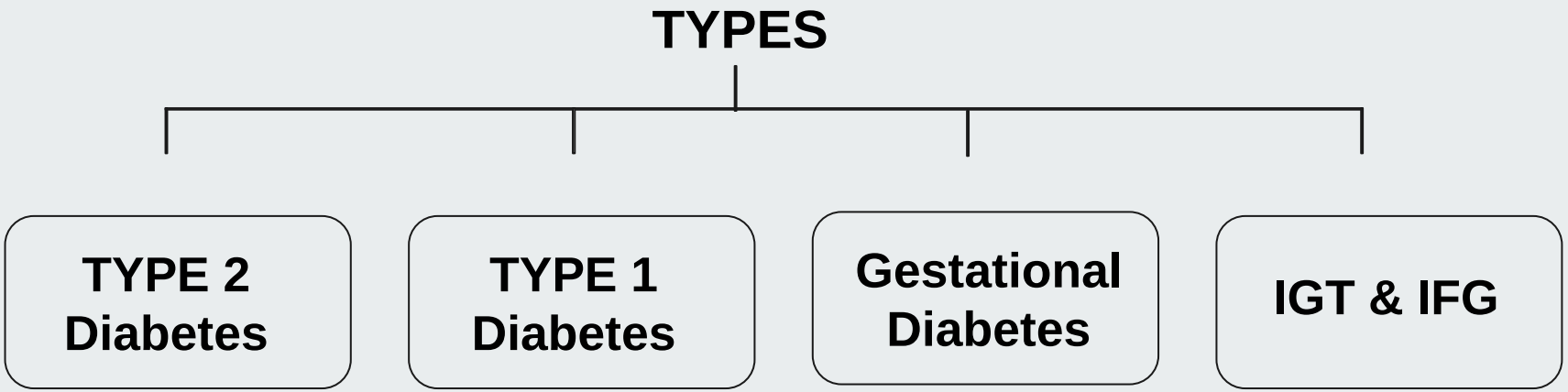
Using classification algorithms in Machine Learning, to solve this binary classification problem.

Motivation

- The trend of using ML techniques in health care has been gaining its popular due to its promising results and reduced cost of diagnosis.
- Given the situation of pandemic, such cost and time effective studies helps to improve the efficiency and effectivity of health care
- It is done by analysing various characteristics of person through different statistical models.

INTRODUCTION

- Diabetes is a medical condition in which a person suffers from high blood sugar levels
- According to International Diabetes Federation Diabetes Atlas, 9th edition, the global diabetes prevalence in 2019 is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. The prevalence is higher in urban(10.8%) than rural (7.2%) areas, and in high-income (10.4%) than low-income countries (4.0%). One in two (50.1%) people living with diabetes do not know that they have diabetes.



Source of Data

- The dataset used for study is taken from **UCI Machine Learning Repository**
- https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
- It has 520 instances, which has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh.
- The considered dataset is labelled dataset, hence supervised models are used.
- This resulted optimized model can be used to predict risk of diabetes of new patients.
- Data snippet is shown below:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Age	Gender	Polyuria	Polydipsia	sudden w	weakness	Polyphagi	Genital th	visual blur	Itching	Irritability	delayed h	partial par	muscle sti	Alopecia	Obesity	class
2	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
3	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
4	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
5	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
6	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive
7	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Positive

Dataset Description

The dataset contains 520 instances with 17 number of attributes

- *Polyuria- frequent and abnormal urination*
- *Polydipsia- temporary or prolonged mouth dryness leading to thirst*
- *Polyphagia- excess hunger*
- *Genital Thrush - yeast infection in private genital parts associated with pain and irritation*
- *Partial Paresis - partial or incomplete paralysis*
- *Alopecia- hair fall out as small patches*
- *Muscle Stiffness- muscle pain, joint pain or stiffness, joint swelling, deformities, and a "pins and needles" sensation in the arms or legs*

S.No	Attribute Name
1	Age
2	Sex
3	Polyuria
4	Polydipsia
5	Sudden Weight Loss
6	Weakness
7	Polyphagia
8	Genital Thrush
9	Visual Blurring
10	Itching
11	Irritability
12	Delayed Healing
13	Partial Paresis
14	Muscle Stiffness
15	Alopecia
16	Obesity
17	Class

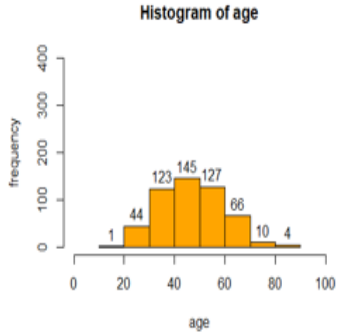
Outcome

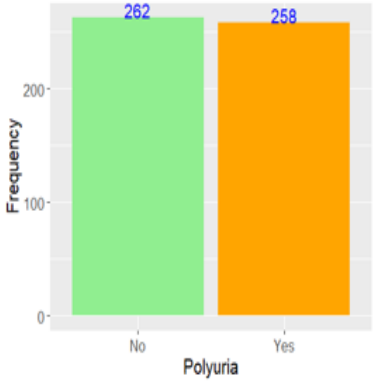
Using the labelled data at hand(which is historical data), we are creating and identifying an optimized model which when deployed to production helps to predict the target class i.e., whether a person is diabetic or not for new patients.

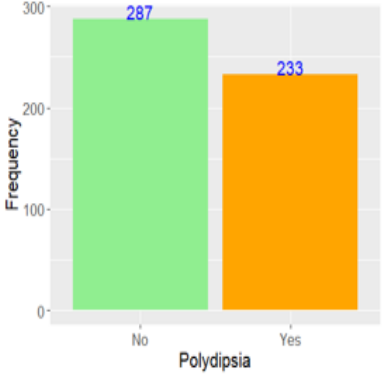
ML Algorithm Flow

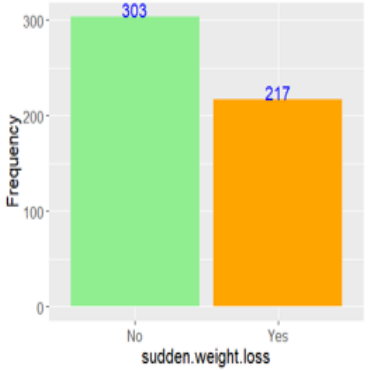
1. Importing the dataset to R
2. Visualization
3. Data Cleansing
4. Feature Engineering
5. Train-Test Split(70:30)
6. Applying classification techniques , testing (before and after cross validation/tuning)
7. Deriving performance metrics and comparison

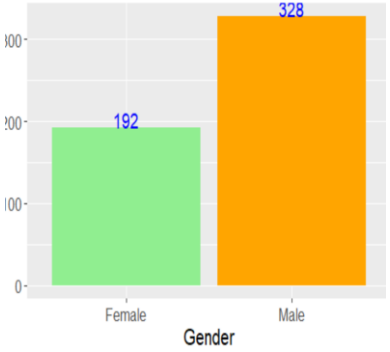
Frequency distribution of input and output variable

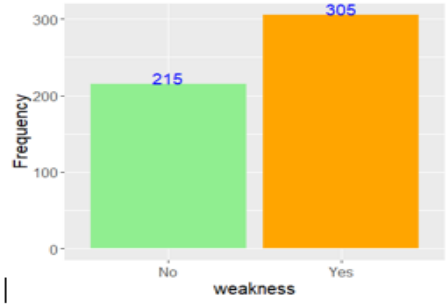


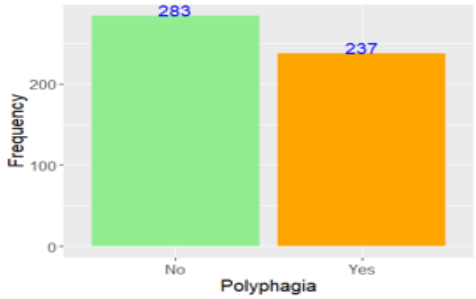


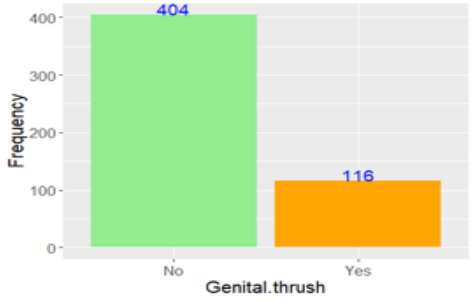


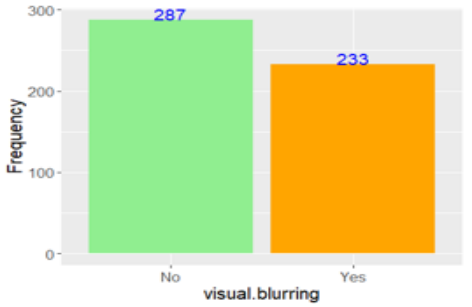


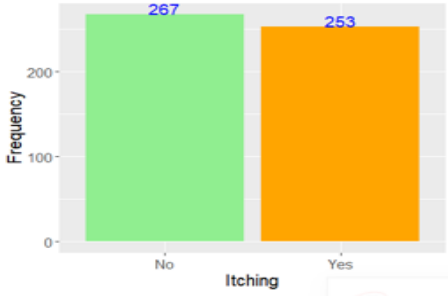


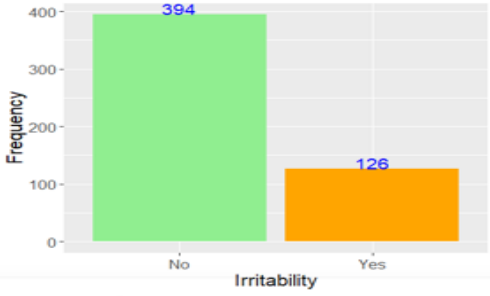

Frequency
300
200
100
0
215
305
No
Yes
weakness

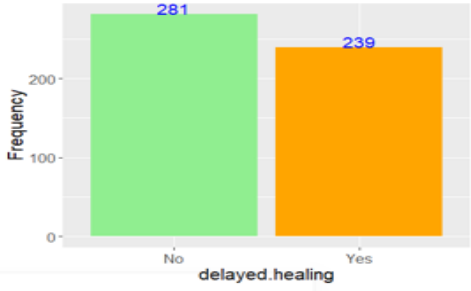

Frequency
200
100
0
283
237
No
Yes
Polyphagia

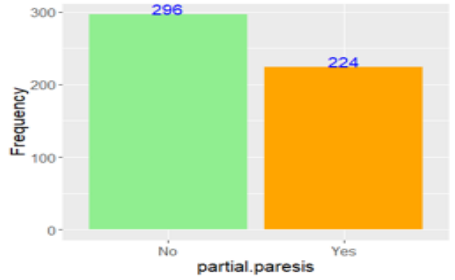

Frequency
400
300
200
100
0
404
116
No
Yes
Genital.thrush

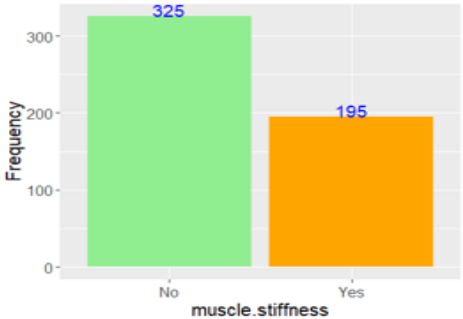

Frequency
300
200
100
0
287
233
No
Yes
visual.blurring

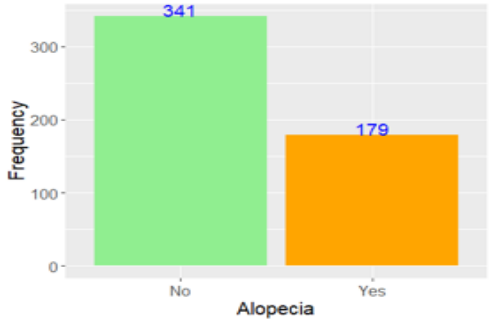

Frequency
200
100
0
267
253
No
Yes
Itching

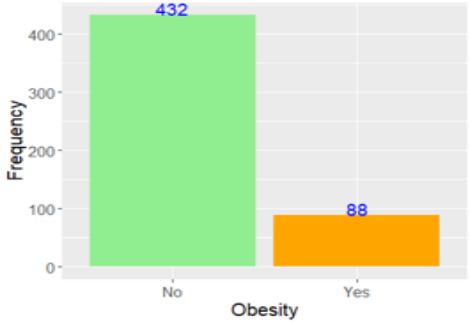

Frequency
400
300
200
100
0
394
126
No
Yes
Irritability


Frequency
200
100
0
281
239
No
Yes
delayed.healing


Frequency
300
200
100
0
296
224
No
Yes
partial.paresis


Frequency
300
200
100
0
325
195
No
Yes
muscle.stiffness


Frequency
300
200
100
0
341
179
No
Yes
Alopecia


Frequency
400
300
200
100
0
432
88
No
Yes
Obesity


Frequency
300
200
100
0
200
320
Negative
Positive
class

Significant attributes(using glm)

Coefficients:

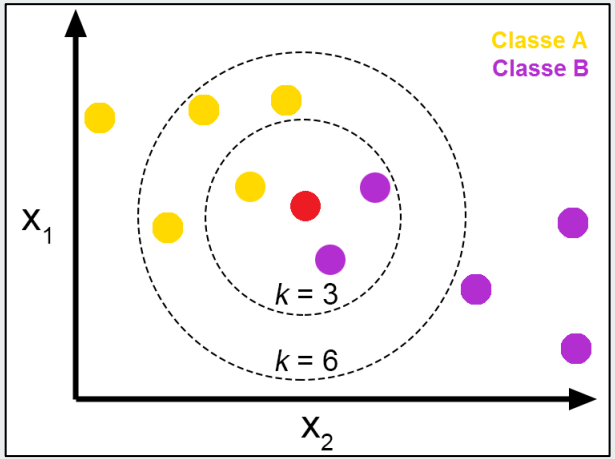
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.6267	0.8401	-0.746	0.455702	
Age	1.5936	0.7696	2.071	0.038381	*
Gender	-4.6332	0.7854	-5.899	3.66e-09	***
Polyuria	4.5104	0.8748	5.156	2.52e-07	***
Polydipsia	5.4424	1.0740	5.067	4.04e-07	***
sudden.weight.loss	0.4561	0.7643	0.597	0.550612	
weakness	0.6165	0.7052	0.874	0.381952	
Polyphagia	0.4528	0.6995	0.647	0.517418	
Genital.thrush	2.2589	0.7280	3.103	0.001916	**
visual.blurring	1.2341	0.8359	1.476	0.139845	
Itching	-2.8602	0.8649	-3.307	0.000943	***
Irritability	2.7915	0.8063	3.462	0.000536	***
delayed.healing	-0.9656	0.7551	-1.279	0.200962	
partial.paresis	1.0451	0.6398	1.633	0.102366	
muscle.stiffness	0.2733	0.8346	0.327	0.743353	
Alopecia	0.5271	0.8055	0.654	0.512853	
Obesity	-0.6665	0.7051	-0.945	0.344536	

ML Techniques used

1.KNN Algorithm
2.Logistic Regression
3. Naive Bayes
4.Support Vector Machine

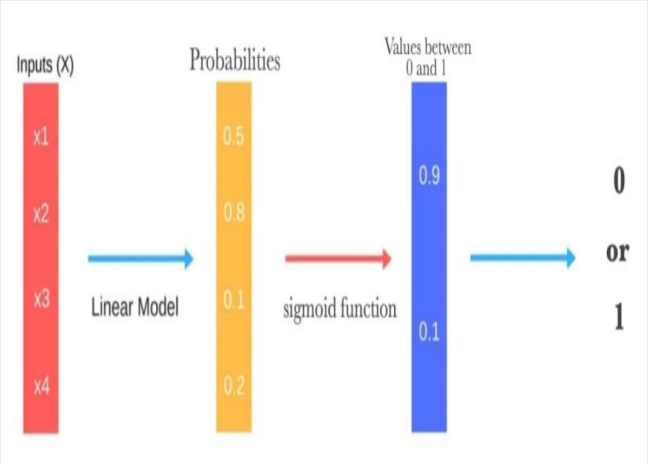
K-NN

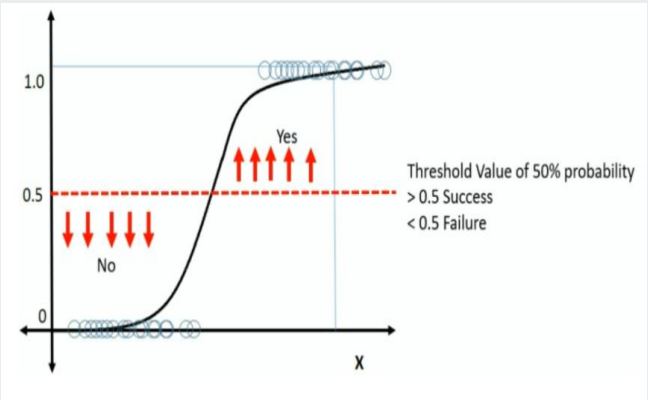
- K- NN is a *Lazy Learning Method* also known as *Instance based Learning*
- Supervised Machine Learning algorithm to identify patterns
- Stores all available cases and classifies new cases based on similarity measure



Logistic Regression

- Logistic Regression is one of the most used Machine Learning algorithms for binary classification
- Logistic Regression measures the relationship between the dependent variable and the one or more independent variables , by estimating probabilities using it’s underlying logistic function.
- These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the sigmoid function.





Naive Bayes

The Naïve Bayes Classifier belongs to the family of probability classifier, using Bayesian theorem.
The reason why it is called ‘Naïve’ because it requires rigid independence assumption between input variables.

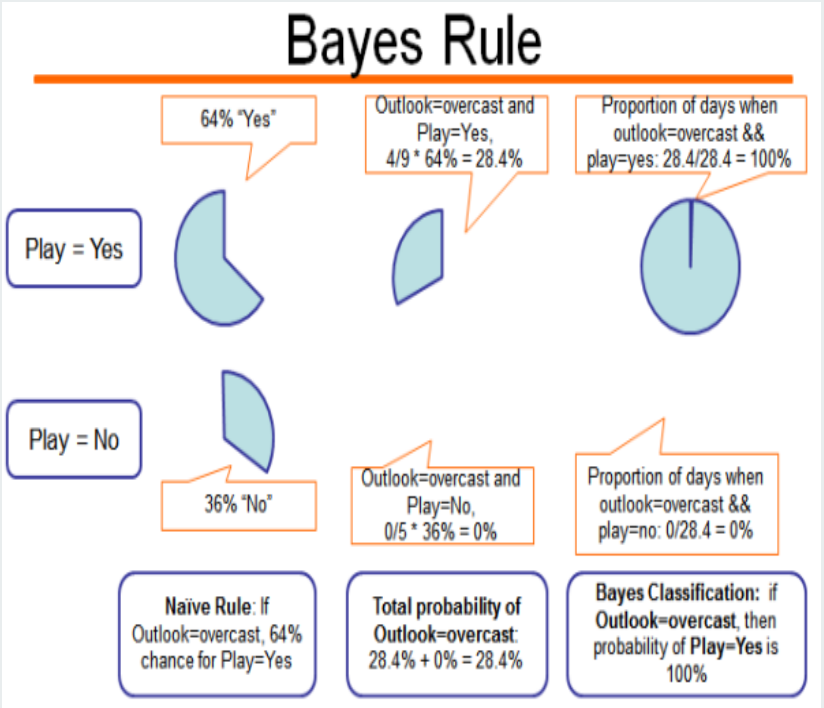
The fundamental Naïve Bayes assumption is that each feature makes an:

Independent & equal contribution to the outcome.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

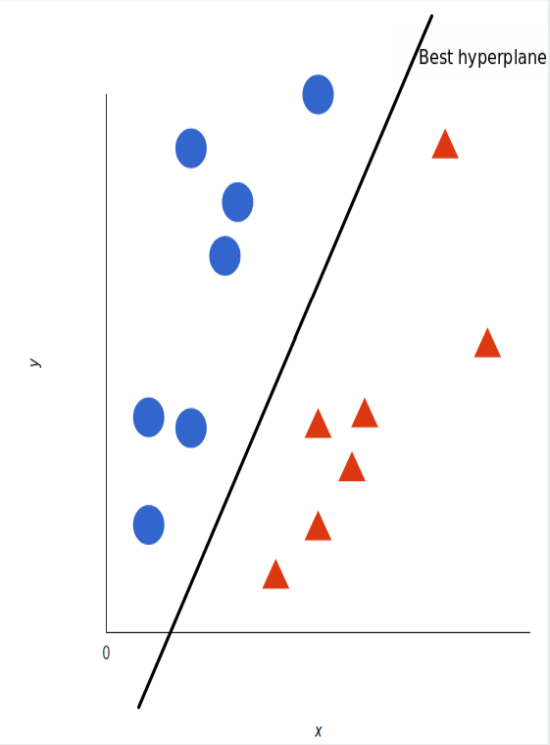
Where A,B are events and P(B)!=0

- P(A|B) is a conditional probability: The likelihood of event A occurring given that B is true.
- P(B|A) is a conditional probability: The likelihood of event B occurring given that A is true.
- P(A) and P(B) are the probabilities of observing A,B independently of each other, this is known as the ‘marginal probability’.



Support Vector Machine

- This is a supervised machine learning technique helps to classify between groups..
- SVM model classifies the different data points using a hyperplane in a N-dimensional space.
- Hyperplanes should be set up in such a way that there is maximum distance between data points.
- Data points close to these hyperplanes holds more significance in the data set.



Observations from results

- Model metrics like Accuracy, F1 score, Specificity, Sensitivity and Precision are compared before and after cross validation
- Since specificity and sensitivity are important performance metrics of a machine learning model in the case of analysing medical data, more focus is laid on those metrics
- Among the four classification models, classification error rates are high in Naive Bayes and KNN classification followed by Logistic Regression
- In SVM model, error rate is minimal along with having optimum levels of Specificity and sensitivity.

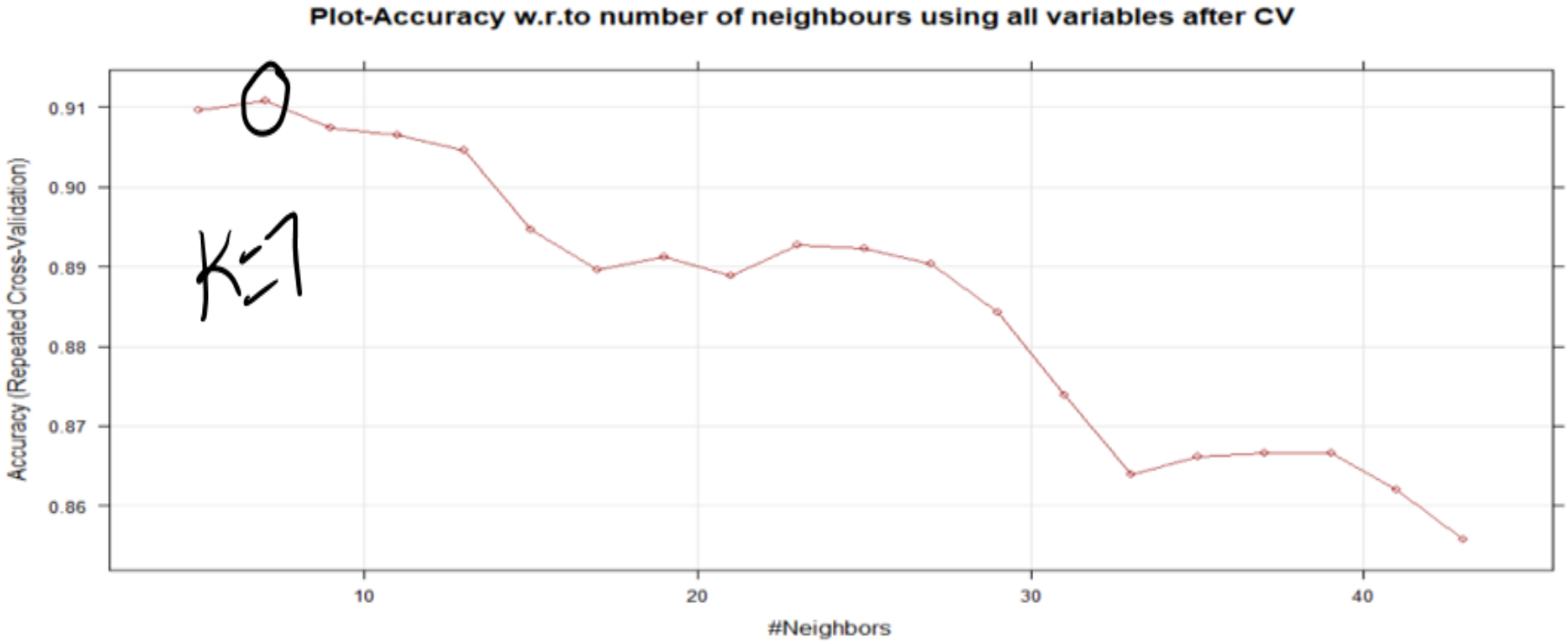


Fig (**e**) line plot showing variation in accuracy of KNN with number of neighbors using all variables after CV (for training data)

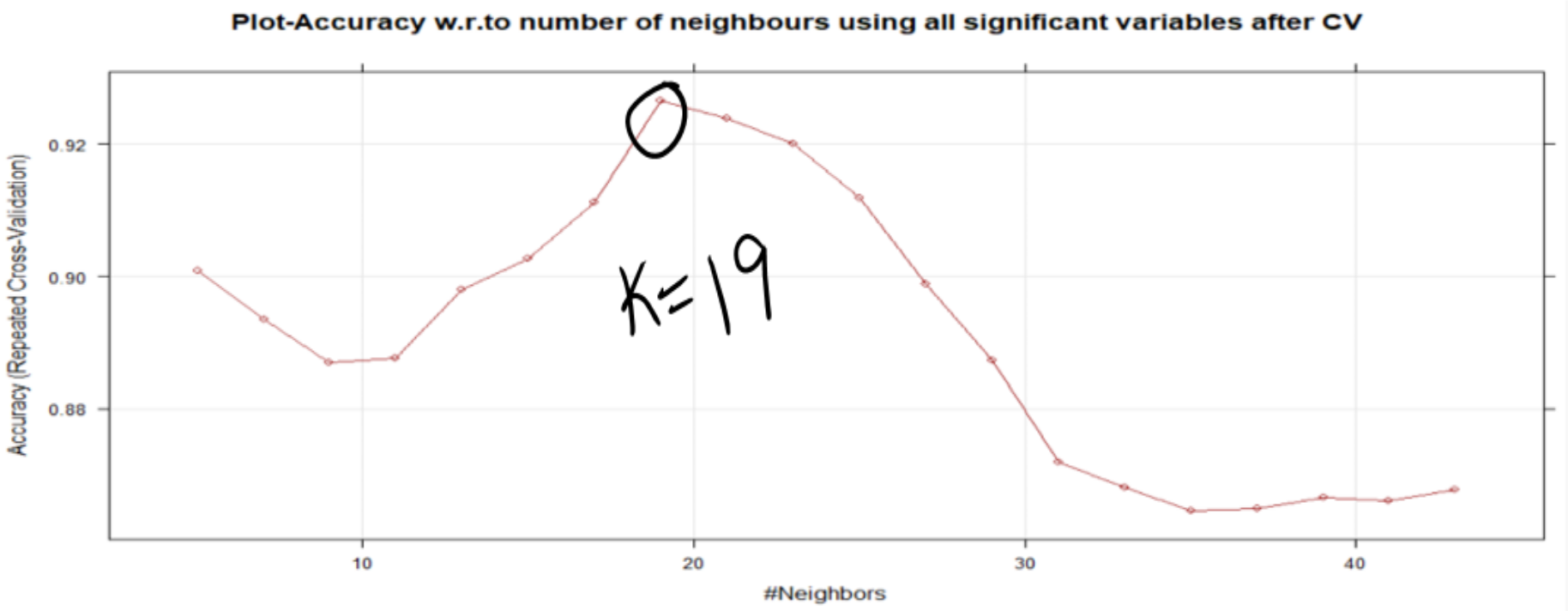


Fig (**f**) line plot showing variation in accuracy of KNN with number of neighbors using all significant variables after CV (for training data)

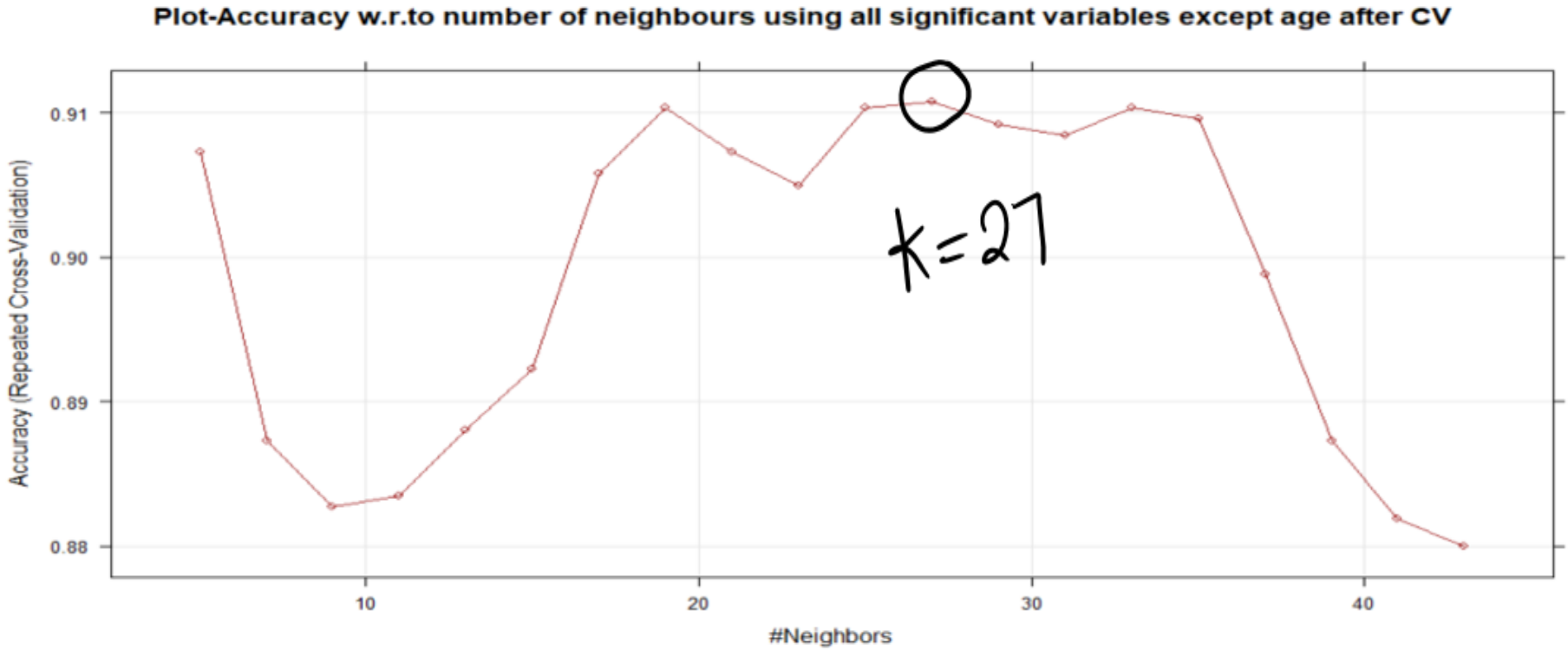
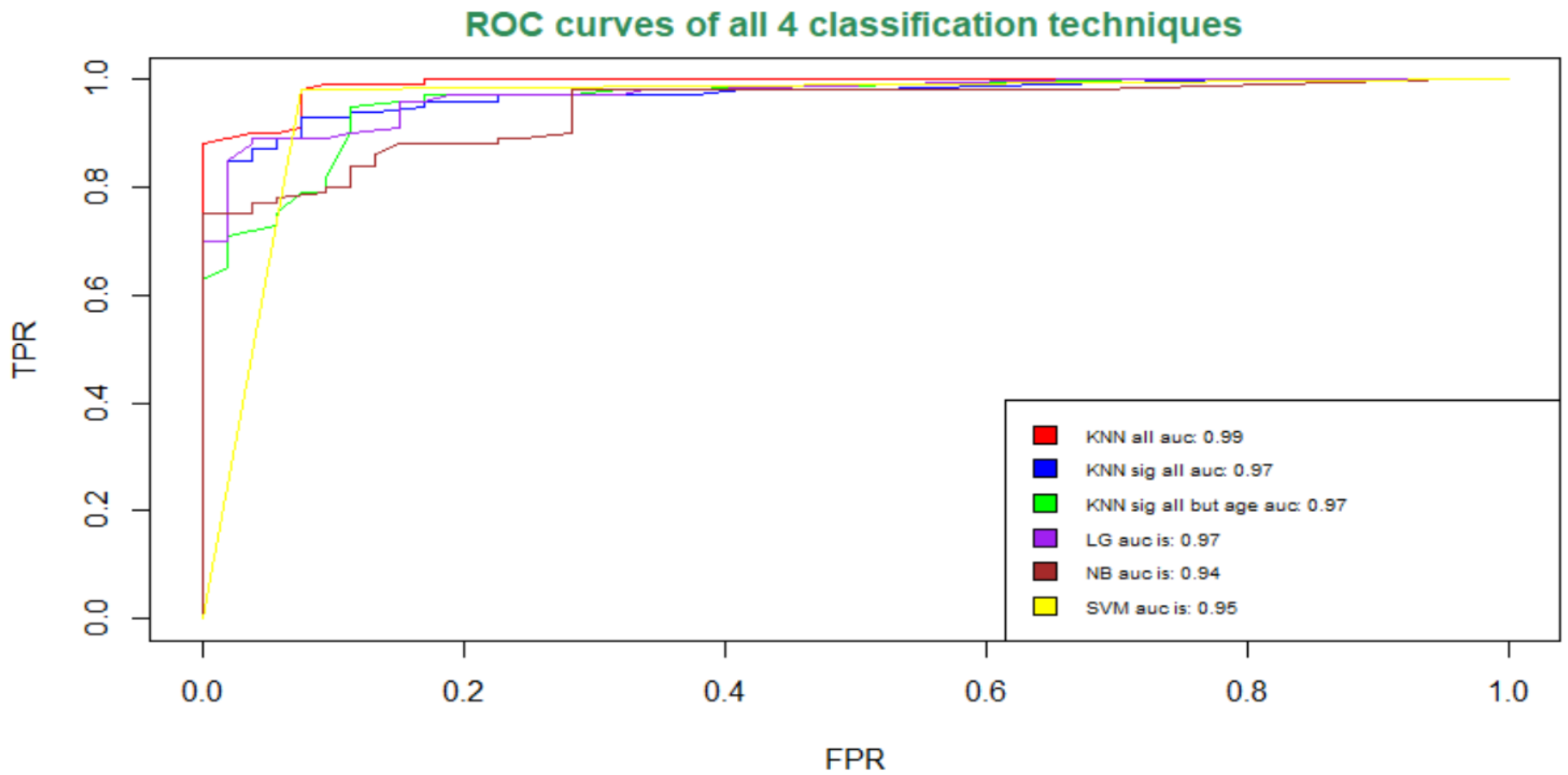


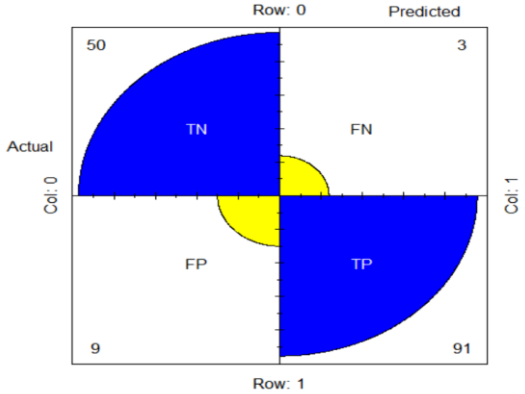
Fig (**g**) line plot showing variation in accuracy of KNN with number of neighbors using all significant variables except age after CV (for training data)

	Accuracy	Recall or Sensitivity	Specificity	F1-score	Classification Error Rate	Precision
KNN - using all variables	92.15686	96.80851	84.74576	93.90426	7.843137	91.00000
KNN - using only significant variables	87.58170	94.50549	77.41935	90.27624	12.418301	86.00000
KNN - using only significant variables except Age	88.23529	91.00000	83.01887	91.00000	11.764706	91.00000
KNN - cross validation using all variables	90.84967	90.00000	92.45283	92.90323	9.150327	95.74468
KNN - cross validation using only significant variables	90.19608	89.00000	92.45283	92.36757	9.803922	95.69892
KNN - cross validation using only significant variables except Age	92.15686	94.00000	88.67925	94.00000	7.843137	94.00000
Logistic Regression - using all variables	92.81046	95.00000	88.67925	94.49735	7.189542	94.05941
Logistic Regression - using significant model	90.84967	94.00000	84.90566	92.98925	9.150327	92.15686
Logistic Regression - cross validation using all variables	92.81046	95.00000	88.67925	94.49735	7.189542	94.05941
Logistic Regression - cross validation using significant model	90.84967	94.00000	84.90566	92.98925	9.150327	92.15686
Naive Bayes Classifier - using all variables	90.19608	95.00000	81.13208	92.43243	9.803922	90.47619
Naive Bayes Classifier - using significant variables	84.96732	88.00000	79.24528	88.49718	15.032680	88.88889
Naive Bayes Classifier - cross validation using all variables	90.19608	95.00000	81.13208	92.43243	9.803922	90.47619
Naive Bayes Classifier - cross validation using significant variables	84.96732	88.00000	79.24528	88.49718	15.032680	88.88889
SVM Classifier - using all variables	96.07843	98.00000	92.45283	96.98969	3.921569	96.07843
SVM Classifier - using significant variables	93.46405	94.00000	92.45283	94.98947	6.535948	95.91837
SVM Classifier - using performance tuning on all variables	98.69281	100.00000	96.22642	98.98990	1.307190	98.03922
SVM Classifier - using performance tuning on significant variables	94.11765	93.00000	96.22642	95.43455	5.882353	97.89474

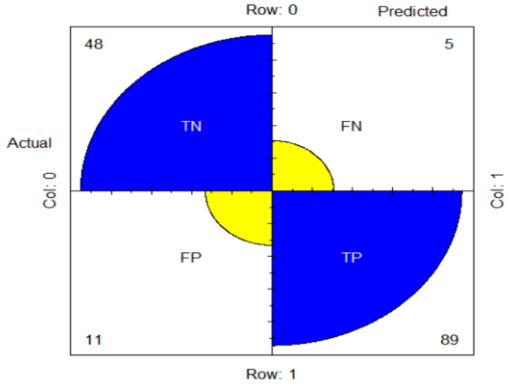


It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

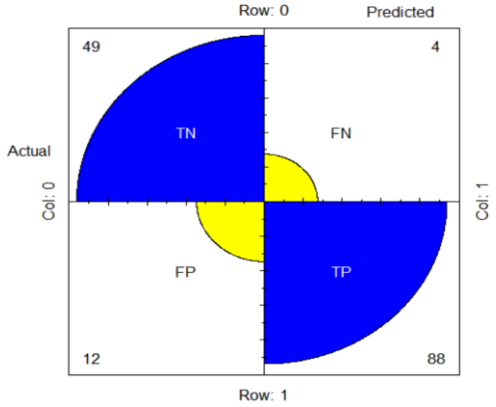
CM-KNN using all variables before CV



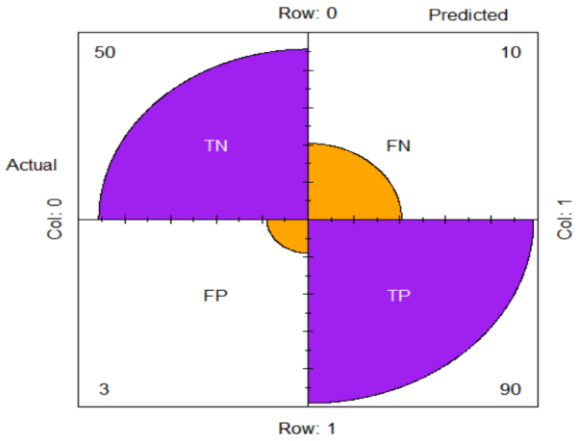
CM-KNN using all significant variables before CV



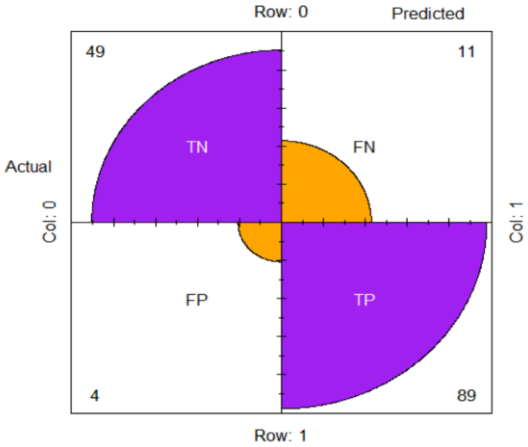
CM-KNN using all significant variables except age before CV



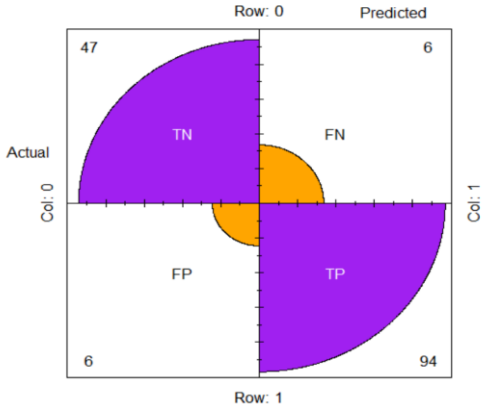
CM-KNN using all variables after CV



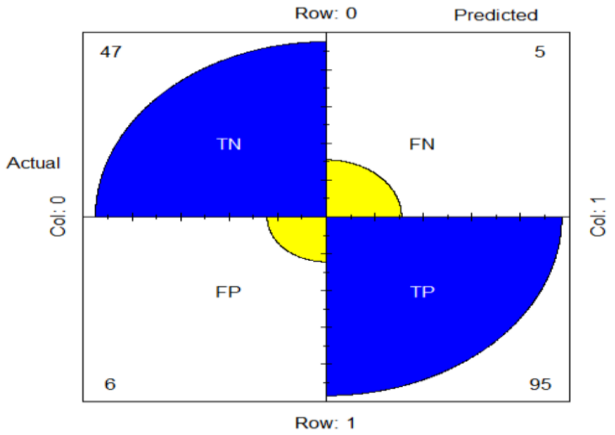
CM-KNN using all significant variables after CV



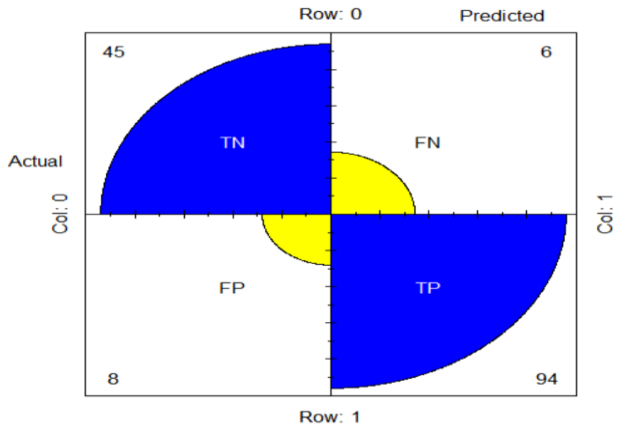
CM-KNN using all significant variables except age before CV



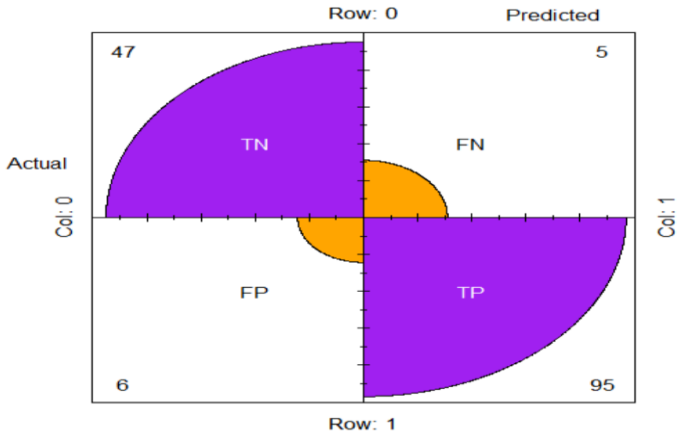
CM-LR using all variables before CV



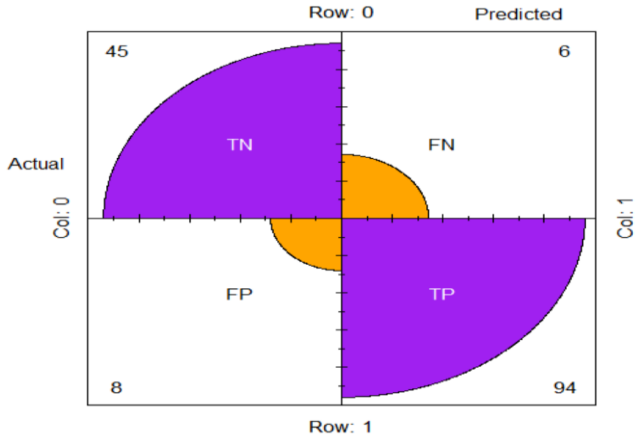
CM-LR using optimum model before CV



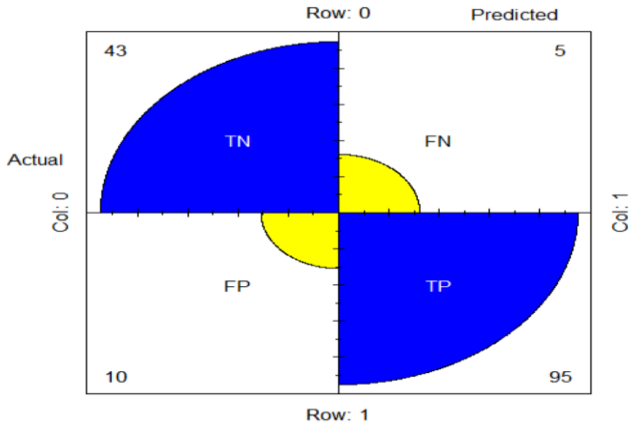
CM-LR using all variables after CV



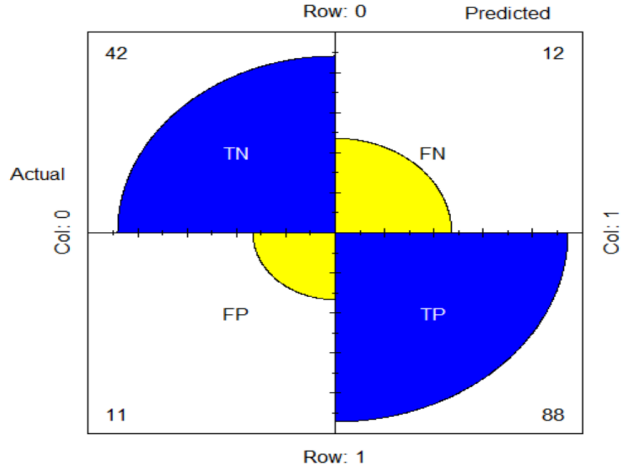
CM-LR using optimum model after CV



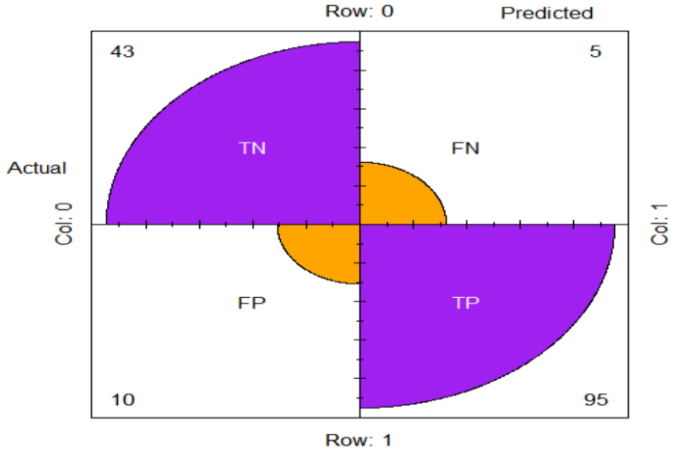
CM-NB using all variables before CV



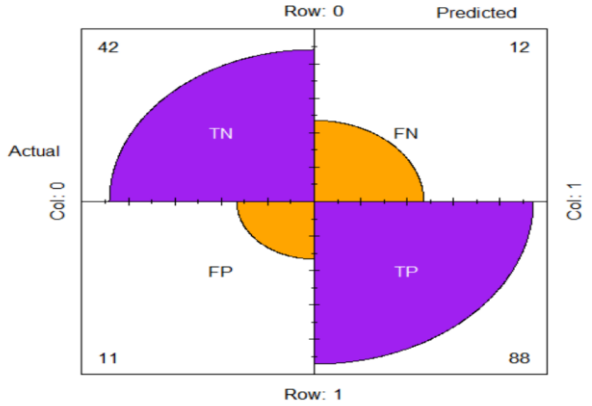
CM-NB using all significant variables before CV



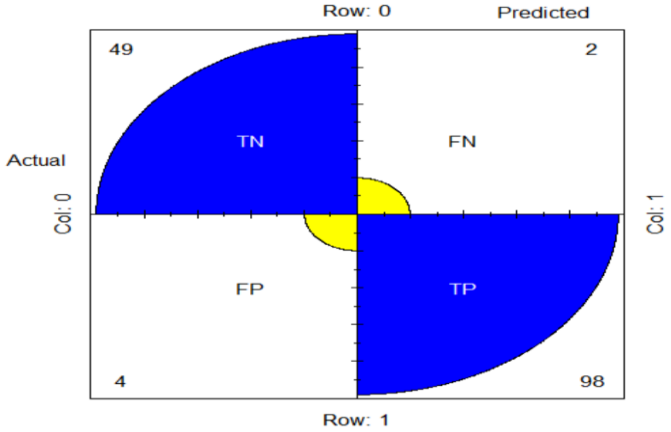
CM-NB using all variables after CV



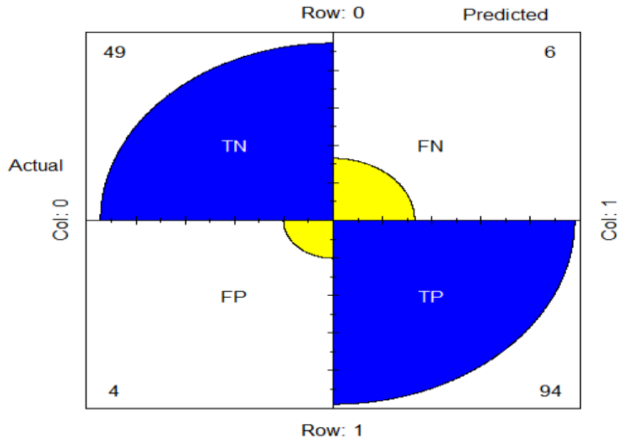
CM-NB using all significant variables after CV



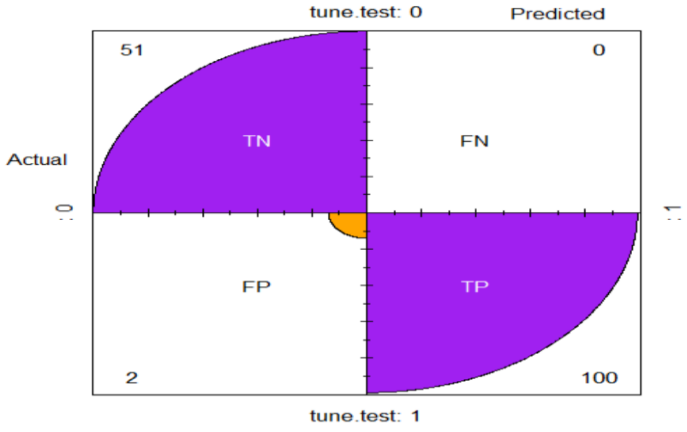
CM-SVM using all variables before CV



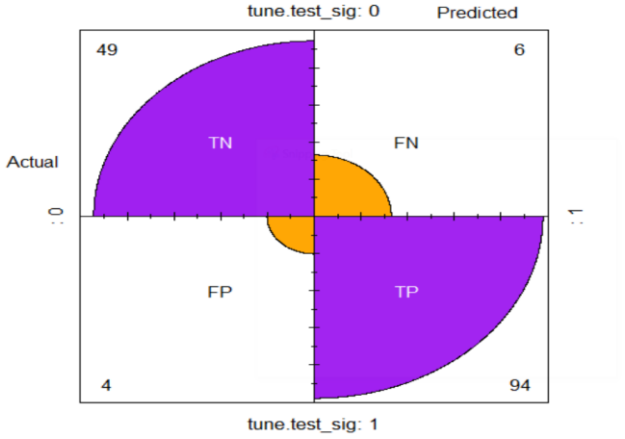
CM-SVM using all significant variables before CV



CM-SVM using all variables after CV



CM-SVM using all significant variables after CV



Conclusion

- Its concluded that, SVM classification model is the best fit model for this particular data set
- Proper use of these techniques facilitate the process of diagnosing right disease at right time among the patients.
- Chance of delaying treatment or not getting proper treatment which results in risking one's life can be reduced.
- During this COVID pandemic, people with comorbidities are the most vulnerable population. These kind of studies using machine learning techniques helps to find out those people and give proper care.

Future enhancements

- Application of random forest, deep learning techniques, hybrid models.

Thank You