Machine Learning for Breast Cancer Diagnosis

Kishan Dalsania(17IT015)Dharmesh Chauhan(17IT013),
Smt. Kundanben Dinsha Patel Department of Information Technology,
CHARUSAT University, Gujarat, India

Abstract— Breast cancer is one among the diseases which make a large number of deaths per annum. It is the foremost common sort of all cancers and therefore the main explanation for deaths of women worldwide. A Classification and data processing methods are very powerful thanks to classification of data. Specially in medical Area, where these methods are utilized in diagnosis as well as analysis to form decisions. during, a performance comparison between different-different machine learning methods used for a classification : Support Vector Machine (SVM), Tree , Naive Bayes (NB), Random Forest , Stochastic Gradient Descent ,Logistic Regression and Neural Network on the Wisconsin carcinoma(2 different type of dataset with different attributes) datasets is used. The main objective is to assess the correctness in classification of data with reference to efficiency and effectiveness of every algorithm in terms of accuracy, precision, recall, and f1 score.

Keywords— Breast Cancer diagnosis, Classification, Accuracy, Neural Network, SVM, Decision Tree, Random Forest, Efficiency.

I. INTRODUCTION

The major explanation for women's death is carcinoma (after lung cancer). 246,660 of women's new cases of invasive carcinoma are expected to be diagnosed within the US during 2016 and 40,450 of women’s death is counted. carcinoma presents about 12% of all new cancer cases and 25% of all cancers in women.

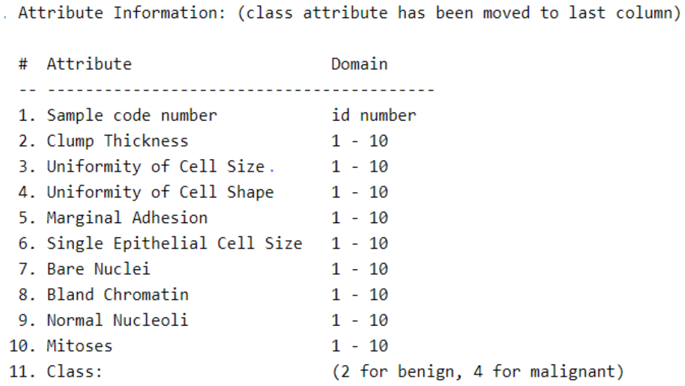
There are many different algorithms for classification and prediction of carcinoma outcomes. this paper provides a comparison between the performance of different classifiers: Support Vector Machine (SVM), Tree, Naive Bayes (NB), Random Forest , Stochastic Gradient Descent ,Logistic Regression and Neural Network which are the foremost influential data processing algorithms. Our aim is to gauge more efficiency and high effectiveness of these algorithms in terms of accuracy, sensitivity, specificity, and precision.

II. BACKGROUND

we initially present the “Breast cancer classification’s” important attributes, After that different machine learning methods used in our classification.

Breast cancer classification (BCC) means to decide the reasonable therapy, that can be forceful or less forceful, contingent upon the class of the disease. To build a decent prophecy, BCC order requies many attributes to measure among them we are divided the attribute as per the importance and available dataset. We have 40 different attributes among them we divided them into 9 attribute and 31 attributes.

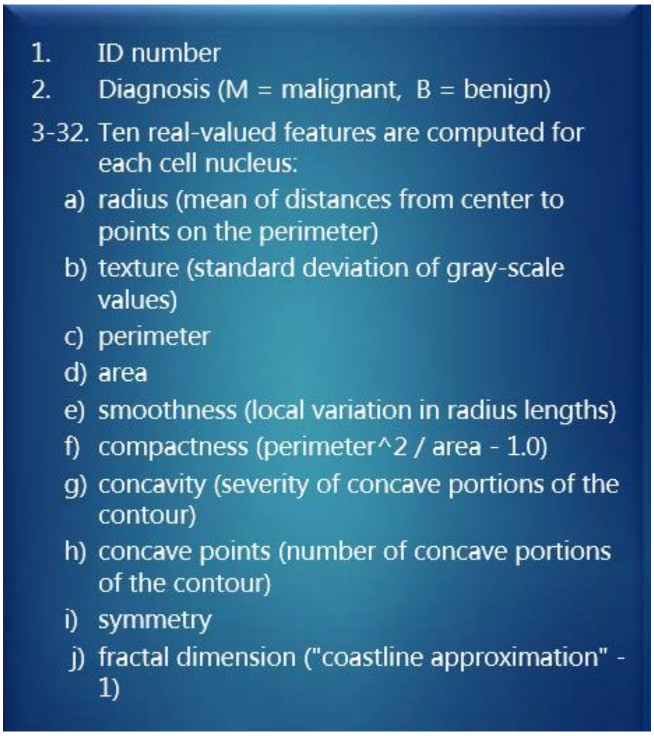
Dataset 1 with 9 attributes which are as follows:



```
 Attribute Information: (class attribute has been moved to last column)

 #  Attribute                     Domain
 -- -----------------------------------------
 1. Sample code number            id number
 2. Clump Thickness               1 - 10
 3. Uniformity of Cell Size       1 - 10
 4. Uniformity of Cell Shape      1 - 10
 5. Marginal Adhesion             1 - 10
 6. Single Epithelial Cell Size   1 - 10
 7. Bare Nuclei                   1 - 10
 8. Bland Chromatin               1 - 10
 9. Normal Nucleoli               1 - 10
10. Mitoses                       1 - 10
11. Class:                        (2 for benign, 4 for malignant)
```

Dataset 2 with 30 attributes which are as follows :



1. ID number
2. Diagnosis (M = malignant, B = benign)

3-32. Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

III. DATA DESCRIPTION

A. Dataset 1 :

The breast cancer Wisconsin (Original) Data set is used in this study. Breast-cancer-Wisconsin has 699 instances (Benign: 458 Malignant: 241), 2 classes (34.5% malignant and 65.52% benign), and 10 integer-valued attributes.

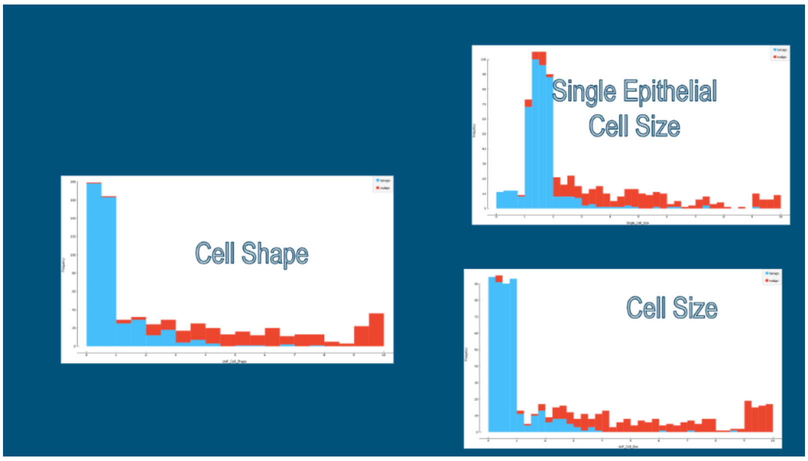
B. Dataset 2 :

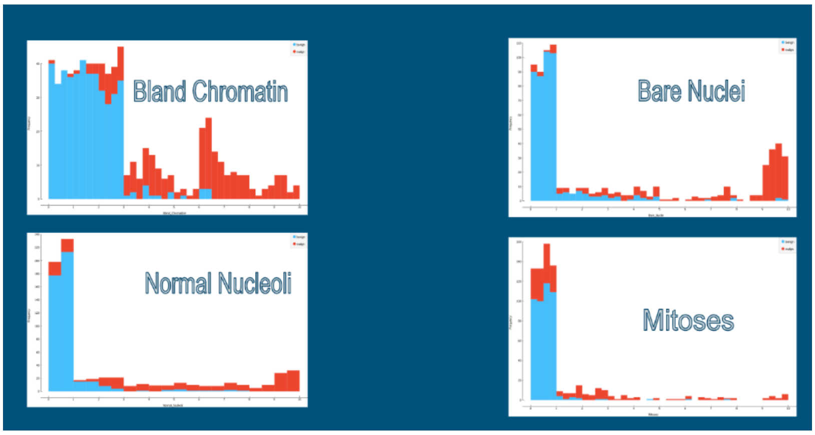
The breast cancer Wisconsin (Diagnostic) Data set is taken in this study and research purpose. Breast-cancer-Wisconsin(Diagnosis) dataset has 569 instances in that, Benign: 357 and Malignant: 212, 2 classes (37.25% malignant & 62.74% benign), and 30 integer-valued attributes.

IV.DATA VISUALIZATION

A. Data Set 1

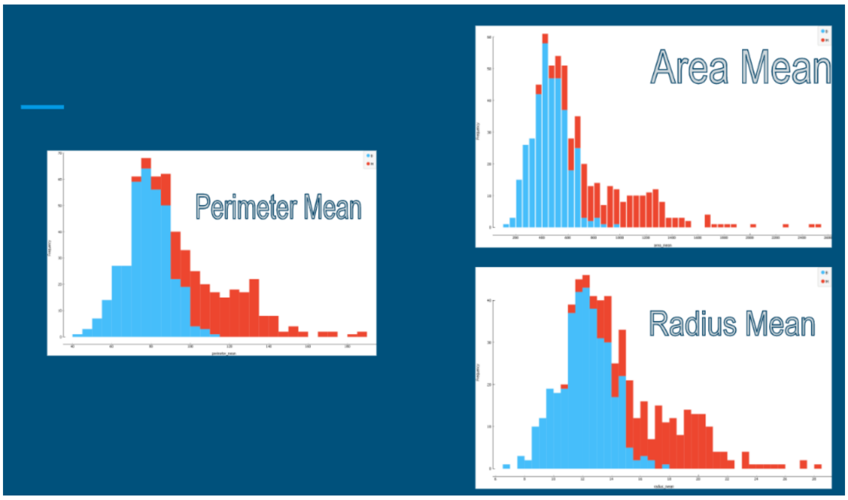
The features distinguish between benign and Malignant fairly well.



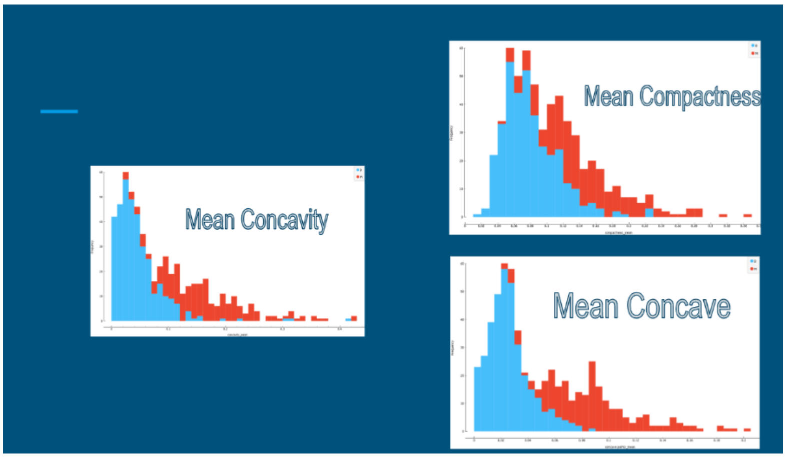


B. Data Set 2

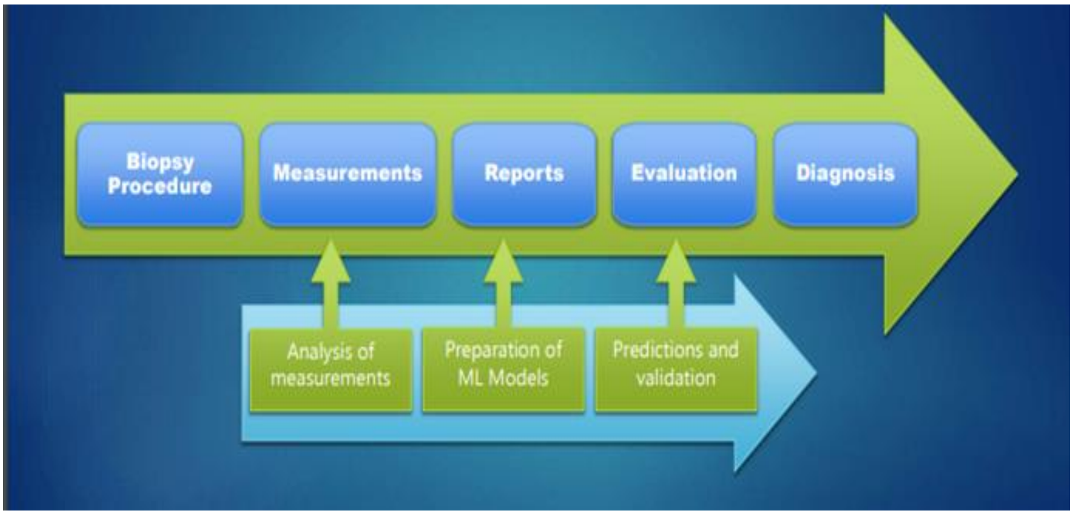
Mean Radius, Mean Perimeter and Mean Area appear to be helpful in classification.Higher the values of each parameter more are the chances of it being malignant.



Mean Concavity, Mean Concave , and Mean Compactness appear to be helpful in classification.Higher the values of each parameter more are the chances of it being malignant.



V. FLOW OF DATA



Biopsy Procedure → Measurements → Reports → Evaluation → Diagnosis

Analysis of measurements → Preparation of ML Models → Predictions and validation

VI. APPROACHES AND EFFECTIVENESS

Machine learning is branch of AI, Machine learning techniques can utilize absolute conditionality, statistics, Boolean logic, probabilities, and unconventional optimization stats to classify a patterns or to build a classification models.

Classification Methods :

- Logistic Regression
- Naïve Bayes
- Tree
- Random Forest
- Neural Network
- Stochastic Gradient Descent
- Support Vector machine

Method of Sampling Data

- Cross Validation Method
  - Here in cross validation we have used 10 folds

Attributes in the Model :

Logistic Regression

- Regularization type : Lasso (L1) Regression

Tree :

- Binary tree
- Min. number of instances in leaves : 2
- Maximal tree depth : 100

Random Forest :

- No of trees : 10

Neural Network :

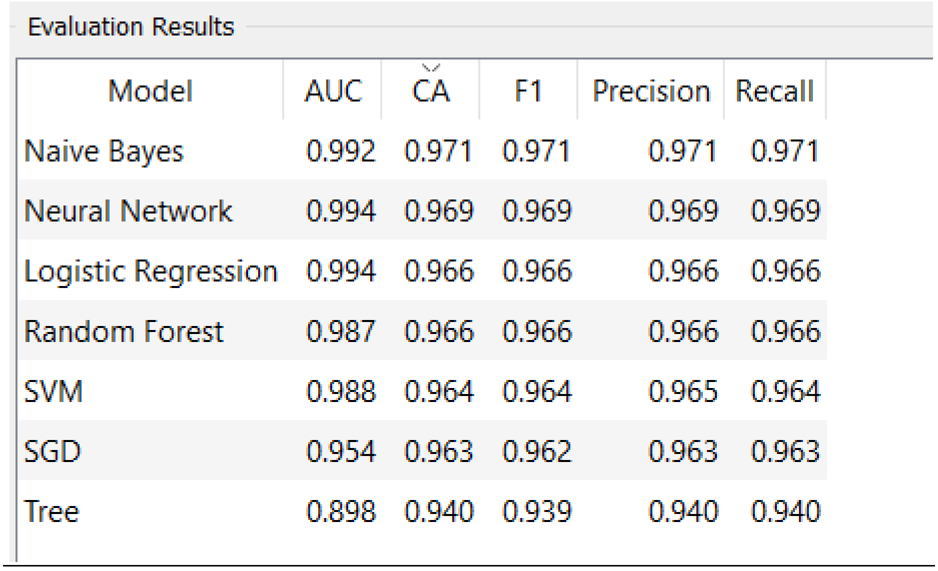
- Neuron in hidden layers : 100
- Activision Function : Relu
- Maximum number of iterations : 200

Stochastic Gradient Descent

- Classification loss Function : squared Epsilon insensitive (where epsilon is 0.10))
- Regularization method : Ridge(L2)
- Learning Rate : constant
- Initial learning rate : 0.0100
- No of iteration : 2511

EFFECTIVENESS (Dataset 1)

- **Performance**

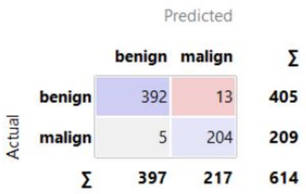


Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.992	0.971	0.971	0.971	0.971
Neural Network	0.994	0.969	0.969	0.969	0.969
Logistic Regression	0.994	0.966	0.966	0.966	0.966
Random Forest	0.987	0.966	0.966	0.966	0.966
SVM	0.988	0.964	0.964	0.965	0.964
SGD	0.954	0.963	0.962	0.963	0.963
Tree	0.898	0.940	0.939	0.940	0.940

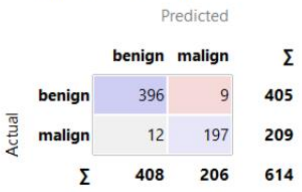
- **Confusion Matrix**

Naive bayes



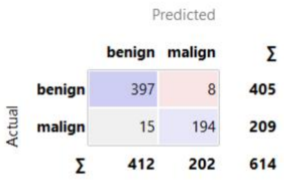
		Predicted		
		benign	malign	Σ
Actual	benign	392	13	405
	malign	5	204	209
	Σ	397	217	614

Logistic Regression



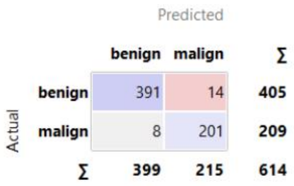
		Predicted		
		benign	malign	Σ
Actual	benign	396	9	405
	malign	12	197	209
	Σ	408	206	614

Stochastic Gradient Descent



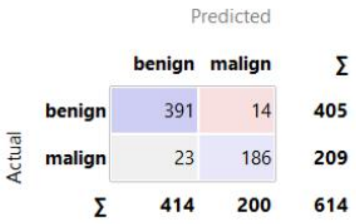
		Predicted		
		benign	malign	Σ
Actual	benign	397	8	405
	malign	15	194	209
	Σ	412	202	614

Support Vector Machine



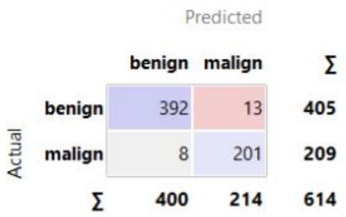
		Predicted		
		benign	malign	Σ
Actual	benign	391	14	405
	malign	8	201	209
	Σ	399	215	614

Tree



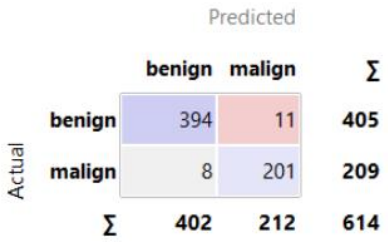
		Predicted		
		benign	malign	Σ
Actual	benign	391	14	405
	malign	23	186	209
	Σ	414	200	614

Random Forest



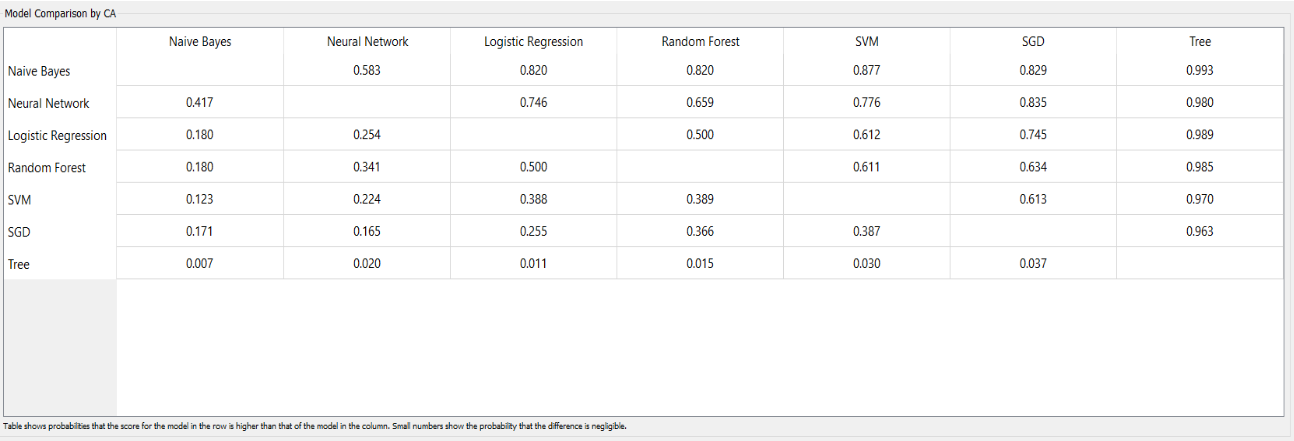
		Predicted		
		benign	malign	Σ
Actual	benign	392	13	405
	malign	8	201	209
	Σ	400	214	614

Neural Network



		Predicted		
		benign	malign	Σ
Actual	benign	394	11	405
	malign	8	201	209
	Σ	402	212	614

- **Comparison of model with Classification Accuracy**



Model Comparison by CA

	Naive Bayes	Neural Network	Logistic Regression	Random Forest	SVM	SGD	Tree
Naive Bayes		0.583	0.820	0.820	0.877	0.829	0.993
Neural Network	0.417		0.746	0.659	0.776	0.835	0.980
Logistic Regression	0.180	0.254		0.500	0.612	0.745	0.989
Random Forest	0.180	0.341	0.500		0.611	0.634	0.985
SVM	0.123	0.224	0.388	0.389		0.613	0.970
SGD	0.171	0.165	0.255	0.366	0.387		0.963
Tree	0.007	0.020	0.011	0.015	0.030	0.037	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

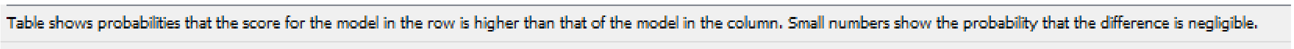
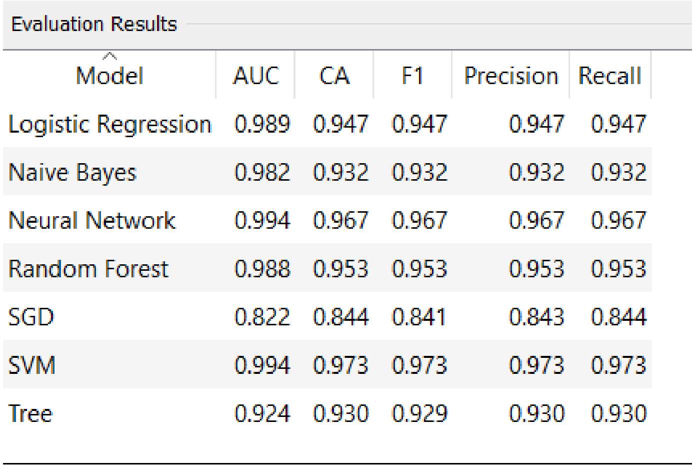


Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

EFFECTIVENESS (<u>Dataset 2)</u>

- **Performance**

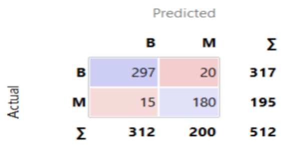


Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.989	0.947	0.947	0.947	0.947
Naive Bayes	0.982	0.932	0.932	0.932	0.932
Neural Network	0.994	0.967	0.967	0.967	0.967
Random Forest	0.988	0.953	0.953	0.953	0.953
SGD	0.822	0.844	0.841	0.843	0.844
SVM	0.994	0.973	0.973	0.973	0.973
Tree	0.924	0.930	0.929	0.930	0.930

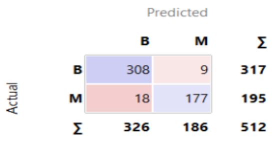
- **Confusion Matrix**

Naive bayes


Predicted

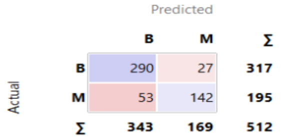
		B	M	Σ
Actual	B	297	20	317
	M	15	180	195
	Σ	312	200	512

Logistic Regression


Predicted

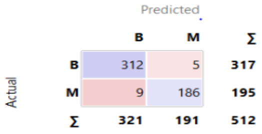
		B	M	Σ
Actual	B	308	9	317
	M	18	177	195
	Σ	326	186	512

Stochastic Gradient Descent


Predicted

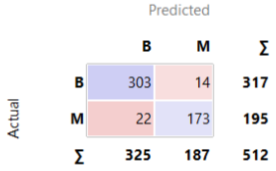
		B	M	Σ
Actual	B	290	27	317
	M	53	142	195
	Σ	343	169	512

Support Vector Machine


Predicted

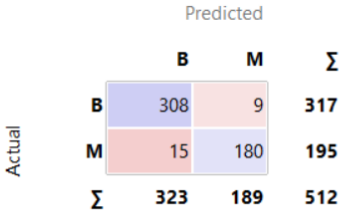
		B	M	Σ
Actual	B	312	5	317
	M	9	186	195
	Σ	321	191	512

Tree


Predicted

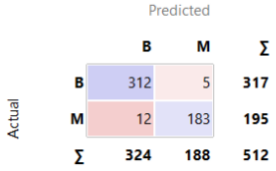
		B	M	Σ
Actual	B	303	14	317
	M	22	173	195
	Σ	325	187	512

Random Forest


Predicted

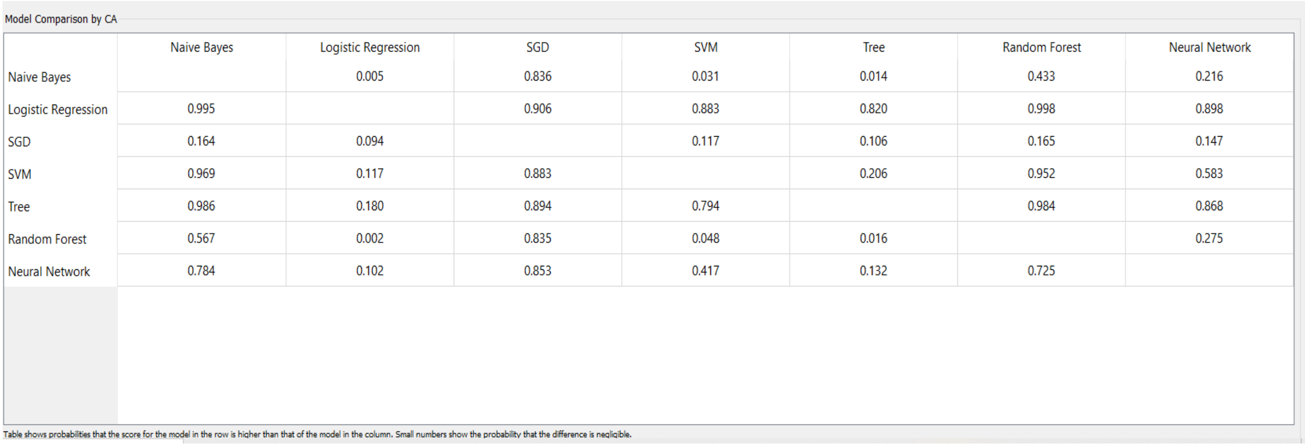
		B	M	Σ
Actual	B	308	9	317
	M	15	180	195
	Σ	323	189	512

Neural Network


Predicted

		B	M	Σ
Actual	B	312	5	317
	M	12	183	195
	Σ	324	188	512

- **Comparison of model with Classification Accuracy**


Model Comparison by CA

	Naive Bayes	Logistic Regression	SGD	SVM	Tree	Random Forest	Neural Network
Naive Bayes		0.005	0.836	0.031	0.014	0.433	0.216
Logistic Regression	0.995		0.906	0.883	0.820	0.998	0.898
SGD	0.164	0.094		0.117	0.106	0.165	0.147
SVM	0.969	0.117	0.883		0.206	0.952	0.583
Tree	0.986	0.180	0.894	0.794		0.984	0.868
Random Forest	0.567	0.002	0.835	0.048	0.016		0.275
Neural Network	0.784	0.102	0.853	0.417	0.132	0.725	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

VII. CONCLUSION

For analyse of health Dataset, different data processing technique and ML methods are available. A big challenge in data processing and machine learning field is to generate very accurate classifiers for a Medical applications. In this we employed Seven main algorithms which are: Support Vector Machine (SVM), Tree , Naive Bayes , Random Forest , Stochastic Gradient Descent ,Logistic Regression and Neural Network on the Wisconsin carcinoma(2 different type of dataset with different attributes) data. We have tried to match efficiency and effectiveness of these algorithms in terms of accuracy, precision, recall, and f1 score to seek out a simplest classification and accuracy of Naive bayes, Neural Network and SVM reaches around 97% in dataset1 while in the dataset 2 SVM has highest accuracy around 97%. Here we can conclude that, SVM has proven it’s effectiveness in carcinoma prediction and achieves the simplest performance in terms of precision, accuracy and low error rate.

SGP Paper

ORIGINALITY REPORT

29%	23%	28%	25%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	cyberleninka.org Internet Source	17%
2	Submitted to Lahore American School Student Paper	4%
3	"Advanced Intelligent Systems for Sustainable Development (AI2SD’2019)", Springer Science and Business Media LLC, 2020 Publication	3%
4	"Machine Learning, Image Processing, Network Security and Data Sciences", Springer Science and Business Media LLC, 2020 Publication	3%
5	Submitted to Charotar University of Science And Technology Student Paper	1%
6	Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science, 2016.	1%