**Breast Cancer Prediction using Machine learning**

**Aayush Bhatta**

Fisk University, Nashville TN, USA

**Abstract**

With the rapid population growth, the risk of death incurred by breast cancer is rising exponentially. Breast cancer is the second most severe cancer among all the cancers already unveiled and the major cause of death in women worldwide. Any improvement in the treatment of breast cancer will lead to a healthy life of people. Therefore, machine learning techniques can contribute to the process of prediction and early diagnosis of breast cancer. So, in this paper, I will be comparing three machine learning techniques i.e., k-nearest neighbors (KNN), random forest classifier, and Decision Tree Classifier (DT) and find out the most effective way of predicting the breast cancer by comparing their accuracy. Moreover, I will be measuring the performance based on the sensitivity, specificity, precision, True positive rate, False positive rate, True negative rate, False Negative rate, and F1 score. Work will be done in the anaconda environment using python. In the end, the accuracy of k Neighbors classifier is 88.3% and decision tree classifier is 91.81%; however, the random classifier achieved the highest score of 94.15%.

**Keywords:** Machine Learning (ML), Breast Cancer, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbor.

**INTRODUCTION**

Breast cancer is the most common cancer among today’s generation. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. Today, one in 5 people worldwide will develop cancer during their lifetime. There are different caused of breast cancer. One of the reasons is the gene. If someone in the family has a history of cancer, then there is the possibility that the coming generation can develop the cancer during their lifetime. More women are diagnosed with breast cancer than any other type of cancer, besides skin cancer. According to WHO, 684,995 people died due to breast cancer. Women are dying more in comparison to men due to breast cancer. Moreover, if we want to see the rate of people diagnosed with invasive breast cancer over the years, then we can see that 1 out of 11 people were diagnosed with the breast cancer in 1975; however, now in 2021, 1 out of 8 people are being diagnosed with the breast cancer. We can see how the rate is increasing over the years, so if there is not early treatment or a way of diagnosing early cancer then it will be too late. Therefore, to prevent or to provide more time to the patients to live is to find out the breast cancer as soon as possible. There might be different ways of doing this; however, in this paper I will be using machine learning to find the more accurate way of predicting the breast cancer among the people based on the breast lump that comes in the breast of most of the people. I will be using number of data collected from Kaggle to make different models to train and test them and find the most efficient way or an efficient technique to find more accurate result using machine learning.

**RELATED WORK**

Many machine learning algorithms are out there for predicting breast cancer. There are a lot of researchers out there who are working on diagnosing the breast cancer using machine learning algorithms. An author Anusha Bharat used machine learning algorithms for breast cancer risk prediction and diagnosis by using support vector machine (SVM), decision tree (CART), Naïve Bayes (NB), and k Nearest Neighbors (KNN). She used Wisconsin breast cancer dataset to perform her algorithm. Similarly, Pedro did systematic review on predicting breast cancer recurrence using Machine Learning Techniques. He used published works that used machine learning techniques in local and open-source databases between 1997 and 2014. He came to know that it is difficult to obtain a representative dataset for breast cancer recurrence and there is not any consensus on the best prediction for the disease. However, high accuracy results can be obtained, which can be beneficial for predicting the disease. In the same way, Montazeri researched on machine learning models in breast cancer survival prediction. He found out that with early diagnosis of breast cancer, the survival rate will increase from 56% to more than 86%. Machine learning can help physicians to reduce the number of false decisions. He used a rule-based classification method with machine learning techniques for the prediction of different types of breast cancer survival. Naive Bayes (NB), random forest, support vector machine, and other 10-cross fold technique were used with the proposed model for the prediction of breast cancer survival. Overall, random forest model came out as a rule-based classification model which was the best model with the highest level of accuracy. Similarly, my research is also focused on finding the best way of predicting the breast cancer using different techniques / algorithms of machine learning.

**PROBLEM**

It has been a huge problem that breast cancer can not be eradicated from our society yet. Therefore, the only way of controlling it is to diagnose it as early as possible, so the patient can get the proper treatment as soon as possible. However, due to various reasons, it becomes difficult to find it out.

Chart

Description automatically generated

In the given figure, we can clearly see how the race of the person is one of the factors of higher mortality rate in black in comparison to other races. There might be different factors behind it, so early prediction / diagnosis of breast cancer can really help in early treatment of those people.

Chart, bar chart

Description automatically generated

Similarly, here we can see how the death rate varies on different age group people. If there was a proper way of predicting the breast cancer as early as possible then the mortality rate of other age group could possibly be decreased. Therefore, my main hypothesis is that if we use machine learning to predict the breast cancer in its early period then we can decrease the mortality rate and increase the survival rate of the people out there.

**DESIGN**

In this project, I am trying to predict the suitable / efficient way of predicting the breast cancer by using different machine learning algorithms. I applied three machine learning algorithms i.e., K Nearest Neighbor classifier, decision tree classifier and random tree classifier. At first, I imported the dataset and visualized my dataset. After that I created the correlation matrix to compare the relationship of various features. Then, from the dataset that I had, I divided some of them as training dataset, and some as testing dataset. Then I created the machine learning model using them. Then, I created the confusion matrix and heatmaps. Additionally, to be more accurate, I find out the accuracy, precision, sensitivity, specificity, and f1 score. After that, I compared the total positive rate and total negative rate by creating the ROC curve. And comes the main part when I used different machine learning models like K Neighbors classifier, decision tree classifier, and random forest classifier to compare the data.

**Visualizing the data**

**Table

Description automatically generated**

**Chart, bar chart, histogram

Description automatically generated**

**Correlation Matrix**

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses

**Table

Description automatically generated**

**RESULT**

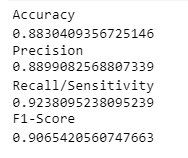
* **Heatmaps**

A heatmap is a graphical representation where individual values of a matrix are represented as colors. A heatmap is very useful in visualizing the concentration of values between two dimensions of a matrix. This helps in finding patterns and gives a perspective of depth.

**Chart, treemap chart

Description automatically generated**

Here, we can see the true positive/negative rate and false positive/negative rate of people suffering from breast cancer. 0 indicates that the person is healthy and 1 indicates that the person is diagnosed with breast cancer.



* **ROC Curve**

ROC curve, also known as Receiver Operating Characteristics Curve, is a metric used to measure the performance of a classifier model. The ROC curve depicts the rate of true positives with respect to the rate of false positives, therefore highlighting the sensitivity of the classifier model.

**Chart, line chart

Description automatically generated**

****

* **Clustering dataset through Scikit Learn**

**Chart, scatter chart

Description automatically generated**

Here, we used the dataset that is available to show on a graph using scikit learn.

* **Machine Learning Models**

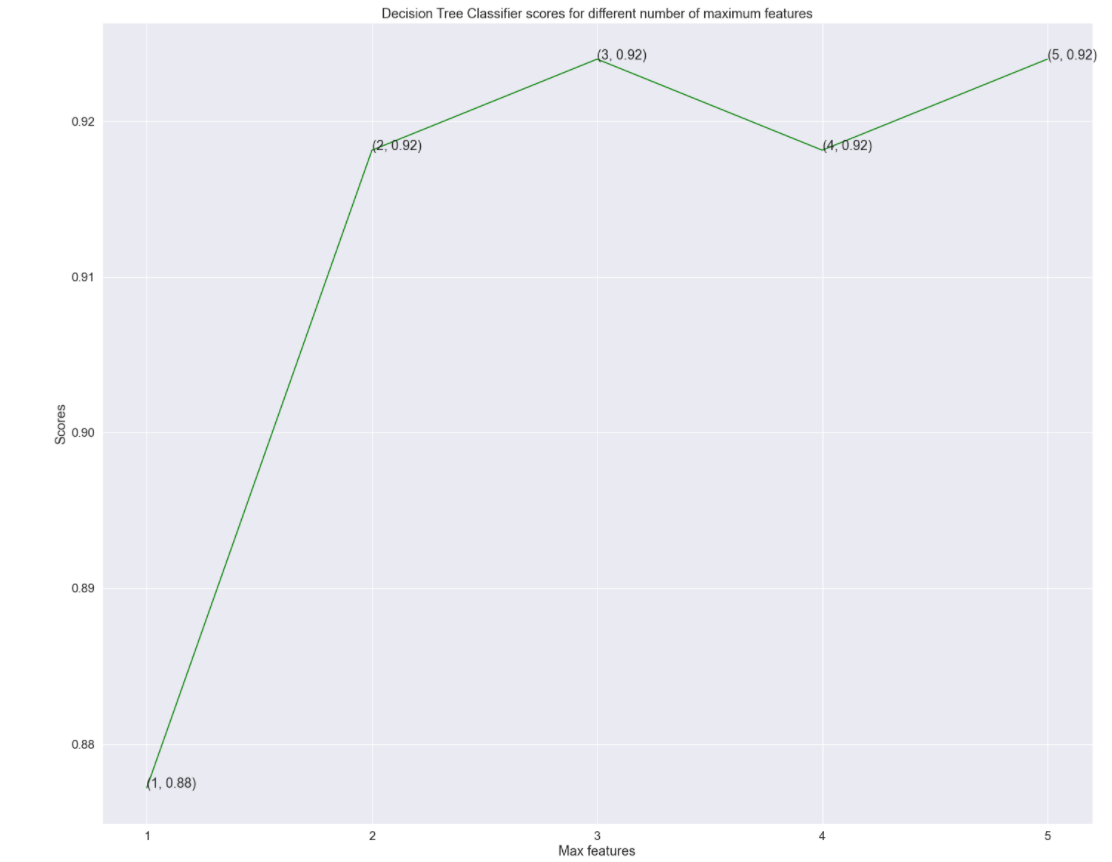
1. **K-Nearest Neighbor Classifier**

**Chart, line chart

Description automatically generated**

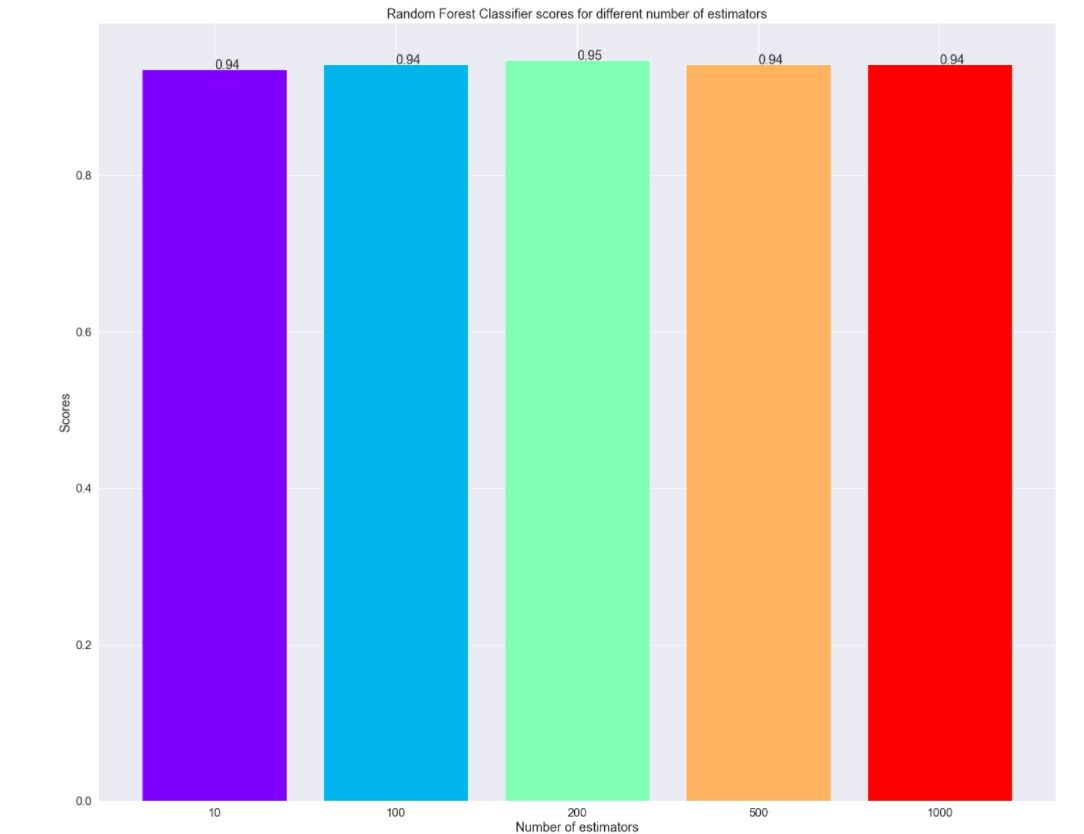
****

1. **Decision Tree Classifier**

****

****

1. **Random Forest Classifier**

****

****

**CONCLUSION**

I used machine learning to predict the breast cancer in this project. After using the dataset and dividing them inro two parts such that I took half of the data for training my model and half of the data for testing my model, I used three machine learning algorithms to find the more accurate techniques to predict the breast cancer. And I concluded that the accuracy of k Neighbors classifier is 88.3% and decision tree classifier is 91.81%; however, the random classifier achieved the highest score of 94.15%.

**FUTURE WORK**

For now, I used the dataset where I was predicting the breast cancer based on the mean radius, mean texture, mean perimeter, mean area, and mean smoothness of the breast lumps on the breast. But there might also be a case where the person can get breast cancer without the appearing of lumps in the breast, so in that case my algorithm might not work because I have not taken that case into consideration; therefore, in the future I will be working on it. Moreover, talking about the project that I have created now, I might use it in the real world by getting in contact with some health professional and showing my model to them to use it in the real life. We can create some web app or mobile app as well for its better use.

**REFERENCES**

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020, September 1). *Breast cancer prediction: A comparative study using machine learning techniques*. SN Computer Science. Retrieved November 24, 2021, from https://link.springer.com/article/10.1007/s42979-020-00305-w#ref-CR2.

*Breast cancer - statistics*. Cancer.Net. (2021, February 24). Retrieved November 24, 2021, from https://www.cancer.net/cancer-types/breast-cancer/statistics.

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. L. H., Abdelouhahid, R. A., & Debauche, O. (2021, September 8). *Machine learning algorithms for breast cancer prediction and diagnosis*. Procedia Computer Science. Retrieved November 24, 2021, from https://www.sciencedirect.com/science/article/pii/S1877050921014629.

Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021, April 16). *Predicting breast cancer 5-year survival using Machine Learning: A Systematic Review*. PloS one. Retrieved November 24, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8051758/.

*Breast cancer statistics: Facts & figures: NBCC*. National Breast Cancer Coalition. (2021, March 12). Retrieved November 24, 2021, from https://www.stopbreastcancer.org/information-center/facts-figures/.

*Using machine learning algorithms for breast cancer risk prediction and diagnosis*. IEEE Xplore. (n.d.). Retrieved November 24, 2021, from https://ieeexplore.ieee.org/document/8739696.

Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016, January 1). *Machine learning models in breast cancer survival prediction*. Technology and Health Care. Retrieved November 24, 2021, from https://content.iospress.com/articles/technology-and-health-care/thc1071.

*Breast cancer mortality rates: Recent figures and Trends - Moose and doc*. Breast Cancer - Moose and Doc. (2019, December 7). Retrieved November 24, 2021, from https://breast-cancer.ca/diag-chnces/.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010, July 16). *Missing data imputation using statistical and machine learning methods in a real breast cancer problem*. Artificial Intelligence in Medicine. Retrieved November 24, 2021, from https://www.sciencedirect.com/science/article/pii/S0933365710000679.