BREAST CANCER PREDICTION USING MACHINE LEARNING

Murikipudi Nikhil (190031117)
Name of the Supervisor: Dr P Lakshmi Prasanna,
Associate Professor ,Department of CSE

Introduction

Objective of your work

The objective of this project is to develop a highly accurate and reliable model that can predict whether a breast mass is malignant or benign based on various features extracted from mammography images and patient information. The model will be trained on a large dataset of historical breast cancer cases and will utilize machine learning algorithms to identify the patterns that distinguish between malignant and benign cases. The ultimate goal of this project is to improve early detection and diagnosis of breast cancer, which can lead to better patient outcomes and potentially save lives.

Origin of your proposal

This project originated from the need to improve the accuracy and reliability of breast cancer prediction, and to provide clinicians with a more powerful tool for diagnosis and treatment planning. The project is driven by the desire to leverage the power of machine learning algorithms to identify patterns and trends in large, complex datasets, and to use this information to develop highly accurate predictive models.

Methods

This Project can be done in total of 5 ways which are given below. After Testing each method, we came to know which method gives most accurate outcome.

Logistic Regression:

Logistic regression is a technique that firstly used for biological studies in the early twentieth century. It has become widespread for social studies too. Logistic regression is also one of the predictive analyses.

K-Nearest Neighbour (KNN):

KNN is a supervised learning technique that means the label of the data is identified before making predictions. Clustering and regression are two purposes to use it. K represents a numerical value for the nearest neighbours. KNN algorithm does not have a training phase.

Decision Tree:

A decision tree (DT) is one of the most common supervised learning techniques. Regression and classification are two main goals to use it. It seeks to solve problems by drawing a tree figure. Features are known as decision nodes, and outputs are leaf nodes.

Block Diagram, Flowchart, Models, Results

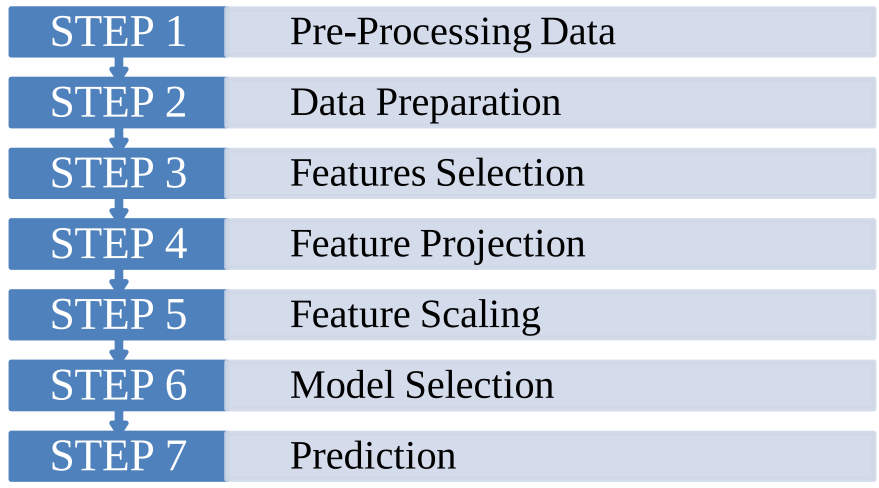
Random Forest:

Random forest is an ensemble learning model that can be used for regression and classification.

Support Vector Machine:

The objective of this algorithm is to find a hyperplane in N-dimensions that classifies the data points.
N dimension diversifies based on the feature numbers.

Figure 1

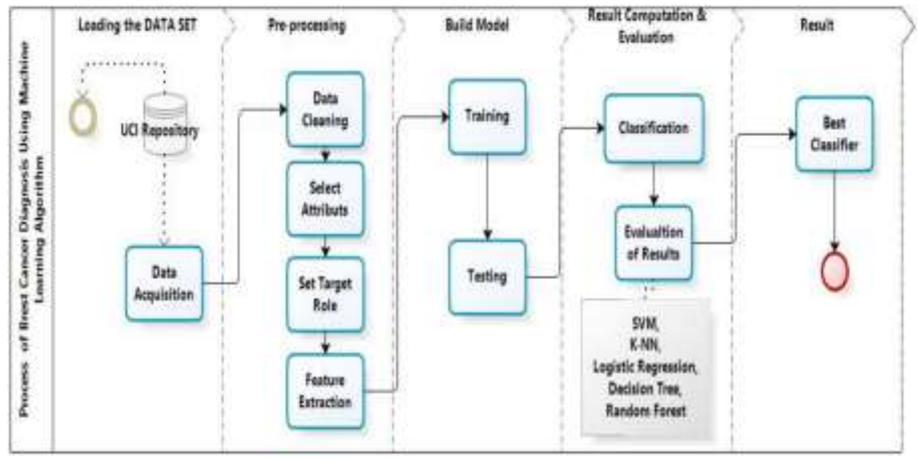


STEP 1 Pre-Processing Data
STEP 2 Data Preparation
STEP 3 Features Selection
STEP 4 Feature Projection
STEP 5 Feature Scaling
STEP 6 Model Selection
STEP 7 Prediction

Process flow involving project completion

This project can be done in 7 steps. After completing first five phases, near 6th phase we use 5 different algorithms and calculate accuracy to conclude phase 7 which is finding best algorithm for predicting breast cancer

Figure 2

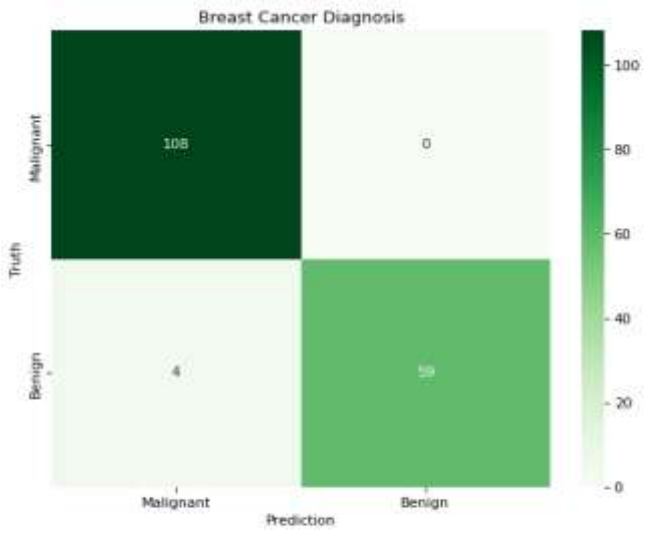


Flow chart indicates complete process of this project

The above flow diagram indicates flow of this project. In this project, logistic regression, k-nearest neighbour, support vector machine, random forest, decision tree, and naïve Bayes classification algorithms were created, and accuracy scores for each of them were obtained. Each algorithm was applied to three different datasets that included various features.

In this project, five distinct machine learning techniques were investigated for breast cancer diagnosis. SVM gave better accuracy results rather than the other methods. The main advantage of this project is that it is very efficient to train.

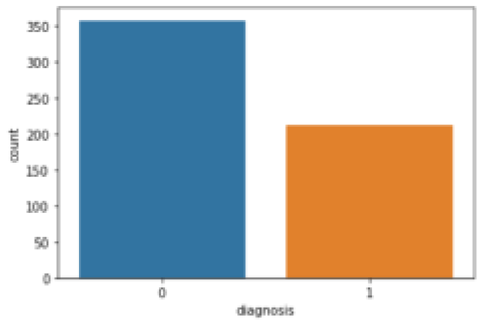
Figure 1

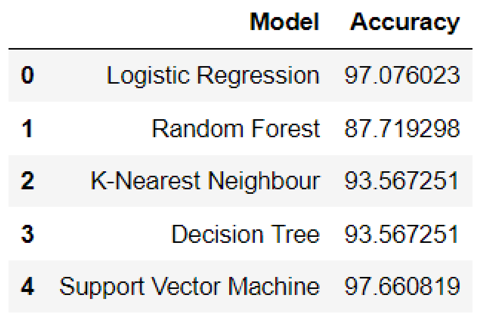


Support vector machine confusion matrix

Each algorithm was applied to three different datasets that included various features. The first dataset covered all independent features, the second dataset included highly correlated features, and the last dataset included low correlated features. Three different datasets were used separately for each machine learning technique, and accuracy results were obtained to make comparisons. SVM gave better accuracy results rather than the other methods.

Figure 2





	Model	Accuracy
0	Logistic Regression	97.076023
1	Random Forest	87.719298
2	K-Nearest Neighbour	93.567251
3	Decision Tree	93.567251
4	Support Vector Machine	97.660819

Bar graph and Accuracy table

In this project we have taken Wisconsin data set as our primary data. After processing data by clearing unwanted data, we have visualized the data in which we found that there are a total of 357 benign cases and 212 malignant cases, five distinct machine learning techniques were investigated for breast cancer diagnosis. SVM gave better accuracy results rather than the other methods.

Conclusion

Discussion/Conclusion

Rapidly spreading breast cancer has widely affected women's lives. Reliable and early detection leads to a reduction in the breast cancer death ratio.In this project in python, we learned to build a breast cancer tumour predictor on the Wisconsin dataset and created graphs and results for the same. As a first step, the dataset was prepared for visualization by removing non-numerical values and normalizing each numeric value. Then, heat map, boxplot, swarm plot, and scatterplots were created by using R studio, Minitab, and Python. Finally Support vector machine gave highest accuracy of all other algorithms which is 97.6%.

Limitations

One of the most significant limitations is the availability and quality of data. Inaccurate data can lead to biased predictions, which can impact the accuracy of the model. Furthermore, the model may face difficulty in predicting breast cancer accurately for women with non-traditional risk factors. Additionally, due to the complex nature of the disease.

Future Direction

As a future improvement, the system can add more features such as recommendation of medicines or treatments based on the severity of the patient. This prediction and recommendation system can help doctors to diagnose and cure the disease more efficiently.

References and Affiliations

References

[1] Wang H., Yoon W.S. Breast cancer prediction using data mining method; Proceedings of the 2015 Industrial and Systems Engineering Research Conference; Nashville, TN, USA. 30 May–2 June 2015.

[2] Nithya R., Santhi B. Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *Int. J. Comput. Appl.* 2011;28:0975–8887. doi: 10.5120/3391-4707.

Acknowledgements

We would like to express our gratitude to Dr Senthil Athithan, HOD for CSE department for providing us with adequate facilities and our instructor Dr. P. LAKSHMI PRASANNA for providing us with adequate planning and support and means by which we can complete this project.