International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Breast Cancer Prediction using Machine Learning Algorithms

Madasu Mallika[1], Dr. K. Suresh Babu[2]

[1]M. Tech (Data Science), Student Department of Information Technology, JNTUHUCESTH, Hyderabad, Telangana - 500085, India
Email: *madasumallika528[at]gmail.com*

[2]Professor of CSE, Department of Information Technology, JNTUHUCESTH, Hyderabad, Telangana - 500085, India
Email: *kare_suresh[at]jntuh. ac. in*

Abstract: ***The project is titled "Breast Cancer Prediction Using Machine Learning Algorithms." Breast cancer affects a significant number of women world-wide and ranks as the second most common cause of death among women. Early detection of breast cancer can drastically improve the prognosis and chances of survival by enabling timely clinical therapy. Furthermore, precise benign tumour classification can help patients avoid unnecessary treatment. The dataset for this study includes several clinical features such as insulin, glucose, resistin, adiponectin, homeostasis model assessment (HOMA), leptin, and monocyte chemoattractant protein-1 (MCP-1), along with age and body mass index (BMI). In this study, we will apply three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) to the Coimbra Breast Cancer Dataset (CBCD). After obtaining the results, a performance evaluation and comparison will be conducted among these different classifiers. This study aims to utilize machine learning algorithms for breast cancer prediction with a focus on identifying the most efficient classifiers through a comprehensive analysis of the confusion matrix, accuracy, and precision.***

Keywords: Machine Learning, Support vector Machine, Random Forest, Logistic Regression

1. Introduction

Among the various cancers that occur in women, breast cancer is found to be the second highest cancer after lung cancer, which is caused for different reasons: age, heredity, life style factors etc. Every year, 9.6 million people are killed by cancer. Conferring to the American Cancer Society, the ladies are affected by breast cancer in comparison to all other cancers already introduced. According to estimates, about 252, 710 women will be diagnosed with invasive breast cancer and approximately 63, 410 women will be diagnosed with in situbreast cancer in the United States in 2017.

Men are also more likely to develop breast cancer. In 2017, it is estimated that roughly 2470 males will be diagnosed with this malignancy in the United States. According to another estimate, around 41, 070 people would die from this disease in 2017. According to recent UK data, 41, 000women are diagnosed with breast cancer each year, whereas just 300 males are diagnosed. Any progress in the prediction and identification of cancer illnesses is vital to living a healthy life.

As a result, high cancer prognosis accuracy is crucial for updating therapeutic elements and patient survival standards. Machine Learning techniques can make a large contribution to the process of prediction and early diagnosis of breast cancer became a research hotspot and has been proved as a strong technique.

In this study, we applied three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR) on the Coimbra Breast Cancer Dataset (CBCD). Following the collection of findings, a performance evaluation and comparison of these various classifiers is performed and deciding on the best classifier.

2. Literature Review

Breast cancer affects the majority of women worldwide, and it is the second most common cause of death among women. Notwithstanding, assuming disease is recognized early and treated appropriately, it is feasible to be restored of the condition. Early recognition of breast cancer can emphatically work on the visualization and chances of endurance by permitting patients to get opportune clinical treatment. Moreover, exact harmless growth characterization can assist patients with staying away from unneeded treatment.

This paper concentrate on involves Convolution Neural Networks for Image dataset and K-Nearest Neighbour (KNN), Decision Tree (CART), Support Vector Machine (SVM), and Naïve Bayes for numerical dataset, whose highlights are acquired from digitized picture of breast mass, as to forecast and break down disease data sets to further develop exactness.

The dataset will be investigated, assessed, and model is prepared as a feature of the interaction. At last, both picture and mathematical test information will be utilized for expectation.

3. Methodology

3.1 Data Source

The dataset used here for predicting breast cancer is taken from kaggle. It is used for implementing machine learning algorithms. The dataset consists of 116 instances of data with the appropriate 10 clinical parameters. A number of clinical features in the dataset of this study including insulin, glucose, resistin, adiponectin, homeostasis model assessment

(HOMA), leptin, monocyte chemoattractant protein-1 (MCP-1), along with their age and body mass index (BMI).

3.2 Data Cleaning

Taking care of missing qualities and exceptions is a basic piece of information pre-processing to guarantee that your machine learning models perform precisely and powerfully. Underneath, I'll frame normal ways to deal with address missing qualities and anomalies in your dataset:

3.2.1 Handling Missing Values

a) Identify Missing Values:
Use Pandas to distinguish missing qualities in your dataset. You can utilize the isnull () strategy to check for missing qualities.

b) Information Attribution:
For missing mathematical qualities, consider attributing them with the mean, middle, or method of the separate section utilizing the fillna () strategy. For missing straight out values, you can supplant them with the most incessant classification (mode) or utilize other proper techniques like forward-fill or in reverse fill.

3.2.2 Handling Outliers

a) Identify Outliers:
Imagine your information utilizing box plots, histograms, or disperse plots to recognize likely exceptions. You can likewise utilize factual strategies like the Z-score or the IQR (Interquartile Reach) to recognize anomalies.

b) Outlier Treatment:
For gentle exceptions, you might decide to keep them in the dataset, particularly assuming that they address legitimate data of interest.

For extreme exceptions, you can consider about one of the accompanying choices:

- Shorten or cover outrageous qualities to a specific limit. Change the information utilizing procedures like log change to lessen the effect of exceptions.
- Eliminate anomalies on the off chance that they are intriguing or on the other hand assuming they essentially influence the model's presentation.

3.3 Algorithm Description

This part portrays around three algorithms used in this system namely Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR).

3.3.1 Support Vector Machine

Support Vector Machine is usually represented as SVM. It is an elegant and Powerful Algorithm. The support vector machine procedure seeks a hyperplane in an N-dimensional space (N — the number of features) that clearly classifies the input points. There are a few hyperplanes that may be utilized to divide the two gatherings of data of interest. We want to find a plane with the best edge, or the best distance between data of interest from the two classes. Boosting the edge distance gives some support; permitting resulting information focuses to be classified with more conviction.

Hyperplanes and Support Vector:
Hyperplanes are choice limits that assist with ordering the important pieces of information. Data points on either side of the hyperplane might belong to distinct classes. Furthermore, the size of the hyperplane is determined by the number of features. When the number of input features is two, the hyperplane is just a line. When the number of input characteristics reaches three, the hyperplane transforms into a two-dimensional plane. When the quantity of characteristics crosses a certain threshold, it becomes impossible to conceive. We want to optimize the margin between the data of interest and the hyperplane in the SVM algorithm.

3.3.2 Random Forest

Irregular timberlands otherwise called arbitrary choice backwoods makes an enormous number of trees that accomplish their result through troupe learning strategies for characterization, relapse. It is the Pack and element haphazardness highlights it utilizations to develop those trees. The arbitrary backwoods enjoy an upper hand over the choice tree which, is that it doesn't over fit the information. The workflow of arbitrary backwoods is given below.

a) From the preparation set, picked K information focuses randomly.
b) From these K pieces of information, produce the choice trees.
c) From produced trees, pick the quantity of N-tree furthermore; rehash steps (i) and (ii).
d) Structure the N-tree that predicts the classification to which the information focuses relate for another data of interest, and dole out the new information point through the classification with the most noteworthy probability.

3.3.3 Logistic Regression

Strategic relapse was presented by analyst DR Cox in 1958 thus originates before the field of machine learning. Strategic relapse is utilized for tackling the order issues. It is a directed Machine learning strategy, utilized in grouping position (for expectations based on preparing data).

Calculated Relapse utilizes a condition like Straight Relapse; however, consequence of strategic relapse is an out and out factor however it is a motivation for other relapse models. Double results can be anticipated from the autonomous factors. It will in general be either Yes or No, 0 or 1, substantial or Sham, etc. in any case, as opposed to giving the particular worth as 0 what's more, 1, it gives the probabilistic characteristics which lie some place in the reach of 0 and 1.

The overall work process is:
(1) Get a dataset
(2) Train a classifier
(3) Make a forecast utilizing such classifier

3.4 Work Flow

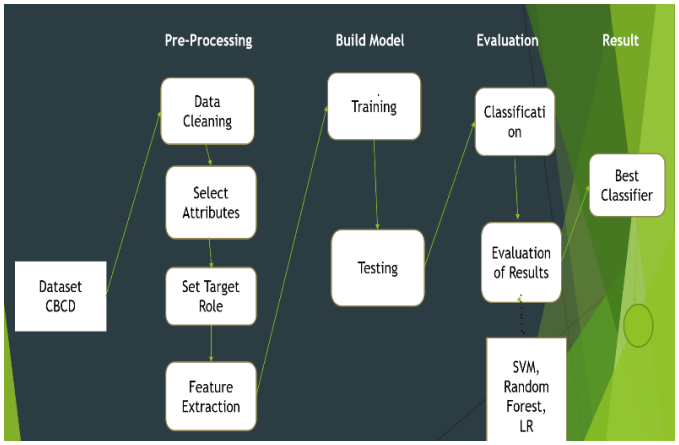

Figure: Work Flow Methodology

4. Model Training & Evaluation

4.1 Split the dataset into training and testing sets (70% training, 30% testing)

To split a dataset into training and testing sets in Python, you can use the train test split function from the scikit-learn library.

4.2 Feature Scaling

Standardize or normalize the features. In machine learning, feature scaling is a crucial pre-processing step that makes sure that features have a constant scale. Standardization (z-score scaling) and normalization (min-max scaling) are two frequently used techniques for feature scaling. You can decide which one best fits your data and your machine learning algorithms' needs. I'll give samples of both techniques here:

i) Normalization (Z-score Scaling)
Features that are scaled according to standardization have a mean of 0 and a standard deviation of 1. When your data contains outliers or your machine learning technique (like Support Vector Machines or k-Nearest Neighbors) expects that features are normally distributed, this is a suitable option.

ii) Normalization (Min-Max Scaling)
Features are scaled to a particular range during normalization, often between 0 and 1. It is a wise decision if your machine learning system such as neural networks, requires inputs that fall within a certain range or if your features have variable unit sizes.

To maintain consistency, apply the same scaling to both the training and testing sets of data after fitting the scaler on the training data to learn the scaling parameters.

Select the scaling technique that best satisfies the needs of your machine learning models and the features of your data. Although standardization is generally a safe option, it might not be appropriate for all datasets.

4.3 Model Selection

Step 1: Train a Logistic Regression model.
Step 2: Train a SVM model.
Step 3: Train a Random Forest model.

4.4Hyperparameter Tuning (Grid Search or other methods)

To obtain the ideal set of hyperparameters for a machine learning model, one method for hyperparameter tweaking is called grid search. Hyperparameters are variables that are chosen in advance of training rather than ones that are learned from the data. Examples include a random forest's number of trees, regularization intensity, and learning rate.
Grid Search is useful since it streamlines the procedure for locating the best hyperparameters. When compared to manual tuning, it methodically explores the hyperparameter space, saving you time and effort. Finding the hyperparameters that result in the greatest model performance on your validation data is one of the benefits of this tool.

4.5 Model Testing

To evaluate the performance of each model, you can calculate various evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

Confusion Matrix:
The effectiveness of a classification model can be assessed using a confusion matrix. It gives a concise description of the predictions made by the model and the actual results. Four elements make up a classic confusion matrix for binary classification:

1) True Positive (TP): The prediction of the positive class by the model was accurate.
2) True Negative (TN): The negative class was successfully predicted by the model.
3) False Positive (FP): A Type I error in which the model predicted the positive class when it should have been projected as the negative class.
4) False Negative (FN): When the model should have predicted the positive class, it instead predicted the negative class (Type II mistake).

5. Conclusion

Early finding of breast cancer can significantly impact its progression and reduce mortality rates by enabling timely therapeutic interventions. Therefore, the primary objective is to conclude breast cancer using Machine Learning algorithms and identify the most effective ones based on factors such as the confusion matrix, accuracy, and precision.

References

[1] https://www.atlantis-press.com/article/125960864.pdf
[2] https://www.researchgate.net/profile/DrKumar11/publication340621Cancer-Prediction-Using-Machine-Learning-Algorithms.pdf

[3] https://research.ijcaonline.org/volume62/number1/pxc3884635.pdf
[4] Dr. B. Santhosh Kumar, T. Daniya, Dr. J. Ajayan, “Breast cancer prediction using machine learning algorithms”, March2020, https://www.researchgate.net/publication/340621198
[5] Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader benyettou,” Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules”, International Journal of Computer Applications (0975 - 8887) Volume 62 - No. 1, January 2013
[6] H. Song, H. Watanabe, X. Xiao and T. Kikkawa, "Influence of Air-gaps between Antennas and Breast on Impulse-Radar Based Breast Cancer Detection," 2019 13th European Conference on Antennas and Propagation (EuCAP), Krakow, Poland, 2019, pp. 1-2