ECE 579 Intelligent Systems, Winter 2023

Final Project Report

Project-Title: Breast Cancer Diagnosis Detection Model

Project Group Members

1. Chinmaya Karthik Botta
2. Monish Lanka
3. Shanmukh Varma Rudraraju

Responsibilities

1. Dataset identification and Initial Research: Shanmukh Varma Rudraraju, Monish Lanka, Chinmaya Karthik Botta
2. Exploratory data analysis: Chinmaya Karthik, Monish Lanka
3. Model Building and Model Evaluation: Shanmukh Varma Rudraraju
4. Literature Survey: Shanmukh Varma Rudraraju, Chinmaya Karthik Botta

Introduction

Breast cancer is one of the leading causes of death among women worldwide. Early detection is crucial for effective treatment and improved survival rates. In this project We are going to develop a classification model which would predict the occurrence of Malignant and Benign respectively and provide accurate diagnosis results. This report presents a detailed overview of our project, including background information, technologies used, methods employed, experiments conducted, and our findings.

Data Description

The "Breast Cancer Wisconsin (Diagnostic) Data Set" [1] on Kaggle is a dataset of patient characteristics and tumour features used to diagnose whether a breast mass is benign or malignant. The table below contains further information about the Dataset. The data was collected by Dr. William H. Wolberg at the University of Wisconsin Hospitals and includes 569 samples of biopsy tissue. Each sample includes 30 features such as radius, texture, and perimeter of the tumour, as well as patient information like age and menopause status. The data is divided into two classes, "Benign" and "Malignant", based on the biopsy diagnosis.

No of Features	Features Classification	No of Classes	Classes Classification	Total Data Points

30	Mean, Standard Error and Worst Mean	2	B = benign
			M = Malignant
				569

Description of technologies

To predict the diagnosis of breast cancer based on the features calculated from the FNA pictures, this dataset has been widely used in machine learning research. Several techniques, such as nearest neighbours, Linear SVM, RBG SVM, decision trees, random forest, neural networks, AdaBoost, Naïve Bayes, Logistic regression, and a hybrid model have been applied to create prediction models. Feature selection and feature extraction research has also used the dataset.

1. **Logistic Regression:** This is a widely used parametric classification model that models the probability of a binary outcome based on one or more predictor variables.
2. **Decision Trees**: This model uses a tree-like structure to partition the feature space into disjoint regions, with each leaf node representing a class label. [3]
3. **Ada Boost**: This is an ensemble model that combines multiple weak classifiers to create a strong classifier. It can handle complex relationships between variables and tends to be more robust to noise but can be sensitive to outliers.
4. **Naïve Bayes**: This is a probabilistic model that assumes independence between the features. It is simple, fast, and works well with high-dimensional data, but can be biased towards features with many levels and may not work well with correlated features.
5. **Random Forest**: This is an ensemble model that uses multiple decision trees and aggregates their predictions. It is robust to noise and can handle high-dimensional data but may not work well with highly correlated features.
6. **Support Vector Machine (SVM)**: This is a model that separates data points into different classes using a hyper plane that maximizes the margin between the classes. [2]
7. **K-Nearest Neighbours (KNN)**: This is a non-parametric model that classifies new data points based on the majority class of their k-nearest neighbours in the training data.
8. **Neural Network Model**: We have used Multilayer Perceptron. This is a feed forward neural network that consists of multiple layers of nodes, with each node in a layer connected to all nodes in the adjacent layers.

Recent Developments

Much progress has been made in the identification of breast cancer using machine learning algorithms in recent years. These advancements have resulted in more accurate and efficient breast cancer detection approaches, which can improve patient outcomes and survival rates.

Here are some recent developments and publications related to machine learning applied to the Breast Cancer Wisconsin (Diagnostic) dataset:

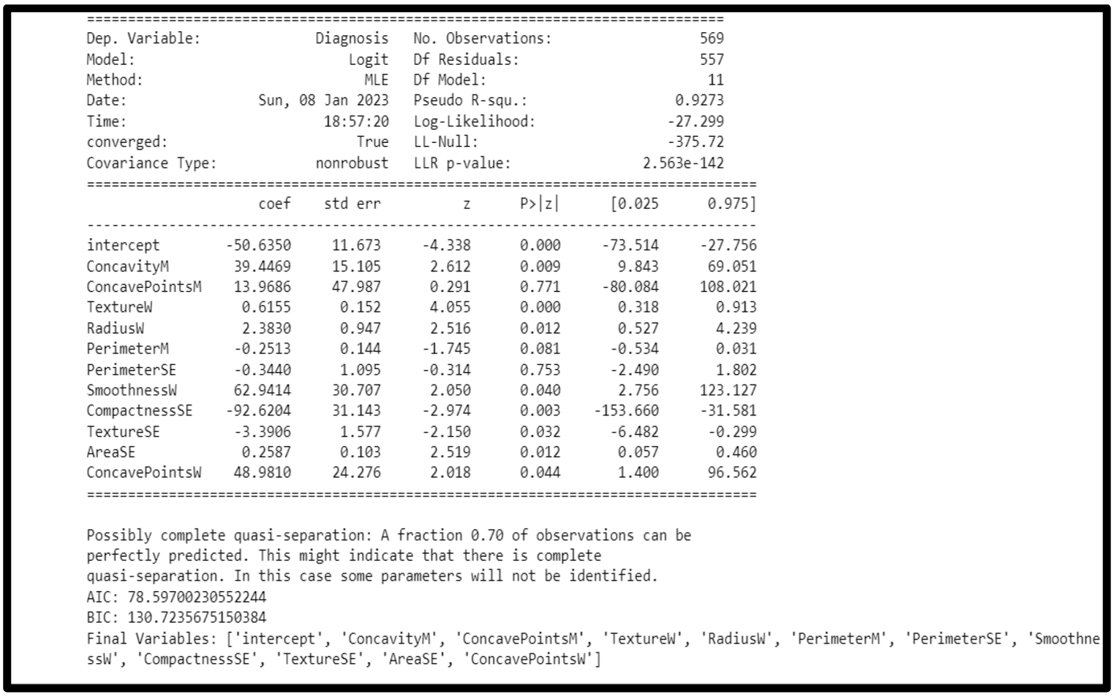
1. "Exploring the role of decision trees in breast cancer diagnosis and prognosis" In this study, the authors used decision trees to predict breast cancer diagnosis and prognosis using the Breast Cancer Wisconsin (Diagnostic) dataset. They compared different decision tree algorithms and found that the CART algorithm produced the best results as per reference [4]
2. "A Comparative Study of Machine Learning Techniques for Breast Cancer Classification”. In this study, the authors compared different machine learning algorithms for breast cancer classification using the Breast Cancer Wisconsin (Diagnostic) dataset. They found that the Random Forest algorithm achieved the highest accuracy as per reference [5]
3. "Breast Cancer Detection using Convolutional Neural Networks". In this study, the authors used convolutional neural networks (CNNs) to detect breast cancer using mammograms. They used the Breast Cancer Wisconsin (Diagnostic) dataset as a benchmark and achieved an accuracy of 97.22% as per reference [6]

Data Pre-processing

Proper data pre-processing is required before applying machine learning algorithms to the breast cancer detection dataset to ensure that the models are accurate and dependable. As the first step the data is cleaned by the removal of all the null values.

Feature Selection

Feature selection is an important stage in developing an accurate prediction model, especially for large and complicated datasets like breast cancer data. To select the most significant characteristics, techniques such as correlation analysis, principal component analysis, and forward/backward selection methods can be applied. We have used forward selection by considering AIC as measure of elimination.



```
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                  569
Model:                          Logit   Df Residuals:                      557
Method:                           MLE   Df Model:                           11
Date:                Sun, 08 Jan 2023   Pseudo R-squ.:                  0.9273
Time:                        18:57:20   Log-Likelihood:                -27.299
converged:                       True   LL-Null:                       -375.72
Covariance Type:            nonrobust   LLR p-value:                 2.563e-142
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
intercept        -50.6350     11.673     -4.338      0.000     -73.514     -27.756
ConcavityM        39.4469     15.105      2.612      0.009       9.843      69.051
ConcavePointsM    13.9686     47.987      0.291      0.771     -80.084     108.021
TextureW           0.6155      0.152      4.055      0.000       0.318       0.913
RadiusW            2.3830      0.947      2.516      0.012       0.527       4.239
PerimeterM        -0.2513      0.144     -1.745      0.081      -0.534       0.031
PerimeterSE       -0.3440      1.095     -0.314      0.753      -2.490       1.802
SmoothnessW       62.9414     30.707      2.050      0.040       2.756     123.127
CompactnessSE    -92.6204     31.143     -2.974      0.003    -153.660     -31.581
TextureSE         -3.3906      1.577     -2.150      0.032      -6.482      -0.299
AreaSE             0.2587      0.103      2.519      0.012       0.057       0.460
ConcavePointsW    48.9810     24.276      2.018      0.044       1.400      96.562
==================================================================================

Possibly complete quasi-separation: A fraction 0.70 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
AIC: 78.59700230552244
BIC: 130.7235675150384
Final Variables: ['intercept', 'ConcavityM', 'ConcavePointsM', 'TextureW', 'RadiusW', 'PerimeterM', 'PerimeterSE', 'Smoothne
ssW', 'CompactnessSE', 'TextureSE', 'AreaSE', 'ConcavePointsW']
```

As per the above image the final variables that we have considered for model building are ConcavityM, ConcavePointsM, TextureW, RadiusW, PerimeterM, PerimeterSE, SmoothnessW, CompactnessSE, TextureSE, AreaSE, ConcavePointsW.

Data normalization

Data normalization is an important pre-processing step in many machine learning and statistical modelling applications. The scalar function is one of the most widely used normalization procedures. The goal of data normalization is to convert the input features of a dataset into a common scale so that the algorithm can function properly. We have used Standard scalar function of Sklearn library for normalizing the data.

Data Validation

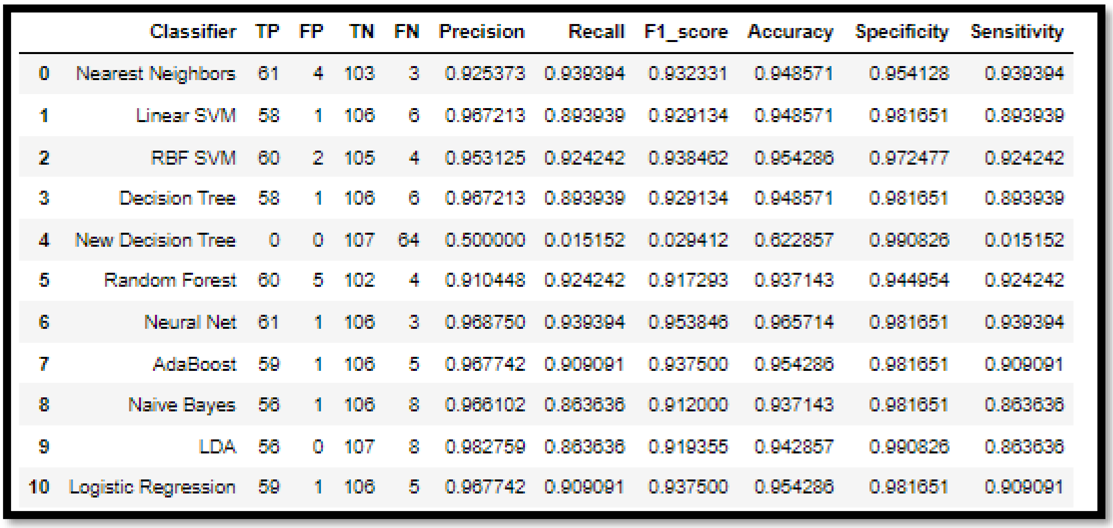
To analyse the effectiveness of our machine learning models, we used train-test split as a data validation technique. To guarantee that the model generalizes effectively to new and unknown data, we divided the dataset into two sets: a training set for training the model and a test set for evaluating its performance.

Train	70%
Test	30%

After the splitting of data we got to know that the dataset is biased towards benign. To address this, we have used **stratify** parameter in train_test_split function.

Methods

We have used 11 different machine learning techniques for predicting the outcome.



	Classifier	TP	FP	TN	FN	Precision	Recall	F1_score	Accuracy	Specificity	Sensitivity
0	Nearest Neighbors	61	4	103	3	0.925373	0.939394	0.932331	0.948571	0.954128	0.939394
1	Linear SVM	58	1	106	6	0.967213	0.893939	0.929134	0.948571	0.981651	0.893939
2	RBF SVM	60	2	105	4	0.953125	0.924242	0.938462	0.954286	0.972477	0.924242
3	Decision Tree	58	1	106	6	0.967213	0.893939	0.929134	0.948571	0.981651	0.893939
4	New Decision Tree	0	0	107	64	0.500000	0.015152	0.029412	0.622857	0.990826	0.015152
5	Random Forest	60	5	102	4	0.910448	0.924242	0.917293	0.937143	0.944954	0.924242
6	Neural Net	61	1	106	3	0.968750	0.939394	0.953846	0.965714	0.981651	0.939394
7	AdaBoost	59	1	106	5	0.967742	0.909091	0.937500	0.954286	0.981651	0.909091
8	Naive Bayes	56	1	106	8	0.966102	0.863636	0.912000	0.937143	0.981651	0.863636
9	LDA	56	0	107	8	0.982759	0.863636	0.919355	0.942857	0.990826	0.863636
10	Logistic Regression	59	1	106	5	0.967742	0.909091	0.937500	0.954286	0.981651	0.909091

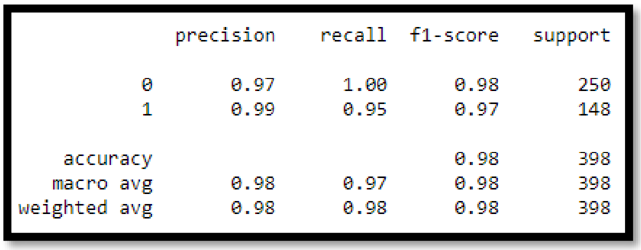
The above image contains the cumulative results of the respective models that we have used for the prediction.

Considering the above results, we have moved on with **logistic regression** [9] for further evaluation due to its simplicity, interpretability, and efficiency. It performs well when the data is linearly separable and easily regularizable to avoid overfitting. Its coefficients can be understood as the impact of each attribute on the anticipated outcome, which improves understanding of the model's behaviour and decision-making process. [8]

Training Results

Classification Report:

In the below report ‘0’ indicates **Benign** and ‘1’ indicates **Malignant** respectively.



```
              precision    recall  f1-score   support

           0       0.97      1.00      0.98       250
           1       0.99      0.95      0.97       148

    accuracy                           0.98       398
   macro avg       0.98      0.97      0.98       398
weighted avg       0.98      0.98      0.98       398
```

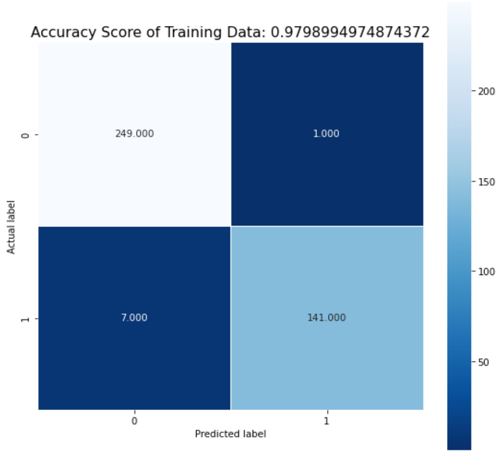
Confusion Matrix:

In the below figure **True Negatives** are under ‘white’ area of the rectangle and **True Positives** are under “Sky Blue” area of rectangle.

The “Dark Blue” area of rectangles indicates **False Positives** and **False Negatives** respectively.

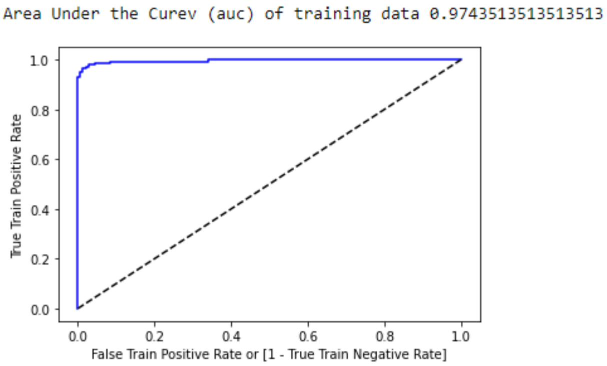
Malignant prediction with **False Negatives** = 7 and **True Positives** = 141.

Accuracy score of the model in terms of training data is very high (97.9%).



ROC Curve:

In the below figure "Blue" line indicates the boundary of AUC (Area under Curve) of Training Data with a percentage of approximately 97.4%.

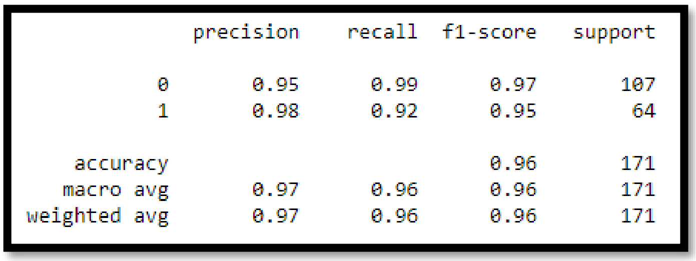


Testing Results

Classification Report:

In the below report '0' indicates **Benign** and '1' indicates **Malignant** respectively.

We can say that the model is performing "Very Good" with **recall** =0.92 in terms of predicting **Malignant.**

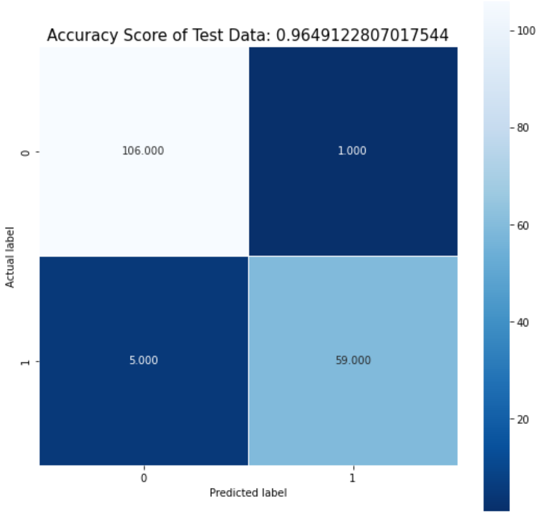


```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97       107
           1       0.98      0.92      0.95        64

    accuracy                           0.96       171
   macro avg       0.97      0.96      0.96       171
weighted avg       0.97      0.96      0.96       171
```

Confusion Matrix:

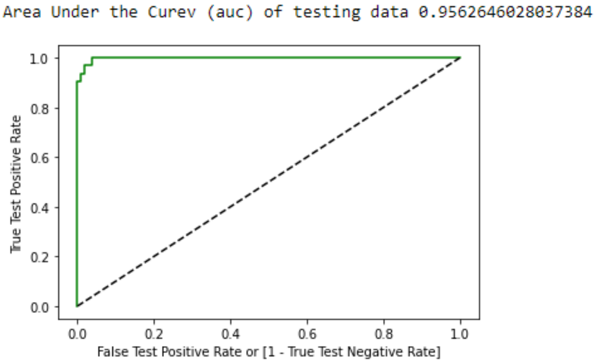
In the below figure **True Negatives** are under 'white' area of the rectangle and **True Positives** are under "Sky Blue" area of rectangle.

The "Dark Blue" area of rectangles indicates **False Positives** and **False Negatives** respectively.



ROC Curve:

In the below figure “Green” line indicates the boundary of AUC (Area under Curve) of Testing Data with a percentage of approximately 95.6%.



Conclusion

In the detection of breast cancer, logistic regression has shown encouraging results. We were able to build a logistic regression model to predict whether a patient has breast cancer or not using a dataset with a variety of variables. The model's performance was assessed using a variety of measures, including accuracy, precision, recall, and F1 score, and it performed well on both the training and testing datasets. Logistic regression offers various advantages in breast cancer detection, including its simplicity, interpretability, and efficiency, and it can be easily modified to address multi-class classification problems.

From this project we have learned several key things. To begin, the significance of data pre-processing in cleaning and formatting data in preparation for analysis. Second, we learned about the procedures in developing a logistic regression model, such as feature selection, data partitioning, and model evaluation using performance measures. Overall, this effort has generated useful insights into the use of machine learning in the identification of breast cancer, which can ultimately improve patient outcomes and save lives.

References

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsindata?resource=download

[2] K. H. R. Al-Hamadi and M. A. Salih, “Breast Cancer Diagnosis Based on SVM Classification,” Journal of Applied Science and Engineering, vol. 21, no. 3, pp. 321–331, 2018.

[3] A. Nasir and M. Mehmood, “Breast Cancer Diagnosis Using Decision Tree Algorithm,” International Journal of Computer Science and Mobile Computing, vol. 5, no. 2, pp. 85–91, 2016.

[4] "Exploring the role of decision trees in breast cancer diagnosis and prognosis" by Díaz-Verdejo, D., Sancho-Gómez, J. L., & López-Guerrero, M. A. (2020).

[5] "A Comparative Study of Machine Learning Techniques for Breast Cancer Classification" by Behera, R. K., & Jena, S. K. (2020).

[6] "Breast Cancer Detection using Convolutional Neural Networks" by Awad, M., Khanna, A., & Musa, A. (2021).

[7] "Breast Cancer Diagnosis Using Machine Learning Techniques: A Review" by Singh, R., & Samanta, D. (2021).

[8] Mohamad, M. S., Zaki, W. M. W., & Mustapha, A. (2020). Application of logistic regression for breast cancer detection using mammogram images. Journal of Physics: Conference Series, 1529(1), 012076

[9] Wang, Q., Wei, Y., & Zhang, H. (2020). A hybrid logistic regression algorithm for predicting breast cancer. Journal of Medical Systems, 44(11), 1-9.