STATISTICAL MACHINE LEARNING CLASSIFICATIONS ON PROSTATE CANCER DATASET

K.Ramakrishna Reddy[1], Dr.G.N.K.Suresh Babu[2]

[1]Ph.D Research Scholar, Visvesvaraya Technological University, Belagavi.
[1]Assistant Professor, Department of Computer Science, Acharya Institute of Graduate Studies, Bengaluru.
[2]Professor, Acharya Institute of Technology, Bengaluru.

Abstract

Cancer is the second prominent cause of death worldwide. Per annum around 6, 50,000 death cases in this current situation due to Prostate cancer. Need to improve determination the causal factors of prostate cancer. In this research work considers a medical dataset containing clinical information on 100 prostate cancer patients by using the inductive learning algorithms. This research work finds Bayes Net classifier gives an optimal results. The Bayes classifier has highest accuracy level which is 84% of accuracy. The lowest accuracy level is 62% of accuracy which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest precision level which is 0.85 of precision level. The lowest precision level is 0.62 of precision level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest recall level which is 0.84 of recall level. The lowest precision level is 0.62 of recall level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest F-Measure level which is 0.84 of F-Measure level. The lowest F-Measure level is 0.76 of F-Measure level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest ROC value level which is 0.93 of ROC level. The lowest ROC level is 0.46 of ROC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest PRC value level which is 0.92 of ROC level. The lowest ROC level is 0.51 of PRC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier.

Key words: WHO, ROC, Bayes Net, Machine learning, PRC

I Introduction

Countless researchers and clinicians are finding their ways to deal with cancer all over the world. Computer scientists have also been able to identify different risk factors and analyze the survival of cancers using various statistical [1] and machine learning techniques [3-10]. Survivability of cancer is well defined as the period from the detection of cancer until the death or conclusion of the study. Countless researchers and clinicians are finding their ways to deal with cancer all over the world. Computer scientists have also been able to identify different risk factors and analyze the survival of cancers using various statistical [11-14] and machine learning techniques [15-17]. Survivability of cancer is well defined as the period from the detection of

cancer until the death or conclusion of the study. In this research work, section 2 contains related works; in section 3 has materials and methods; in section 4 presents results and discussions and finally section 5 presents conclusion of this research work.

II Literature Survey

In the medical domain, treatments/medicines play a crucial role in determining the survival of patients. [18,19]They can define whether the patient can survive or not based on his condition.[20,21] Patients tend to improve when they receive the right treatment suited for him based on his circumstances. Various clinicians have tried to analyze survival based on different sets of medications [9,20]. Still, the significance of treatments in prostate cancer prognosis is yet to be examined with the help of data mining techniques. Data mining is expanding fast in different fields, including healthcare, and it can extract some interesting information in determining the best set of treatments for patients. Sequence mining is an area in data mining that extracts some set of frequent sequences in a set of temporal data. It has been used by some researchers to gain some insights into the biomedical area [15, 16, 22]. Machine learning techniques, on the other hand, have been proved to give better performance than conventional survival models employed in earlier studies [11, 23]. Thus, through this study, we made an attempt to analyze different machine learning classification techniques to create a survival prediction model.

III Materials and Methods

This section focuses on the materials and methods of research work. Here, the prostate cancer dataset borrowed from one of the leading dataset repository such as kaggle repository. The dataset contains 100 patients' records. Such as 100 observations and 10 variables which are as follows:

Table 1: Meta data of Prostate Cancer dataset

S.No	Label	Data type
1	Id	Integer
2	Radius	Integer
3	Texture	Integer
4	Perimeter	Integer
5	Area	Integer
6	Smoothness	Float
7	Compactness	Float
8	Symmetry	Float
9	Fractal dimension	Float
10	Diagnosis_result	Character

Methodology:

Here this research work focuses on the above mentioned dataset using following statistical machine learning algorithms in 10 cross fold validation in one of the leading open source data mining tool namely Weka 3.9.5.

- Bayes Net(BN)
- Naïve Bayes(NB)
- Naïve Bayes Multinomial(NBM)
- Niave Bayes Multinomial Text(NBMT)
- Niave Bayes Multinomial Updateable(NBMU)
- Naïve Bayes Updateable(NBU)

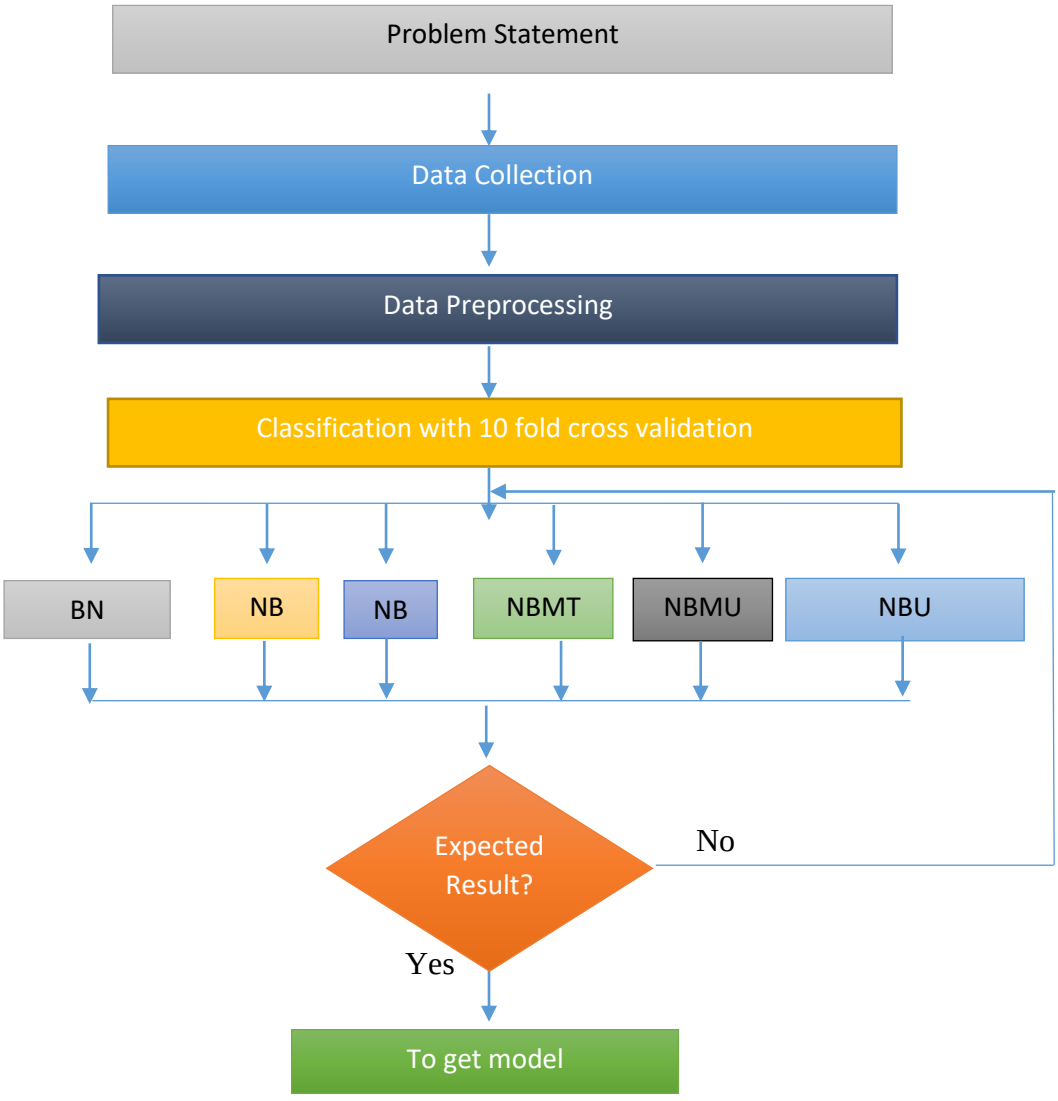

Figure 1: Proposed System Architecture

IV Results and Discussions

This section focuses on the results and discussions of this research work. The below picture shows that the attribute distribution of borrowed dataset from Kaggle dataset namely Prostate cancer dataset.[2]

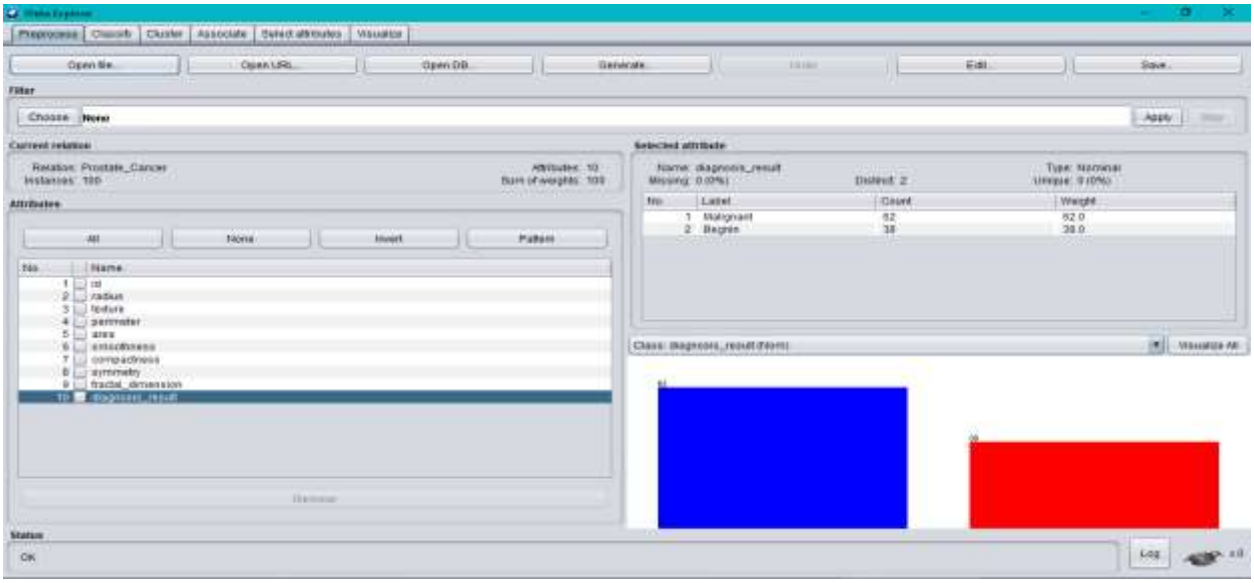

Figure 2: Distribution of attributes on Weka.3.9.0

The below table shows that the various outcomes of statistical machine learning algorithms in 10:90 fold cross validation.

Table 2: Various Bayes Classifiers and their measurements

Classifiers	Accuracy	Precision	Recall	F- Measure	ROC	PRC	Time taken to build model (In Sec.)
Bayes Net	84	0.85	0.84	0.84	0.93	0.92	0.05
Naïve Bayes	79	0.81	0.79	0.79	0.89	0.88	0.01

Naïve Bayes Multinomial	81	0.82	0.81	0.81	0.88	0.88	0.02
Niave Bayes Multinomial Text	62	0.62	0.62	0.76	0.46	0.51	0
Niave Bayes Multinomial Updateable	81	0.82	0.81	0.81	0.88	0.88	0
Naïve Bayes Updateable	79	0.81	0.79	0.79	0.9	0.88	0

The Bayes Net classifier produces 84% of accuracy value,0.85 of precision value,0.84 of recall value,0.84 of F-Measure value,0.93 of receiver operating characteristic curve value,0.92 of precision recall value and it takes 0.05 time consumption to build a model. The Naïve Bayes classifier produces 79% of accuracy value,0.81 of precision value,0.79 of recall value,0.79 of F-Measure value,0.89 of receiver operating characteristic curve value,0.88 of precision recall value and 0.01 seconds takes a time consumption to build a model. The Naïve Bayes Multinomial classifier produces 81% of accuracy value,0.82 of precision value,0.81 of recall value,0.81 of F-Measure value,0.88 of receiver operating characteristic curve value,0.88 of precision recall value and 0.02 seconds takes a time consumption to build a model. The Naïve Bayes Multinomial Text classifier produces 62% of accuracy value,0.62 of precision value,0.62 of recall value,0.76 of F-Measure value,0.46 of receiver operating characteristic curve value,0.51 of precision recall value and 0 seconds takes a time consumption to build a model. The Naïve Bayes Multinomial Updateable classifier produces 81% of accuracy value,0.82 of precision value,0.81 of recall value,0.81 of F-Measure value,0.88 of receiver operating characteristic curve value,0.88of precision recall value and 0 takes a time consumption to build a model. The Naïve Bayes Updateable classifier produces 79% of accuracy value,0.81 of precision value,0.79 of recall value,0.79 of F-Measure value,0.90 of receiver operating characteristic curve value,0.88 of precision recall value and 0 seconds takes a time consumption to build a model.

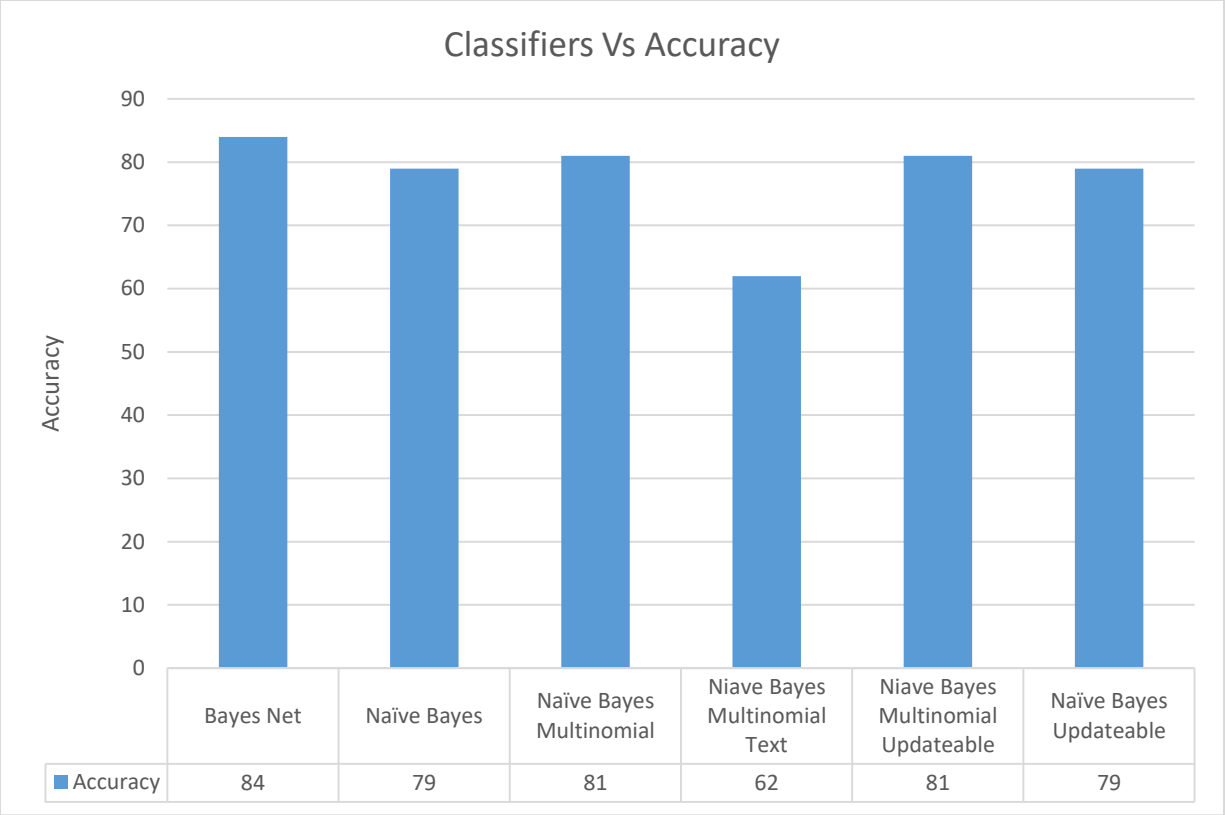

Figure 3: Various Bayes algorithms and their accuracy values

The above diagram shows that the various statistical classifiers and their accuracy levels. The Bayes classifier has highest accuracy level which is 84% of accuracy. The lowest accuracy level is 62% of accuracy which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes and Naïve Bayes Updateable has produces next lowest accuracy level as well as same accuracy level which is 79% of accuracy level. The Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has same accuracy level which is 81% of accuracy.

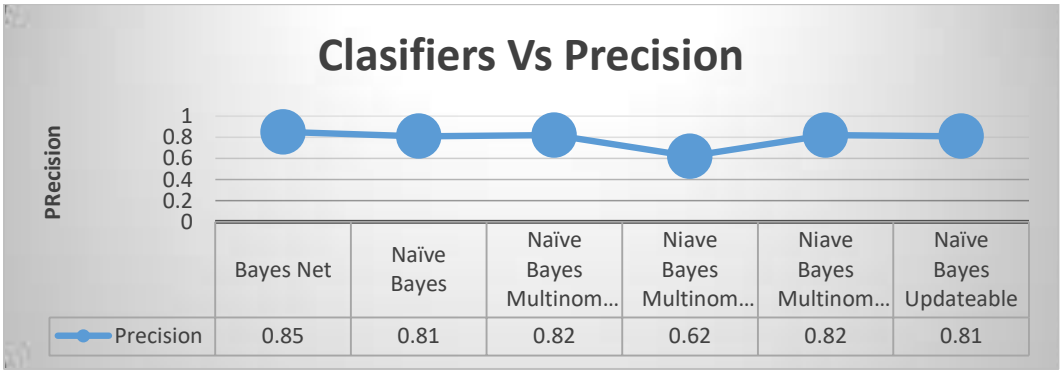

Figure 4: Various Bayes algorithms and their precision values

The above diagram shows that the various statistical classifiers and their precision levels. The Bayes classifier has highest precision level which is 0.85 of precision level. The lowest precision level is 0.62 of precision level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes and Naïve Bayes Updateable has produces next lowest precision level as well as same precision level which is 0.81 of precision level. The Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has same precision level which is 0.82 of precision value.

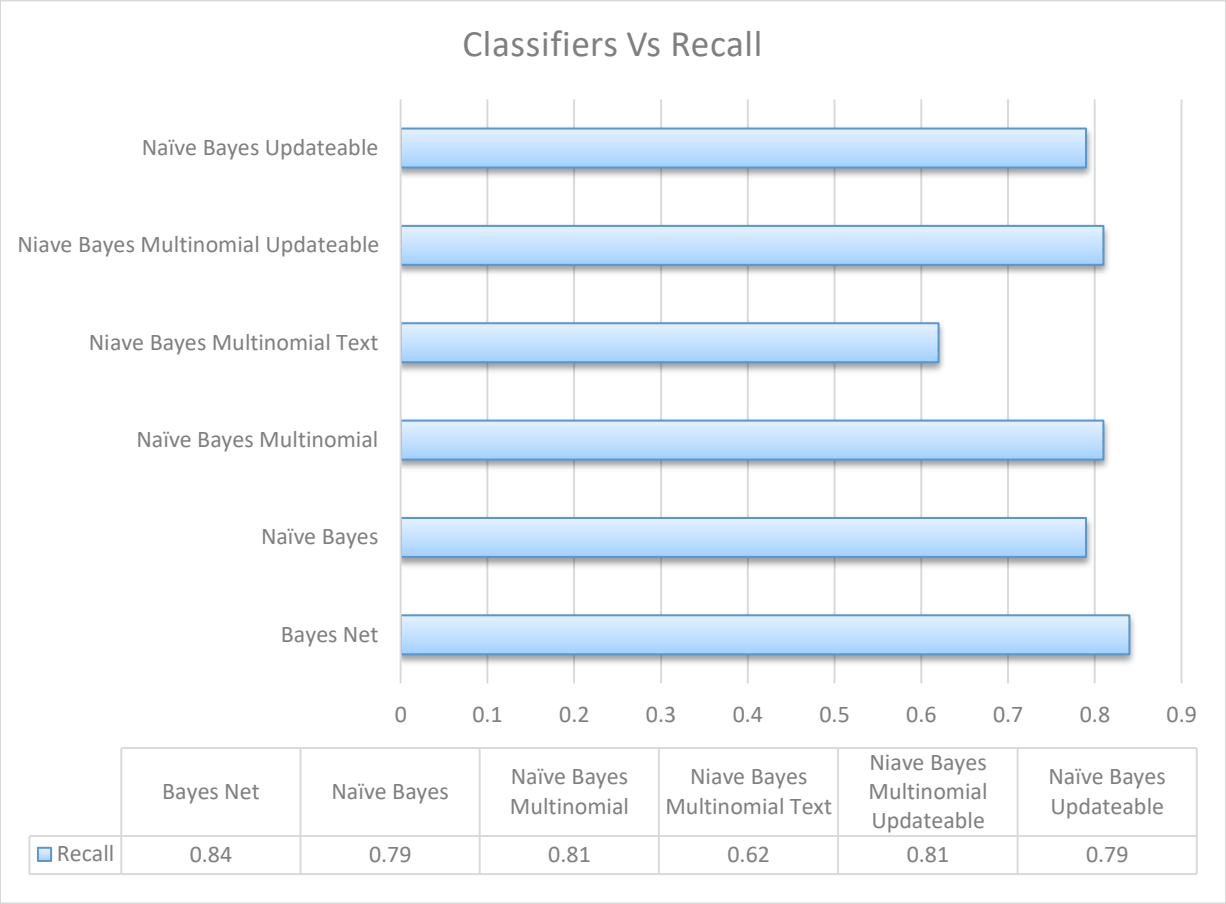

Figure 5: Various Bayes algorithms and their Recall values

The above diagram shows that the various statistical classifiers and their recall levels. The Bayes classifier has highest recall level which is 0.84 of recall level. The lowest precision level is 0.62 of recall level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes and Naïve Bayes Updateable has produces next lowest recall level as well as same recall level which is 0.79 of recall level. The Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has same recall level which is 0.81 of recall value.

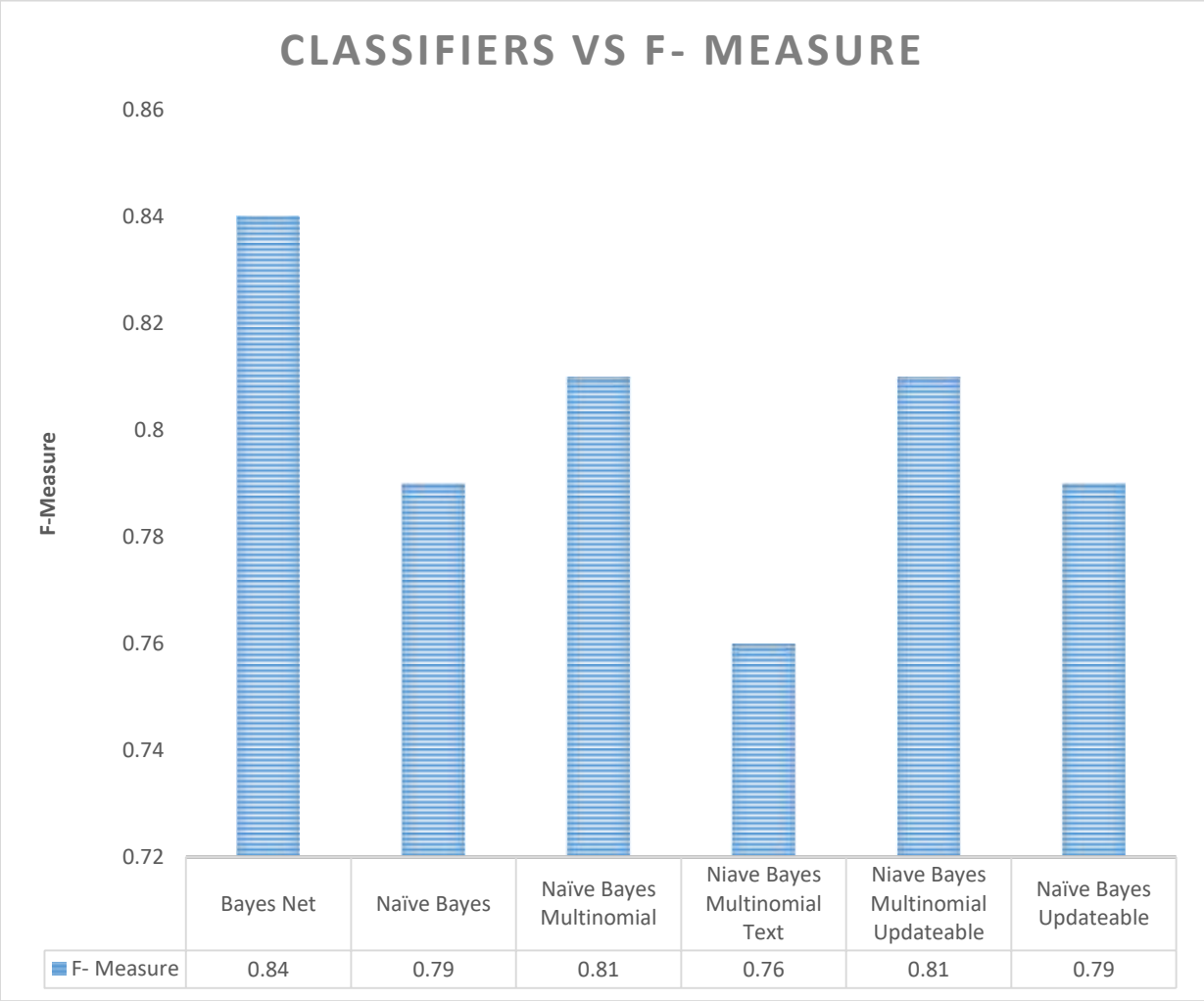

Figure 5: Various Bayes algorithms and their F-Mesure values

The above diagram shows that the various statistical classifiers and their F-Measure levels. The Bayes classifier has highest F-Measure level which is 0.84 of F-Measure level. The lowest F-Measure level is 0.76 of F-Measure level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes and Naïve Bayes Updateable has produces next lowest recall level as well as same F-Measure level which is 0.79 of F-Measure level. The Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has same F-Measure level which is 0.81 of F-Measure value.

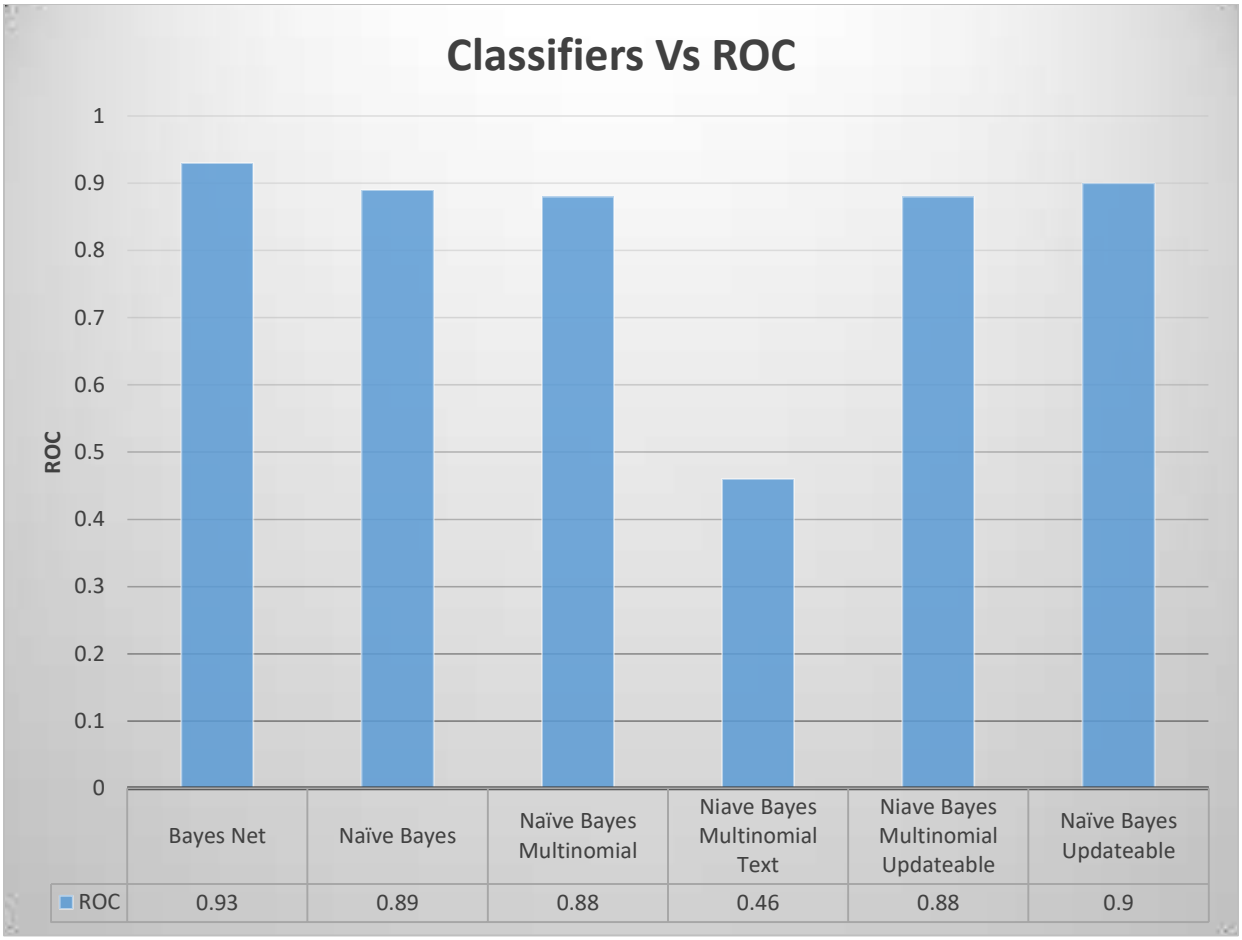

Figure 6: Various Bayes algorithms and their ROC values

The above diagram shows that the various statistical classifiers and their ROC values. The Bayes classifier has highest ROC value level which is 0.93 of ROC level. The lowest ROC level is 0.46 of ROC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes and Naïve Bayes Updateable has produces next lowest ROC value level as well as same ROC level which is 0.88 of ROC level. The Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has more or less same ROC value level which is 0.89 and 0.90 of ROC values.

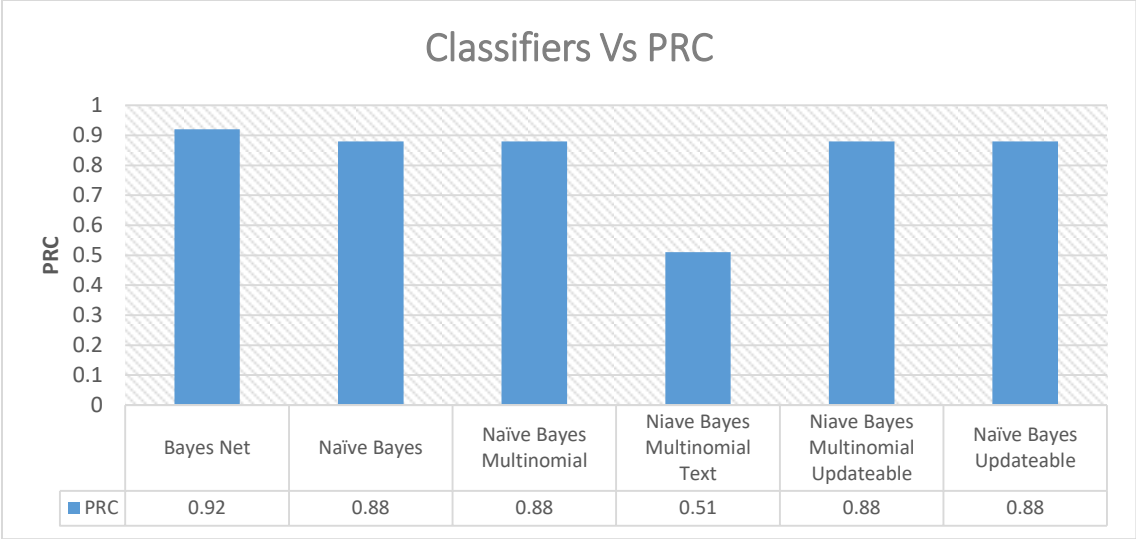

Figure 7: Various Bayes algorithms and their PRC values

The above diagram shows that the various statistical classifiers and their PRC values. The Bayes classifier has highest PRC value level which is 0.92 of ROC level. The lowest ROC level is 0.51 of PRC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Naïve Bayes , Naïve Bayes Updateable, Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable classifiers has produces next lowest PRC value level as well as same PRC level which is 0.88 of PRC level.

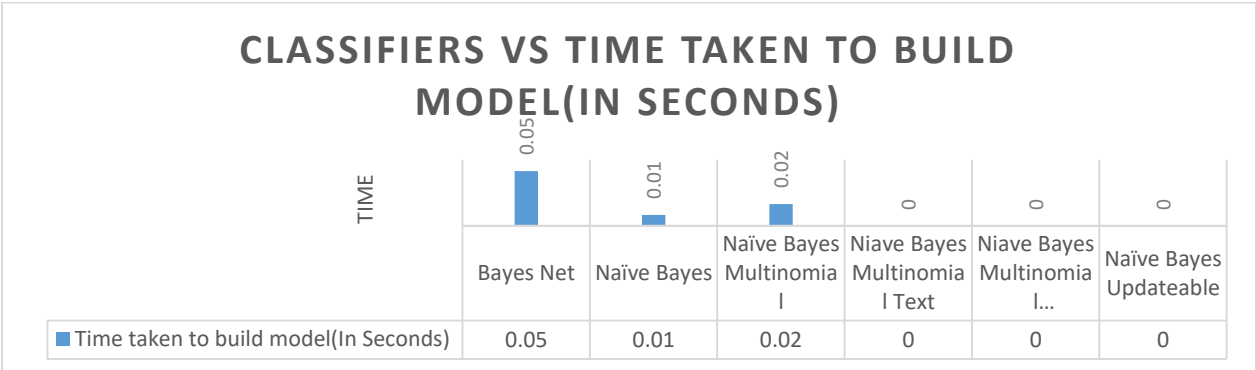

Figure 8: Various Bayes algorithms and their time taken to build models

The above diagram shows that the various statistical classifiers and their time consumption to build the models. The Bayes classifier has taken more time to build a model which is 0.05 seconds. The Niave Bayes Multinomial Text, Niave Bayes Multinomial Updateable and Naïve Bayes Updateable takes zero seconds to build a model.

V. Conclusion

This research work concludes that the The Bayes classifier has highest accuracy level which is 84% of accuracy. The lowest accuracy level is 62% of accuracy which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest precision level which is 0.85 of precision level. The lowest precision level is 0.62 of precision level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest recall level which is 0.84 of recall level. The lowest precision level is 0.62 of recall level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest F-Measure level which is 0.84 of F-Measure level. The lowest F-Measure level is 0.76 of F-Measure level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest ROC value level which is 0.93 of ROC level. The lowest ROC level is 0.46 of ROC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. The Bayes classifier has highest PRC value level which is 0.92 of ROC level. The lowest ROC level is 0.51 of PRC level which is produced by Naïve Bayes Multinomial Text classifier of Bayes classifier. This system recommends that the Bayes net classifier produces optimal results compare with other models.

References

[1] Ishleen Kaur, M.N. Doja, Tanvir Ahmad, Time-range based sequential mining for survival prediction in prostate cancer, Journal of Biomedical Informatics,Volume 110,2020,103550,ISSN 1532-0464,https://doi.org/10.1016/j.jbi.2020.103550.

[2] https://www.kaggle.com/sajidsaifi/prostate-cancer

[3]‘Prostate cancer curable if detected early’, The Hindu, Hyderabad, SEPTEMBER 27, 2014.

[4] Prostate Cancer ICMR. Available at: http://cancerindia.org.in/prostate-cancer/.

[5] R.J. Kate, R. Nadig,Stage-specific predictive models for breast cancer survivability,Int. J. Med. Inf., 97 (2017), pp. 304-311

[6] M. Jajroudi, T. Baniasadi, L. Kamkar, F. Arbabi, M. Sanei, M. Ahmadzade,Prediction of survival in thyroid cancer using data mining technique,Technol. Cancer Res. Treat., 13 (4) (2014), pp. 353-359

[7] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, C.-N. Lin,The application of data mining techniques to oral cancer prognosis.J. Med. Syst., 39 (5) (2015)

[8] M.S. Santos, P.H. Abreu, P.J. García-Laencina, A. Simão, A. Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients,J. Biomed. Inform., 58 (2015), pp. 49-59

[9] K. Malhotra, S.B. Navathe, D.H. Chau, C. Hadjipanayis, J. Sun, Constraint based temporal event sequence mining for Glioblastoma survival prediction, J. Biomed. Inform., 61 (2016), pp. 267-275

[10] A. Shknevsky, Y. Shahar, R. Moskovitch, Consistent discovery of frequent interval-based temporal patterns in chronic patients’ data, J. Biomed. Inform., 75 (2017), pp. 83-95, 10.1016/j.jbi.2017.10.002

[11] A.J. Steele, S.C. Denaxas, A.D. Shah, H. Hemingway, N.M. Luscombe, Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease, PLoS ONE, 13 (8) (2018), p. e0202344

[12] N. Shukla, M. Hagenbuchner, K.T. Win, J. Yang, Breast cancer data analysis for survivability studies and prediction, Comput. Methods Programs Biomed., 155 (2018), pp. 199-208
[13] K. Park, A. Ali, D. Kim, Y. An, M. Kim, H. Shin, Robust predictive model for evaluating breast cancer survivability, Eng. Appl. Artif. Intell., 26 (9) (2013), pp. 2194-2205
[14] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Comput. Biol. Med., 59 (2015), pp. 125-133
[15] S. Walczak, V. Velanovich, Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks, Decis. Support Syst., 106 (2018), pp. 110-118
[16] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst., 74 (2015), pp. 150-161
[17] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, A. Dag, A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival, Expert Syst. Appl. (2019), p. 112863, 10.1016/j.eswa.2019.112863
[18] K.-J. Wang, B. Makond, K.-H. Chen, K.-M. Wang, A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, Appl. Soft Comput., 20 (2014), pp. 15-24
[19]Mohammed J. Zaki, Wagner Meira Jr.,Data Mining and Analysis, Cambridge University Press (2014)
[20]K. Fukui, Y. Okada, K. Satoh, M. Numao, Cluster sequence mining from event sequence data and its application to damage correlation analysis, Knowl.-Based Syst. (2019), 10.1016/j.knosys.2019.05.012
[21]M. Taub, R. Azevedo, A.E. Bradbury, G.C. Millar, J. Lester, Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment, Learn. Instruct., 54 (2018), pp. 93-103, 10.1016/j.learninstruc.2017.08.005
[22] K. Amin, J.S. Shah, Sequential sequence mining technique in large database of gene sequence, 2010 International Conference on Computational Intelligence and Communication Networks (2010), 10.1109/cicn.2010.142
[23] J. Zhang, Y. Wang, C. Zhang, Y. Shi, Mining contiguous sequential generators in biological sequences, IEEE/ACM Trans. Comput. Biol. Bioinf., 13 (5) (2016), pp. 855-867, 10.1109/tcbb.2015.2495132