AN APPROACH USING MACHINE LEARNING MODEL FOR BREAST CANCER PREDICTION

Fatema Nafa, Enoc Gonzalez and Gurpreet Kaur

Department of Computer Science, Salem State University, Salem, MA, USA

ABSTRACT

Breast cancer is one of the most common diseases that causes the death of several women around the world. So, early detection is required to help decrease breast cancer mortality rates and save the lives of cancer patients. Hence early detection is a significant process to have a healthy lifestyle. Machine learning provides the greatest support to detect breast cancer in the early stage, since it cannot be cured and brings great complications to our health system. In this paper, novel models are generated for prediction of breast cancer using Gaussian Naive Bayes (GNB), Neighbour's Classifier, Support Vector Classifier (SVC) and Decision Tree Classifier (CART). This paper presents a comparative machine learning study based to detect breast cancer by employing four different Machine Learning models. In this paper, experiment analysis carried out on a Wisconsin Breast Cancer dataset to evaluate the performance for the models. The computation of the model is simple; hence enabling an efficient process for prediction. The best overall accuracy for breast cancer detection is achieved equal to 94%. using Gaussian Naive Bayes.

KEYWORDS

Machine Learning, Breast Cancer, Representation Learning, Gene Embeddings.

1. INTRODUCTION

Cancer that initiates in the cells of the breasts is called breast cancer, and it occurs more in women and rarely in men. Statistics indicate that breast cancer-related complications are the top causes of death among women, and the significant cases of breast cancer are attributed to a shortage of information. In America, breast cancer is the leading cause of cancer-related deaths, and the mortality rate is relatively high compared to the neighbouring countries in America.

According to [1], about 1 in 8 women in the United States is predicted to develop breast cancer during her lifetime. Statistically, this number is very high, and it shows the importance of studying and analysing the factors associated with breast cancer. Understanding breast cancer and what causes it has helped increase the survival rates, with the deaths associated with it declining.

According to [2], In the United States, breast cancer has become the second leading cause of cancer death in women, only second to lung cancer. The older the person, the higher the risk of getting breast cancer, with women being more likely to develop breast cancer than men. Factors such as family genetics also play a big part in breast cancer. People with close relatives older the person, the higher the risk for breast cancers who have been diagnosed with breast cancer are more likely to develop the disease at some point in their life.

David C. Wyld et al. (Eds): AI, AIMLNET, BIOS, BINLP, CSTY, MaVaS, SIGI - 2022
pp. 155-161, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121815

The symptoms and warnings of breast cancer people can experience vary from everyone. What this means for people is that in some cases an individual might not experience what is considered "typical" symptoms, and this makes breast cancer screenings the most important. According to [3], 99% of breast cancer occurs in women, with the lifetime of breast cancer risk for men being 1 in 1000. This means that while women are more likely to get breast cancer, men are also victims of this.

Although science has done a great job at saving so many lives from breast cancer, there's still much more to do, and detecting breast cancer in very early stages is crucial in order to increase the chances of survival of both women and men affected by the disease.

2. RELATED WORK

There has been a growing demand for machine learning models in the biomedical domain, each model aims at addressing challenges associated with a particular set of factors. The models included but not limited to, clustering, classification, neural networks, and associated rule mining [4]– [7] Many of them show good accuracy for the result. There are two kinds of machine learning models: predictive models and descriptive models [8]. A predictive model is to predict unknown variables of interest and it is applied to supervised learning. On the other hand, descriptive models are used to discover patterns in data, and it is applied to unsupervised learning [9]. In this work supervised learning models have been used for prediction.

Many scientists designate themselves to develop appropriate approaches for the detection of breast cancer. However, different machine learning algorithms have been used with different breast cancer dataset and the result varies based on the algorithm and the dataset used by different researchers. The research associated with the prediction task is outlined in brief as follows. Authors in [10], authors used several machine learning models to predict breast cancer and found that logistic regression performed best. Work by [11] proposed using three machine learning algorithms, Naïve Bayes, random forest and K-Nearest Neighbour for breast cancer prediction. In the result, K-Nearest Neighbour (KNN) has better performance.

Another study [12] tested different machine learning models such as SVM, KNN, DT, Logistic Regression, and Random Forest for breast cancer prediction. The highest effectiveness was determined to be 89 percent for random forest. Nikita Rane et al [13] tested six different machine learning algorithms for breast cancer prediction. Naive Bayes, Random Forest, Artificial Neural Network, Nearest Neighbour, Support Vector Machine, and Decision Tree. They classified the cancers as benign and malignant.

According to [14] authors surveyed machine learning techniques based on the Scopus database. Examples of the techniques are Support Vector Machine, Logistic Regression, and Decision Tree model, it used for breast cancer classification task. Also, ML techniques widely used for developing CAD systems are Decision Tree (DT), Naive Bayes, nearest neighbour, Artificial Neural Network (ANN), Support Vector Machines, and set Classifiers have been used for breast cancer classification.

Using the background and related work, a missing element to existing research was to assess the sensitivity, specificity, and accuracy of the performance and quality of the proposed models. This work presents a comparative study of the efficiency for four classifiers: SVM, Naive Bayes, Neighbour's, and Decision Tree, which are the most popular machine learning techniques. Also, the optimization of these models has been investigated for breast cancer detection.

3. PROBLEM DEFINITION

The dataset provides the attribute of women and their incidence of breast cancer (diagnosis). Performing exploratory data analysis on the dataset and come up with insights on the factors that cause breast cancer. Also, in this paper we'll analyze the data and evaluate different Machine Learning models to predict whether a specific set of symptoms will be high risk breast cancer or not which is predicting breast cancer at early stage.

4. METHODS

This section will introduce the proposed models to solve the problem. Overview of the system presented in Figure 1.1. The first step that needed to be completed is the preprocessing of the dataset. Different python libraries have been used in order to handle different issues. In this step, exploratory data analysis level has been applied to clean and prepare the data. A dummy variable (diagnosis), handle the missing values, visualize the data, and create heatmaps and a scatter matrix. In additionally, a Multicollinearity has been checked because of one of the assumptions of the proposed model is that there isn't any Perfect multicollinearity. Multicollinearity is where one of the variables is highly correlated with another explanatory variable. To check Multicollinearity, a correlation matrix using the corr() python function used. The function creates a matrix with each variable having its correlation calculated for all the other variables. The visualization for the matrix created using heatmaps. if you travel diagonally down the matrix all the associations should be one, as it is calculating the correlation of the variable with itself. heatmaps can quickly help identify the highly correlated variables, by just looking for the darker colors. So, the correlation between diagnosis other variables was found. More details can be found in the result section.

Second step, another level of in-depth exploration of the data before building the models, in this step the data get explored to see how the data is distributed and if there are any outliers. Third step, build the model, Random Forest Regression model is used. Then 4 machine learning models imported, the best accuracy gained by this model. last step, evaluating the performance of the model using Confusion Matrix. and explored the data we can proceed to the next part, building the model.in additionally, the interpretation of the prediction result needs a discussion to see if the result make sense.

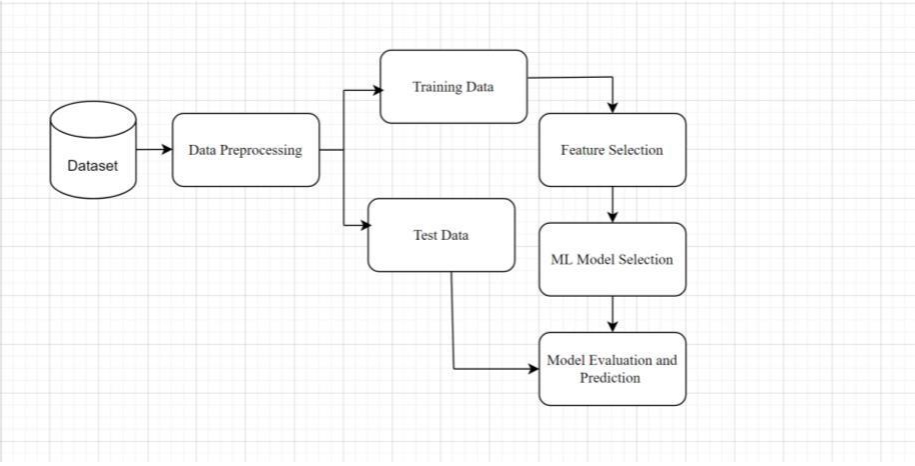

Figure 1. Overview of the proposed Model.

4.1. Model Inputs

Supervised Learning algorithms, it is a machine learning algorithm that learn to anticipate new results based on previous learning by learning patterns from pre-existing data. Existing data is identified using machine learning methods such as probability-based, function-based, rule-based, tree-based, instance-based, and so on.

Naïve bayes is a machine learning model based on based on Bayes' Theorem with an assumption of independence among predictors. In other words, the attributes are independent of each other. Naïve Bayes predicts datasets with the assumption that attributes belonging to a class that is independent of each other. This study uses Gaussian Naïve Bayes algorithm which works well with both continuous and discrete datasets. And Bayes refers to the statistician and philosopher Thomas Bayes theorem [15]. The NB theorem can be expressed mathematically as follows:

$$P(A/B) = (P(B/A)\ P(A))/(P(B)) \quad (1)$$

P (A | B): Probability of occurrence of event A given the event B is true.
P (A) and P (B): Probabilities of the occurrence of event A and B respectively.
P (B | A): Probability of the occurrence of event B given the event A is true.

Naïve bayes has been used in different application such as Real time Prediction, Text classification, Spam Filtering and Sentiment Analysis [6], [16], [17].

5. RESULTS

5.1. Dataset

An experiment was conducted on dataset. The dataset is Wisconsin Breast Cancer dataset from UCI Machine

Learning Repository. There are around 30 numeric attributes of features in the dataset. The total size was 156.7 KB. Table 1.1 provides a descriptive statistic to describe and summarize the data. It uses quantitative approach describes and summarizes data numerically. We can observe that the data set contain 569 rows and 31 columns. The missing values are none.

Table 1. Descriptive Statistics about the Dataset.

Dataset statistics	
Number of variables	31
Number of observations	569
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	165.7 KiB
Average record size in memory	298.2 B

Variable types	
Categorical	1
Numeric	30

we should analyze and "get to know" the dataset. in other words, get familiarize with the dataset, to gain some understanding of the potential features and to see if data cleaning is needed. After the statical information has been presented. The pre-processing step needed to be applied. In this dataset, reordering for some columns has been performed, delete ID column. The diagnosis

feature represents the number of Benign and Malignant cases. It has been replaced with Benign = 0 and ‘Malignant = 1 to make all the data numerical type.

Using descriptive method that generates descriptive statistics that summarize the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN values. This method tells us a lot of things about a dataset. Table1.2 shows the statistics that are generated by the describe () method: count tells us the number of non-empty rows in a feature. mean tells us the mean value of that feature. std tells us the Standard Deviation Value of that feature. min tells us the minimum value of that feature. 25%, 50%, and 75% are the percentile/quartile of each feature. This quartile information Max tells the maximum value of that feature Correlation is a statistical technique that can show whether and how strongly pairs of variables are related/interdependent [18], [19]. Figure 3. show Pearson correlation coefficient. Pearson correlation coefficient is a measure of the strength linear association between two variables. Looking at the heatmap along with the correlation matrix we can identify a few highly correlated variables as in Figure 3. This is an extremely high correlation and marks it as a candidate to be removed. Logically it makes sense that these two are highly correlated.

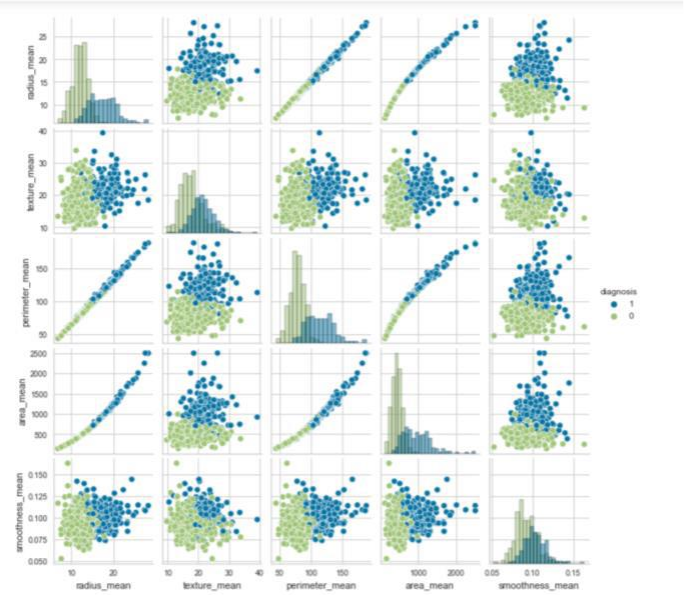

Figure 2. Correlation between Variables.

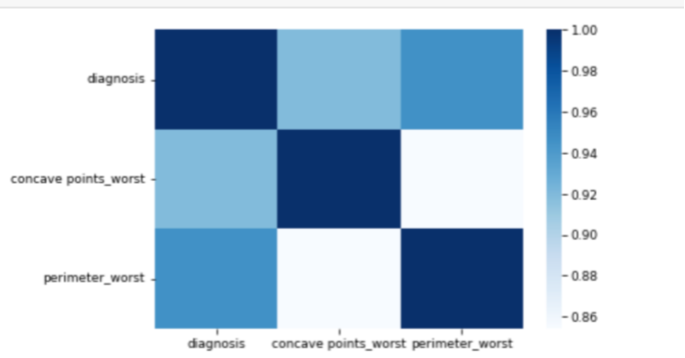

Figure 3. Correlation between concave points worst, perimeter worst and diagnosis.

5.2. Performance Metrics/Confusion Matrix

The performance metrics of the classification model were calculated based on precision, recall, and accuracy and are presented in Table 2. TP and TN specify the numbers of diabetes and normal patients that were correctly classified, respectively, while FN and FP specify the numbers of normal and diabetes patients that were incorrectly classified, respectively. 10-fold cross-validation was used to train and test the dataset for the entire classification model.

6. CONCLUSION AND DISCUSSION

As a result, Gaussian Naive Bayes (NB) has been used effectively to predict whether a patient will have breastcancer based on some features. The model has achieved an accuracy of 93% as shown in Table 2. The most important features are the one with high values. Based on the result, it looks that variable6 level has the most significant influence on the model. the second highest value is variable9 and the last one is variable8. The three features have a positive influence on the prediction their higher values are correlated with person being diabetes. I noticed a lot of information needed to be discussed with medical expertise to make sure that it is correct.

Table 2. The Accuracy for the Prediction model.

Algorithm	Recall	precision	F1-Score	Accuracy	Run Time
DecisionTreeClassifier	90.00%	99.00%	97.00%	92.00%	0.059s
Support Vector Machine	87.00%	99.00%	90.00%	92.00%	0.056s
Gaussian Naive Bayes (NB).	**93%**	**100%**	**90%**	**94.00%**	**0.016s**
KNeighborsClassifier	91.00%	91.00%	92.00%	91.00%	0.028s

REFERENCES

[1] "Breast Cancer Statistics | How Common Is Breast Cancer?" https://www.cancer.org/cancer/breast-cancer/about/how-common-isbreast-cancer.html (accessed May 13, 2022).

[2] J. A. Ajani et al., "Gastric cancer, version 2.2022, NCCN clinical practice guidelines in oncology," J. Natl. Compr. Canc. Netw., vol. 20, no. 2, pp. 167–192, 2022.

[3] P. A. McElfish et al., "Diabetes and hypertension in Marshallese adults: results from faith-based health screenings," J. Racial Ethn. Health Disparities, vol. 4, no. 6, pp. 1042–1050, 2017.

[4] R. Khan, Y. Qian, and S. Naeem, "Extractive based Text Summarization Using K-Means and TF-IDF.," Int. J. Inf. Eng. Electron. Bus., vol. 11, no. 3, 2019.

[5] M. A. Ibrahim, M. U. G. Khan, F. Mehmood, M. N. Asim, and W. Mahmood, "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification," J. Biomed. Inform., vol. 116, p. 103699, 2021.

[6] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," Expert Syst. Appl., vol. 165, p. 113679, 2021.

[7] N. Alami, M. Meknassi, N. En-nahnahi, Y. El Adlouni, and O. Ammor, "Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling," Expert Syst. Appl., vol. 172, p. 114652, 2021.

[8] M. Kantardzic, Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[9] B. Lantz, Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd, 2019.

[10] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," J. Robot. Control JRC, vol. 2, no. 3, pp. 115–118, 2021.

[11] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," J. Robot. Control JRC, vol. 2, no. 3, pp. 115–118, 2021.

[12] R. Rawal, “Breast cancer prediction using machine learning,” J. Emerg. Technol. Innov. Res. JETIR, vol. 13, no. 24, p. 7, 2020.
[13] R. Hazra, M. Banerjee, and L. Badia, “Machine learning for breast cancer classification with ann and decision tree,” in 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2020, pp. 0522–0527.
[14] E. H. Houssein, M. M. Emam, A. A. Ali, and P. N. Suganthan, “Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review,” Expert Syst. Appl., vol. 167, p. 114161, 2021.
[15] T. Bayes and D. Hume, “BAYES’S THEOREM,” in Proceedings ofthe British Academy, 1763, vol. 113, pp. 91–109.
[16] E. Ezpeleta, U. Zurutuza, and J. M. Gómez Hidalgo, “Does sentiment analysis help in bayesian spam filtering?,” in International Conference on Hybrid Artificial Intelligence Systems, 2016, pp. 79–90.
[17] R. Mallik and A. K. Sahoo, “A novel approach to spam filtering using semantic based naive bayesian classifier in text analytics,” in Emerging technologies in data mining and information security, Springer, 2019, pp. 301–309.
[18] M. Ezekiel and K. A. Fox, “Methods of correlation and regression analysis: linear and curvilinear,” 1959. [19] S. L. Crawford, “Correlation and regression,” Circulation, vol. 114, no. 19, pp. 2083–2088, 2006.