International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

Comparative Study of Classification Algorithms of Data Mining for Possibilities of Breast Cancer

Bindu Trikha[1], Meghna Gupta[2]
[1, 2]*I.T. Department, IMS Ghaziabad (University Courses Campus), Ghaziabad, Uttar Pradesh*

Abstract: Data Mining is the technique of finding new information from the existing data on the basis of patterns that has been shown in the data to predict some conclusion from the data. Data mining has a variety of tools to find out patterns and relationship among data that can be used to predict the final results. One of the most hazardous disease among women is Breast Cancer which if not diagnosed timely, results in loss of life. This cancer is formed in the cells of the breasts. It causes the cells to change its characteristics and ends up in abnormal growth of cells that becomes a tumor. It can occur both in men and women but quite common in woman. This paper is aimed to predict the stage of breast cancer on the basis of data set for *AI for Social Good: Women Coders' Bootcamp* ***organized by Artificial Intelligence for Development in collaboration with UNDP Nepal.***
Keywords: Classification, possibilities of breast cancer, comparative analysis of algorithms using R

I. INTRODUCTION

In recent years there is a rapid growth in the occurrence of breast cancer. Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year, and also causing a large number of cancer-related deaths among women. According to "World Health Organization" statistics in 2018, it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. In order to improve breast cancer outcomes and survival, early detection is very important. Based on certain diagnostic parameters it can be predicted whether the cancer symptoms are benign or malignant. For predicting we have certain classification and prediction algorithms in data mining. In this paper we have selected the top three best out of all Algorithms which provides the comparative analysis over the topic. Here we would be comparing the accuracy of the three algorithms. These algorithms are taken into consideration to make our research more precise and it shows a clear comparison and most importantly the analysis and productive results.

A. Algorithms Selected For Comparison

We have used K- nearest neighbour, C5.0 Algorithm and Rpart Algorithm along with R language to diagnose the stages of Breast Cancer to predict whether the cancer is in beginning stage and it can be treated or in the hazardous stage i.e. it can affect other body parts also.

B. K Nearest Neighbour Algorithm

It is one of the simplest machine learning algorithms based on the idea that "objects that are 'near' each other will also have similar characteristics. So we can predict the nearest neighbour of any object if we know about the characteristics feature of any related object. It works by storing all the available data and classifies new data using a distance function.

$$d(p,q)= \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \ldots\ldots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

where $\mathbf{p} = (p_1, p_2,..., p_n)$ and $\mathbf{q} = (q_1, q_2,..., q_n)$ are two points in Euclidean n-space.

C. C5.0 Decision Tree

C5.0, is an algorithm that extracts informative patterns from data. A C5.0 model works by splitting the sample based on the field that provides the maximum **information gain**. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are re-examined, and those that do not contribute significantly to the value of the model are removed or pruned.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

D. RPart Algorithm

The rpart algorithm works by splitting the dataset recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable. Splitting rules can be constructed in can be constructed in many different ways, all of which are based on the notion of impurity- a measure of the degree of heterogeneity of the leaf nodes.

E. Software used for Comparison

We have done our experiments with C5.0 Decision Tree, K Nearest Neighbour and R Part Algorithm with R Language. Default settings are used for all compared ensemble methods. We were aware that the accuracy of some methods on some data sets can be improved when parameters were changed. However, it was difficult to find another uniform setting good for all data sets. Therefore, we did not change default settings since the default produced high accuracy on average.

F. Data Set Used

We have applied the algorithms on the dataset for Breast Cancer provided by Artificial Intelligence for Development in collaboration with UNDP Nepal. Machine learning finds extensive usage in pharmaceutical industry especially in detection of cancer cells growth. R finds application in machine learning to build models to predict the abnormal growth of cells in form of lump thereby helping in detection of cancer and benefiting the health system.

G. Acknowledgments

1) Dr. William H. Wolberg, University of Wisconsin Hospitals, Madison
2) Merishna Singh Suwal, Data Scientist at Arch Analytics, Kathmandu, Nepal

H. Context

We have used a data set of 569 patients and thereby interpreting results. The data set consists of 569 observations and 6 variables (out of which 5 numeric variables and one categorical variable) which are as follows:

1) *Diagnosis:* The diagnosis of breast tissues (1 = malignant, 0 = benign) where malignant denotes that the disease is harmful
2) *Mean_radius:* mean of distances from center to points on the perimeter
3) *Mean_texture:* standard deviation of gray-scale values
4) *Mean_perimeter:* mean size of the core tumor
5) *Mean_area:* mean area of the core tumor
6) *Mean_smoothness:* mean of local variation in radius lengths

II. EXPERIMENTAL RESULTS & DISCUSSION

A. Implementation of KNN Algorithm on the Data Set

```
> data_train_n=data_n[1:500,]
> data_test_n=data_n[501:569,]
> data_train_target=data1[1:500,6]
> data_test_target=data1[501:569,6]
> model_knn=knn(train=data_train_n,test=data_test_n,cl=data_train_target,k=13)
> model_knn
[1] B B B B M B M B M B B B B M B B B M B B M B B M B B M M B B M B M B B B M B B B B B B B B M B B B M M B
M B B B B B B B B B M M B M B M B
Levels: B M
> result_knn=confusionMatrix(model_knn,data_test_target)
> result_knn
Confusion Matrix and Statistics
          Reference
Prediction  B  M
         B 47  2
         M  1 19
```

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VII, July 2019- Available at www.ijraset.com

Accuracy : 0.9565
95% CI : (0.8782, 0.9909)
No Information Rate : 0.6957
P-Value [Acc > NIR] : 6.492e-08
Kappa : 0.8959
Mcnemar's Test P-Value : 1
Sensitivity : 0.9792
Specificity : 0.9048
Pos Pred Value : 0.9592
Neg Pred Value : 0.9500
Prevalence : 0.6957
Detection Rate : 0.6812
Detection Prevalence : 0.7101
Balanced Accuracy : 0.9420
'Positive' Class : B

B. Implementation of C5.0 Algorithm on the set

```
> model_c50=C50::C5.0(data_train[,-6],data_train[,6])
> model_c50
```

Call:
C5.0.default(x = data_train[, -6], y = data_train[, 6])
Classification Tree
Number of samples: 500
Number of predictors: 5
Tree size: 10
Non-standard options: attempt to group attributes

```
> predict_c50=predict(model_c50,data_test)
> predict_c50
```

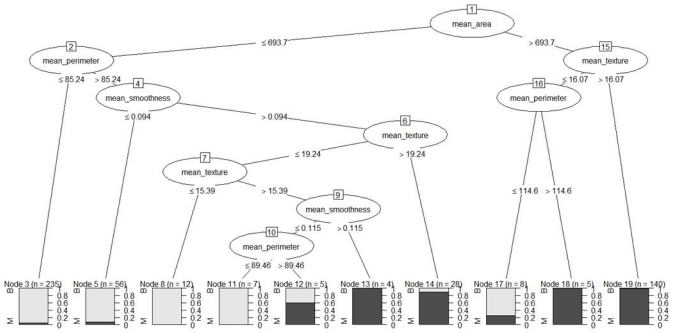
[1] M B B B M B M B M B B B B M B B B M B B M B B B B B M M B B M B M B B B M B B B B B B B B M B B B M M B M M B B B B B B M M M B M B M B
Levels: B M

```
> result_c50=confusionMatrix(predict_c50,data_test_target)
> result_c50
```

Confusion Matrix and Statistics
Reference
Prediction B M
B 45 2
M 3 19
Accuracy : 0.9275
95% CI : (0.8389, 0.9761)
No Information Rate : 0.6957
P-Value [Acc > NIR] : 2.89e-06
Kappa : 0.8311
Mcnemar's Test P-Value : 1
Sensitivity : 0.9375
Specificity : 0.9048
Pos Pred Value : 0.9574
Neg Pred Value : 0.8636
Prevalence : 0.6957
Detection Rate : 0.6522

Detection Prevalence : 0.6812
Balanced Accuracy : 0.9211
'Positive' Class : B

> plot(model_c50)



C. Implementation of Rpart Algorithm

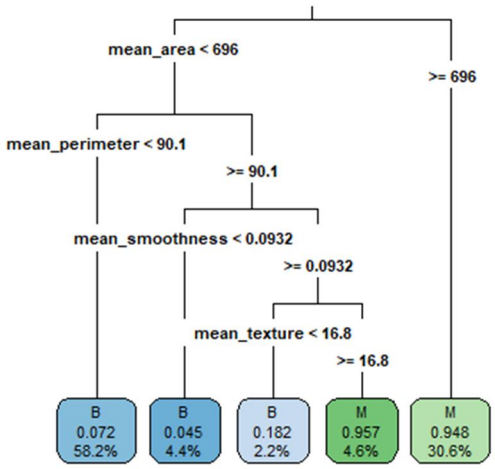
> model_rpart=rpart(diagnosis~.,data=data_train,method="class")

> model_rpart
n= 500
node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 500 191 B (0.61800000 0.38200000)
2) mean_area< 696.25 347 46 B (0.86743516 0.13256484)
4) mean_perimeter< 90.115 291 21 B (0.92783505 0.07216495) *
5) mean_perimeter>=90.115 56 25 B (0.55357143 0.44642857)
10) mean_smoothness< 0.09321 22 1 B (0.95454545 0.04545455) *
11) mean_smoothness>=0.09321 34 10 M (0.29411765 0.70588235)
22) mean_texture< 16.835 11 2 B (0.81818182 0.18181818) *
23) mean_texture>=16.835 23 1 M (0.04347826 0.95652174) *
3) mean_area>=696.25 153 8 M (0.05228758 0.94771242) *

> predict_rpart=predict(model_rpart,data_test,type="class")
> predict_rpart
294 262 4 89 554 21 456 106 425 132 8 88 59 363 16 190 32 527 403 189 386 107 330 230 220 261 562 485 111
B B B B M B M B M B B B B M B B B M M B M B B B B B M M B
195 531 5 509 114 264 417 431 118 104 20 11 222 202 31 376 546 249 352 248 380 491 277 461 391 340 139 297 347
B M B M B B M M B B B B B B B B M B B B M M B M M B B B B
126 87 193 356 401 529 400 496 24 434 327
B B B B M M M M B M B
Levels: B M

> result_rpart=confusionMatrix(predict_rpart,data_test_target)
> result_rpart
Confusion Matrix and Statistics
Reference
Prediction B M
B 44 2
M 4 19

Accuracy : 0.913
95% CI : (0.8203, 0.9674)
No Information Rate : 0.6957
P-Value [Acc > NIR] : 1.41e-05
Kappa : 0.8
Mcnemar's Test P-Value : 0.6831
Sensitivity : 0.9167
Specificity : 0.9048
Pos Pred Value : 0.9565
Neg Pred Value : 0.8261
Prevalence : 0.6957
Detection Rate : 0.6377
Detection Prevalence : 0.6667
Balanced Accuracy : 0.9107
'Positive' Class : B



> rpart.plot(model_rpart,type=3,digits=3,fallen.leaves = TRUE)

III. COMPARISON

	K Nearest Neighbour	C5.0 Algorithm	RPart Algorithm
Accuracy	0.9565	0.9275	0.913
95% CI	(0.8782,0.9909)	(0.8389,0.9761)	(0.8203,0.9674)
No –Information Rate	0.6957	0.6957	0.6957
P-value[Acc> NIR]	6.492e-08	2.89e-06	1.41e-05
Kappa	0.8959	0.8311	0.8
Mcnemar's Test P-value	1	1	0.6831
Sensitivity	0.9792	0.9375	0.9167
Specificity	0.9048	0.9048	0.9048
Pos Pred value	0.9592	0.9574	0.9565
Neg Pred value	0.9500	0.8636	0.8261
Prevalence	0.6957	0.6957	0.6957
Detection rate	0.6812	0.6522	0.6377
Detection Prevalence	0.7101	0.6812	0.6667
Balanced Accuracy	0.9420	0.9211	0.9107
'Positive' Class	B	B	B

IV. CONCLUSION

In this paper, we have presented an intelligent and effective breast cancer prediction methods using data mining. We studied an efficient approach for the extraction of significant patterns from the breast cancer data warehouses for the efficient prediction of breast cancer. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. The proposed work can be further enhanced and expanded for the automation of Breast Cancer prediction. Real data from Health care organizations and agencies needs to be collected and all the available techniques will be compared for the optimum accuracy.

In this study, K Nearest Neighbour, C5.0 Decision Tree, Rpart algorithms were implemented on a Breast Cancer Dataset to predict the potential risk in the future. Based on the three types of scenario results, KNN achieves better performance. It clearly states that the highest Balanced Accuracy is of KNN for the Data Set, so it is preferable to use KNN for this type of Data set. Whereas the C5.0 shows slightly better in the corresponding terms to RPart.

Moreover, we can conclude that the presence of breast cancer can be predicted through such methods which help us in being aware and analysing the stuff in a more efficient manner.

REFERENCES

[1] Hamid Karim Khani Zand, A Comparitive Survey on Data Mining Techniques for Breast Cancer Diagnosis and Prediction, Indian Journal of Fundamental and Applied Sciences (2231-6345) Volume 5– No.10, 2015.

[2] Sunita B.Aher, lobo L.M.R.J, Comparitive study of Classification Algorithms, International Journal of Information Technology and Knowledge Management ,Volume 5, No.2, July-december 2012

[3] Shelly Gupta, Dharminder Kumar, Anand Sharma, Data Mining Classification techniques applied for Breast Cancer Diagnosis and Prognosis, Indian Journal of Computer Science and Engineering, Vol.2 No.2, Apr-May 2011

[4] Nwokocha, Nathan, Ledisi Kabari & Agaba, Francis, Prediction of Breast Cancer Disease Using Decision Tree Algorithm International Journal of Innovative Information Systems & Technology Research 7(1):34-38, Jan.-Mar., 2019

[5] SagarS. Nikam, A Comparative Study of Classification Techniques in Data Mining Algorithms, Oriental Journal of Computer Science & Technology, ISSN: 0974-6471 April 2015, Vol. 8, No. (1): Pgs. 13-19

AUTHORS PROFILE

Ms. Bindu Trikha is working as an Assistant professor, IMS Ghaziabad University courses Campus. She has an approximate 15 years of academic experience. Her areas of interest are Data analysis, Data mining & Algorithm designing.

Ms. Meghna Gupta is working as an Assistant professor, IMS Ghaziabad University courses Campus. She has an approximate 5 years of academic experience. Her areas of interest are Data Mining & Algorithms designing.