Recurrence of Cancer Prediction using ML techniques

Khalil Ahmad, S.
Department of ComputerScience
Air University
PAF Complex, E9, Islamabad
170420@students.au.edu.pk

Dil Khan, K.
Department of ComputerScience
Air University
PAF Complex, E9, Islamabad
170344@students.au.edu.pk

Rehman, H.
Department of ComputerScience
Air University
PAF Complex, E9, Islamabad
170288@students.au.edu.pk

***Abstract*— Recurrence is an important keystone in breast cancer behavior, the most common cancer among the women is breast cancer except skin cancer. Breast cancer can occur in both men and women but its far more common in women. According to a survey 13 percent women are affected by it all over the world. This research evaluates several Machine learning to detect and predict the recurrence of breast cancer; and compare the used models by using different metrics like accuracy, precision, etc.**

Keywords—Cancer, Machine Learning, Predictive models, classification

I. INTRODUCTION

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can ease the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has directed many researchers, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods.

Breast cancer (BC) is one of the most common cancers among women worldwide, representing most new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem today. Almost 13% women are affected by breast cancer all over the world according to the survey of World Health Organization (WHO).

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of kind or gentle tumors can prevent patients undergoing unnecessary treatments. So, the correct diagnosis of BC and classification of patients into malicious or gentle groups is the subject of much research. we are trying to use a dataset of patients that had the BC, got treatment and won the battle against this disease. We will be predicting the recurrence of BC by using different models of Machine Learning (ML) and comparing the results like accuracy score, Entropy, Gini index, Precision, information gain, etc.

Classification and information mining strategies are a powerful method to group information. Particularly in clinical field, where those strategies are generally utilized in conclusion and investigation to decide.

II. RELATED LITRETURE

we are not the first who is doing this work nor the last one. There is a lot of people who are working in this field. we also took help from the work of those people. A literature review shows that there have been some studies identified with this theme using statistical approaches, artificial neural networks however could find only a couple of studies utilizing data mining techniques like Decision trees and Random Forest. Abreu, Santos and Andrade used decision trees, naïve Bayes, logistic regression, K-means Algorithm, begging and boosting to predict the recurrence of breast cancer [1]. Ahmad LG, Eshlaghy AT, Poorbrahimi A, Ebrahimi M and Razavi AR used decision tree, support vector machine and 10-Fold cross validation for predicting breast cancer recurrence [3]. Chang ming, Velaria viassolo, Nicole Probst-Hensch, Pierre O. Chappuis, Ivo D. Dinov and Maria C. Katapodi used classification and Clustering approaches of ML techniques that are model-based and model-free for predictive analytics. The model-based approaches included GLM, LOGIT, LDA, and QDA while the model-free predictive analytics involved ADA, RF, KNN [2].

A. Abbreviations and Acronyms

- Breast Cancer (BC)
- Machine Learning (ML)
- World Health Organization (WHO)
- Generalized Linear Models (GLM)
- Logistic Regression (LOGIT)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Adaptive Boosting (ADA)
- Random Forest (RF)
- K- nearest Neighbors (KNN)
- Decision Tree (DT)

B. Equations

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(fi - yi)^2 \quad (1)$$

Where

N= number of data points.
fi =value returned by model.
yi =actual value for data point i.

$$E(S) = -\sum_{i=1}^{c} Pi \log_2 Pi \quad (2)$$

Where

Pi =probability of class.

$$IG = E(Y) - E(Y|X) \quad (3)$$

Where

E(Y)= Entropy of selected class.

$$Accuracy = \frac{TR+TN}{TP+TN+FP+FN} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Specificity = \frac{TN}{TN+FP} \quad (7)$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \quad (9)$$

Where

TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative respectively.

III. Proposed Methodology

The materials we used for this task are Google Colab for coding and breast cancer recurrence data from the UCI repository. Our methodology involves different machine learning approaches such as Random Forest, Decision tree, K-nearest neighbor, and naïve Bayes.

We collected the dataset from the UCI repository and perform some preprocessing such as missing values and scaling. As we have three options for missing values i.e. remove the instances with missing values, but we already have the small dataset, the other option is to replace the missing values with the mode of the feature having missing values but it may affect the prediction and gives ambiguous results, we used the third one we applied an ML algorithm on the feature having missing values by making it dependent on other features of the dataset including the label. That gives us results and we put those results for missing values in the dataset. Missing values were found in the feature "node-caps" After performing preprocessing, we trained some ML classification models on the dataset and compare the results. We have 9 independent features and one class label in the dataset.

1. Age
2. Meno-pause
3. Tumor-size
4. Inv-nodes
5. Node-caps
6. Deg-malign
7. Breast
8. Breast-quad
9. Irradiate
10. Class (label)

A. *Block diagram*

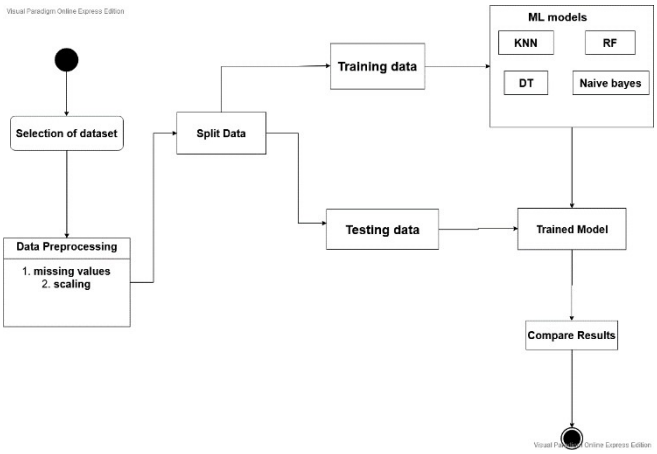

Fig 1. Block diagram of methodology

IV. Experiments and Results.

A. ***Random Forest***

Table 1. confusion matrix RF

	No-recurrence	**recurrence**
No-recurrence	18	3
recurrence	2	6

Table 2. classification report for RF

	precision	recall	F1-score	Support
No-recurrence	0.90	0.86	0.88	21
recurrence	0.67	0.75	0.78	8
accuracy			0.83	29
Macro-avg	0.78	0.80	0.79	29
Weighted avg	0.84	0.83	0.83	29

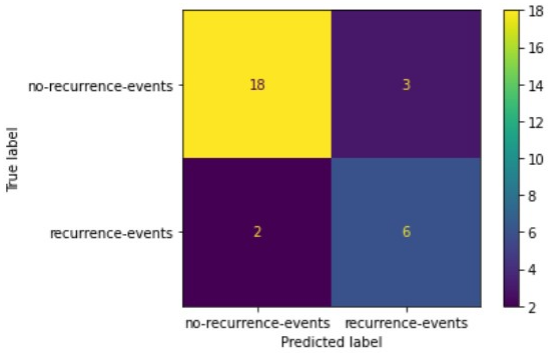

Fig 2. confusion matrix

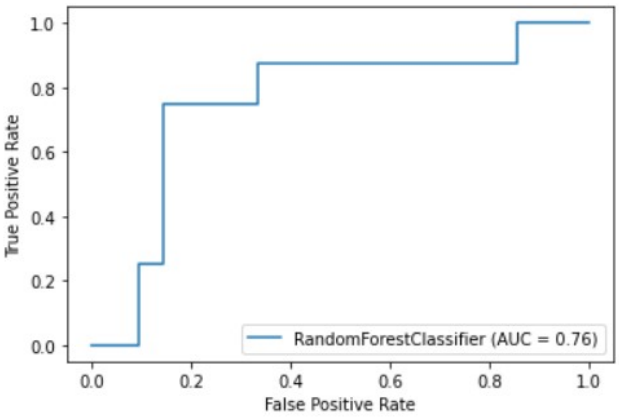

Fig 3. ROC curve for RF

Accuracy =82.75%

B. *K-Nearest Neighbor*

Table 3. confusion matrix KNN

	No-recurrence	recurrence
No-recurrence	19	2
recurrence	3	5

Table 4. classification report for KNN

	precision	recall	F1-score	Support
No-recurrence	0.86	0.90	0.88	21
recurrence	0.71	0.62	0.67	8
accuracy			0.83	29
Macro-avg	0.79	0.76	0.78	29
Weighted avg	0.82	0.83	0.82	29

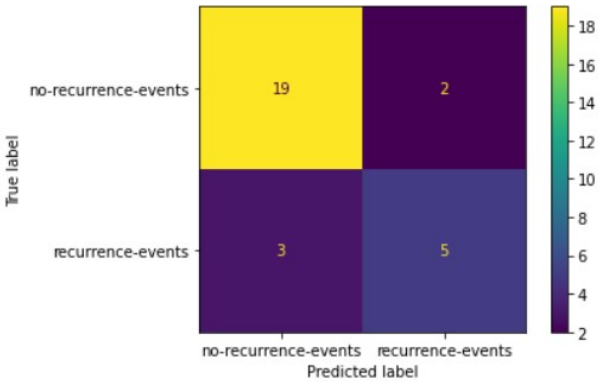

Fig 4. confusion matrix

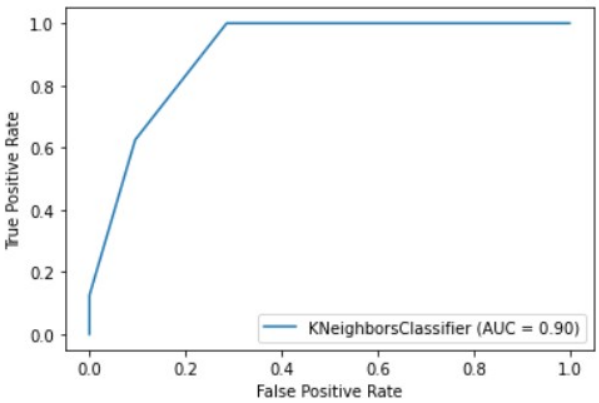

Fig 5. ROC curve for KNN.

Accuracy= 82.75%

C. *Decision Tree*

Table 5. confusion matrix for DT

	No-recurrence	recurrence
No-recurrence	22	0
recurrence	5	2

Table 6. classification report for DT

	precision	recall	F1-score	Support
No-recurrence	0.81	1	0.90	22
recurrence	1	0.29	0.44	7
accuracy			0.83	29
Macro-avg	0.91	0.64	0.67	29
Weighted avg	0.86	0.83	0.79	29

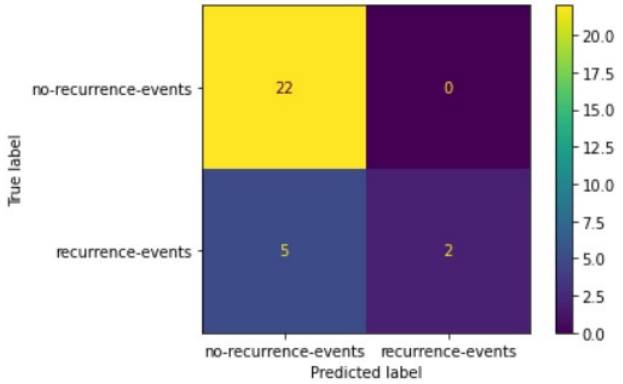

Fig 6. confusion matrix

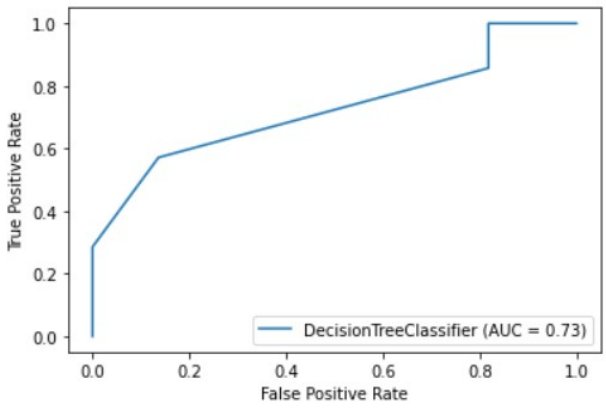

Fig 7. ROC curve for DT

Accuracy=82.75%

D. *Naïve Bayes*

Table 7. confusion matrix for Naïve Bayes

	No-recurrence	**recurrence**
No-recurrence	18	3
recurrence	3	5

Table 8. classification report for Naïve Bayes

	precision	recall	F1-score	Support
No-recurrence	0.86	0.86	0.86	21
recurrence	0.62	0.72	0.72	8
accuracy			0.79	29
Macro-avg	0.74	0.74	0.74	29
Weighted avg	0.79	0.79	0.79	29

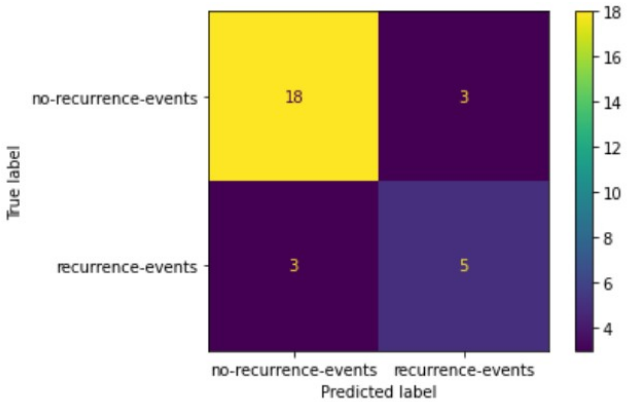

Fig 8. confusion matrix

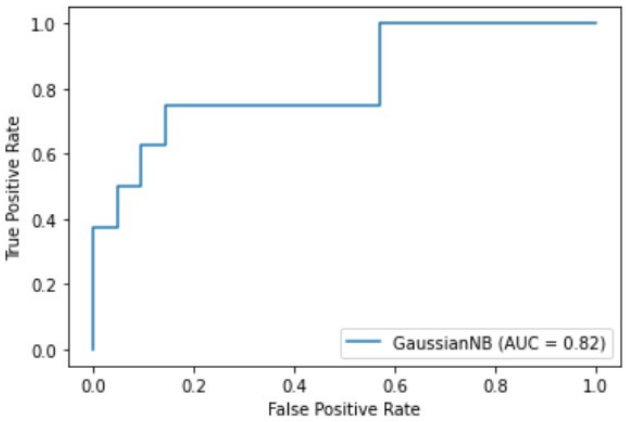

Fig 9. ROC curve for Naïve Bayes

Accuracy=79.31%

V. Discussion

Firstly, trained a RF model using 80% of the data that model was 76% accurate, and the random state we used was 10. After that we set the random state variable 8 and check the accuracy. This time accuracy decreased to 70% then we used the 10% data for testing purpose and the model was 79% accurate because by increasing the testing data the accuracy also increases, we set n_estimater variable to 100. Using 10% data for testing and by setting the n_estimators' value.

The next model is KNN when the model was trained on 80% data and the value of the n_neighbor set by 1. The model was 98% accurate and for 90% training data model become 100% accurate its now the model is over fitting.

For value of n_neighbor 2 the model accuracy decreased to 74% and 79% for 80% and 90% training data respectively.

For decision tree the accuracy 74% and 83% for 20% and 10% testing data respectively and we used entropy (information gain) as criterion and max depth of tree was 3.

The last one is naïve bayes we did same as above first train model for 20% testing data and later for 10%. The accuracy for 80% training data is 74% and for 90% training data accuracy increased to 79%.

VI. Conclusion

Predicting recurrence is a key point in BC context. It's a fact that the researchers have tried to address this problem, but it remains an open challenge. Based on the analysis, the works utilizing the different ML approaches have an advantage over working with a single technique.

The important aspect of this problem is not completely addressed but the implementation of ML strategies is only possible if the data has a enough records and also a balanced dataset will be a plus point. In this task, we used four different techniques for predicting the recurrence of Breast cancer. After comparing the results, found that the random forest is the best classifier predictor with the test dataset. if we compare the accuracy then the RF and DT both have the same accuracy but after comparing the values of the confusion matrix, RF predicted more accurately than DT.

VII. Acknowledgements

We are very thankful to the UCI machine learning repository for providing the dataset.
We also thankful to Dr. Mehdi Hassan for guiding us. Without his support, it would have been difficult for us to prepare the paper.

References

[1] Henriques Abreu, Pedro & Santos, Miriam & Henriques Abreu, Miguel & Aveleira Andrade, Bruno & Silva, Daniel. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Computing Surveys. 49. 1-40. 10.1145/2988544.

[2] Ming, C., Viassolo, V., Probst-Hensch, N. *et al.* Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. *Breast Cancer Res* **21,** 75 (2019). https://doi.org/10.1186/s13058-019-1158-4

[3] Ghasem Ahmad, Leila & Eshlaghy, A. & Pourebrahimi, Alireza & Ebrahimi, Mansour & Razavi, A.. (2013). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health & Medical Informatics. 4. 124-130.

[4] Rana, M., Chandorkar, P., Dsouza, A., Kazi, N. J. I. I. J. o. R. i. E., & eISSN, T. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. 2319-1163.

[5] Navlani, A. (2018, December 4). Naive Bayes Classification using Scikit-learn. Datacamp. https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

[6] Robinson, S. (n.d.). K-Nearest Neighbors Algorithm in Python and Scikit-Learn. Stack Abuse. https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn

[7] Navlani, A. (2018a, December 4). Decision Tree Classification in Python. Data Camp. https://www.datacamp.com/community/tutorials/decision-tree-classification-python

[8] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, Procedia Computer Science, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.224.