Analysis of Breast Cancer Dataset

Nayan Shivhare
Department of Computer Science
Northern Arizona University
Flagstaff, Arizona, USA
ns977@nau.edu
02-04-2021

***Abstract*—Cancer is one of the most dangerous incurable diseases, and according to WHO breast cancer (BC) is one of the most common cancers among women worldwide and it is the second leading cause of death. The purpose of the paper is to predict whether the cancer is benign or malignant by analyzing the Breast Cancer dataset using data visualization and Machine Learning Methods. In this paper we performed data visualization to get better understanding of the features in data. We implemented four Machine Learning Models including Logistic Regression, K-Nearest Neighbor, Decision Tree and Random Forest are implemented and accuracy obtained from all the four models were compared to get the best model for this dataset. The results of our analysis showed that Logistic Regression is the best model with highest accuracy amongst all four models.**

***Index Terms*—Machine Learning, Analysis, Prediction, Cancer, Data Visualization, Accuracy**

I. Introduction

In Healthcare environment, cancer is one of the significant reasons for death in women. About 2.6% deaths in women's are caused due breast cancer. It was found that early diagnosis of cancer, with early treatment, will improve diagnosis and chance of survival significantly, as it can promote timely clinical treatment to patients. So, using machine learning on the breast cancer dataset we can predict and analyze critical features beforehand for the correct diagnosis, preventing patients from undergoing unnecessary treatments, and help in the timely diagnosis of breast cancer.
Data Science and Machine Learning are the most popular and exciting research areas. It can help us to predict and understand future data. Data Science is critical in keeping track of patients' health and alerting them to the measures that need to be done to deter diseases from developing. It is very difficult to deal with large data manually so for understanding and predicting future data, Machine Learning and Data science is the most straightforward technique. Moreover, to have general idea about the dataset, data analysis and visualization plays a crucial role [3].

The rest of the paper is organized as follows: Section II discusses all the previous work done on Breast Cancer Dataset; Section III provides a overview of the dataset used in this project and the process performed to make it ready for prediction modelling; Section IV explains the importance of Data Visualization and various types of plots; Section V discusses the prediction models implement for this dataset; Section VI explains the results obtained after the analysis.

II. Related Work

Survivability of breast cancer disease has been predicted in [5] by comparing three data mining methods. The aim was to research how to use technological advancements to built ML models for predicting breast cancer survivability. To measure the unbiased estimate of the prediction model, 10-fold cross validation was used. After analysis, results showed that the decision tree is the best predicition model with the accuracy of 93.6% and artifical neural network is the second best predicion model with the accuracy of 91.2% and at last Logistic Regression with the accuracy of 89.2%.

A paper [2] by Agarap presented a comparative analysis of six machine learning algorithms on the breast cancer dataset. The dataset was splitted in 70:30 ratio for implementing the ML algorithms. After measuring all the metrics, results showed that Multi-layer Perceptron is the best algorithm with the test accuracy of 99.04%.

Ak presented a paper [3] to perform comparitive analysis on breast cancer dataset with the help of Machine Learning and Data Mining techniques. Seven Prediction Models (Logistic Regression, K-NearestNeighbor, Support Vector Machine, Naive Bayes, Random Forest, Decision Tree and Rotation Forest) were implemented using python and the results showed that the Logistic Regressions is the best model amongst the seven models as it has the accuracy of 98.1% [3].

III. Dataset

The dataset used for this project has been acquired from UCI Machine Learning Repository. The dataset is related to breast cancer and consists of 569 rows and 33 columns. There are few steps involved in the organization of the process. First and foremost, the process is to then understand the data, convert it to the proper format, and removing the missing values. The libraries used for reading, putting it in a specific format, and cleaning the data will be pandas and Numpy. For analyzing the dataset, we acquired all the required libraries and imported the CSV file. The basic exploration of the breast cancer dataset is done to understand the size and the type of the variables (numerical or categorical variables). To understand the data, python libraries will be used. The dataset is a categorical

dataset that consists of 33 columns and 569 rows as shown in figure 2.

```
Data columns (total 32 columns):
id                         569 non-null int64
diagnosis                  569 non-null object
radius_mean                569 non-null float64
texture_mean               569 non-null float64
perimeter_mean             569 non-null float64
area_mean                  569 non-null float64
smoothness_mean            569 non-null float64
compactness_mean           569 non-null float64
concavity_mean             569 non-null float64
concave points_mean        569 non-null float64
symmetry_mean              569 non-null float64
fractal_dimension_mean     569 non-null float64
radius_se                  569 non-null float64
texture_se                 569 non-null float64
perimeter_se               569 non-null float64
area_se                    569 non-null float64
smoothness_se              569 non-null float64
compactness_se             569 non-null float64
concavity_se               569 non-null float64
concave points_se          569 non-null float64
symmetry_se                569 non-null float64
fractal_dimension_se       569 non-null float64
radius_worst               569 non-null float64
texture_worst              569 non-null float64
perimeter_worst            569 non-null float64
area_worst                 569 non-null float64
smoothness_worst           569 non-null float64
compactness_worst          569 non-null float64
concavity_worst            569 non-null float64
concave points_worst       569 non-null float64
symmetry_worst             569 non-null float64
fractal_dimension_worst    569 non-null float64
dtypes: float64(30), int64(1), object(1)
```

Fig. 1. Imformation about dataset

It is very important to deal with missing values(if any) in the dataset for better prediction and accuracy. Sometimes the missing values are in the form of NAN, nan, and ?. We need to dig deep and find out the type of missing value in data and get rid of it. In the dataset, there is one column 'unamed' which contains all the null values. So, we removed the entire column for cleaning purpose. After removing that column, there was no missing values in data that means only 'unamed' column was having the missing values.
Next, the dataset consists of some numerical variables and some categorical variables. For implementing machine learning models it is very important to convert all the categorical variables to numeric. There are few methods to perform this but for the project, we will be using label encoding. This encoding approach is used to encode categorical variables. It simply converts each value of the column to a number. Pandas library can be used in combination with the .cat.code method in this approach to convert categories to numbers.

The dependent variable in the dataset is 'diagnosis' and it contains two values i.e. 'Benign' and 'Malignant'. We encoded the dependent variable into 0 and 1 using label encoder.

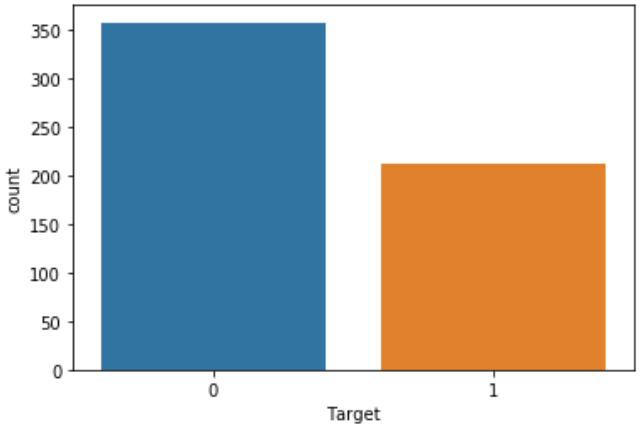

Fig. 2. Count-plot of Target Variable

To normalize/scale the dataset, we performed Feature Scaling. Feature Scaling is a process that is performed on all the independent variables of the data. Sometimes the dataset we are dealing with very high in magnitudes and to handle this we need to apply it to scale to the attributes. For performing the feature-wise scaling in an independent way we used StandardScaler technique.

IV. Data Visualization

After preparing the data for analysis next step is to have a better understanding of each attribute and how it is related to each other and with the dependent variable, we need to use some python visualization libraries such as matplotlib, seaborn, plotly, etc. For analyzing the data we use various plots such as box plot, count plot, scatter plot, histogram, heat map, and pair plot.
Data visualization is a graphical representation of the data for better understanding. To understand and gain insights from the dataset, we can translate the dataset into visual format so that it becomes easier for humans. Data Visualization is important because it makes it possible to identify trends and patterns [8]

A. Histogram:

It shows the frequency on the y-axis and the x-axis shows the column for which you want to know the frequency. To create a histogram we will be using the hist() function.

B. Box Plot:

It is also known as whisker plot and it is used to show the summary of the data including minimum, maximum, first quartile, and third quartile. It is plotted using the boxplot() function,

C. Scatter plot:

This plot is used to plot the data points. Here each value in data is represented by a dot. A method of matplotlib module is used to plot the scatter plots but the two arrays need to be of the same length.

D. Count Plot:

This plot shows the count of an item based on a certain type of category. It is plotted using seaborn library

E. Heatmap:

It is a 2D graphical of data. It is plotted using the seaborn library for analyzing the correlation between the attributes. The motive here is to provide a summary of the data in different colors.

F. Pair Plot:

This plot will show us both the distribution of single variables and the relation between the 2 variables. This plot consists of various pairwise bivariate distribution plots in one graph.

V. Baseline Model

A baseline model is one that is easy to set up and has a good chance of producing good outcomes. When you begin a project, the first thing you can do is think about any possible obstacles that can arise. Baselines, even though they aren't the final iteration of the model, help you to iterate easily by spending as little time as possible [4].

VI. K-Fold Cross Validation

The K-Folds approach is common and simple to understand, and it usually produces a less skewed model than other approaches [10]. The K-Fold CV is a method of splitting a data set into a K number of sections/folds, with each fold serving as a testing set at some stage [9]. The value of K in this project is 5. As a result, the dataset is divided into 5 splits. The first fold is used to test the model, while the rest are used to train it in the first iteration. The second iteration uses the second fold as the testing set and the remainder as the training set. This procedure is repeated until each of the five folds has been used as test set.

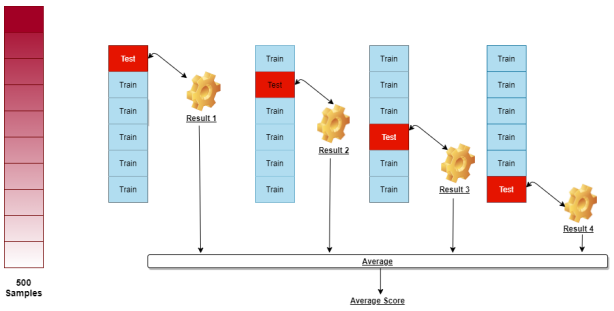

Fig. 3. K-fold Cross Validation- Dividing sample into K number of folds

VII. Prediction Models

A. Logistic Regression:

This prediction model is used for categorical dependent variables to fit models. The result of the logistic regression model between 0 and 1. It is also known as the logit model which is an extension of the linear regression model. This model is used when the dependent variable is binary and in this dataset, our dependent variable is binary(0 or 1). Basically, to characterize data and illustrate the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is used [11].

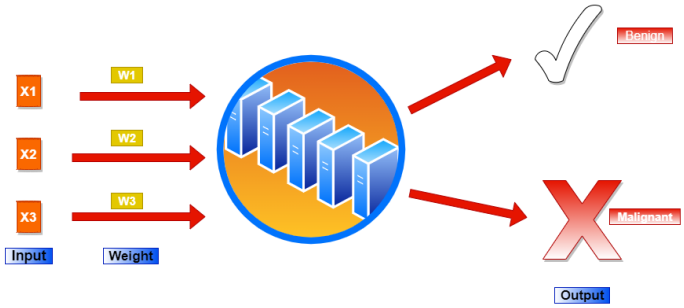

Fig. 4. Logistic regression

B. K-Nearest Neighbors

It is a supervised machine learning algorithm that is used for both classification and regression problems. This algorithm collects all the data points together which are close to it. Moreover, the feature which has a high degree of variation plays a key role in determining the distance. It is also known as a lazy learner as it delays the modeling process until it is necessary to label and classify the data. KNN can also be helpful in solving problems with methods that rely on finding identical artifacts if you have enough computational power to manage the data you're using to make predictions quickly [7].

C. Random Forest

This algorithm is one of the easiest and flexible machine learning algorithm. It is supervised algorithm which is used for both regression and classification problems. When increasing the leaves, the random forest brings more randomness to the model. When breaking a node, it looks for the best feature in a random subset of features rather than the most appropriate feature [6].The random forest begins with a common machine learning technique known as a "decision tree," which refers to our slow learner in ensemble terms [4].

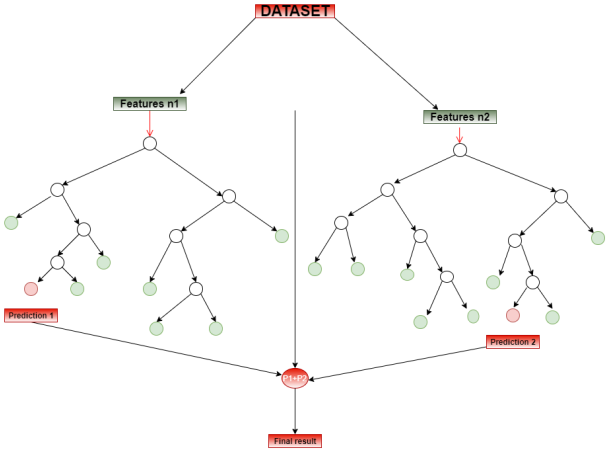

Fig. 5. Random Forest

D. Decision Tree

This is one of the popular algorithms which is also used for both regression and classification problems. It's quick to grasp the data and make some good interpretations because decision trees also imitate human level reasoning. Here, function is represented by each node, connection is represented by branch and outcome is represented by each leaf [1].

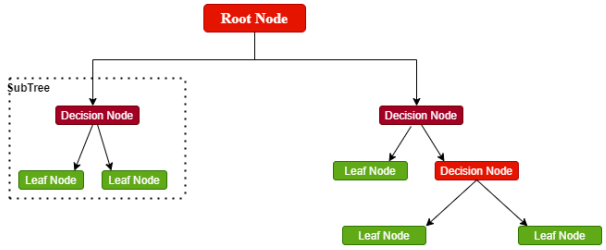

Fig. 6. Decision Tree

VIII. Results and Analysis

Necessary steps were performed to make the data ready for further analysis. We implemented a baseline model and it gave the accuracy of 63%. This baseline model accuracy means that there is high chances of improvement in accuracy by implementing some other models. And, when comparing all other machine learning algorithms on your problem, the scores from these algorithms provide the necessary point of comparison. For comparative analysis, we implemented four Predictive Models i.e Logistic Regression, K-Nearest Neighbor, Decision Tree and Random Forest. Also, we used 5-Fold Cross Validation for evaluating the efficiency of these models. We found accuracy for all these models and shown in the table I

The table consist of accuracy of all the four models at each fold and the average of accuracies for each model. We can conclude that the best predictive model is Logistic Regression with best accuracy(98.5%) amongst four models.

TABLE I
ACCURACY FOR MODELS WITH 5-FOLD CV

Folds	LR	RF	KNN	DT
Fold 1	97.4%	91.6%	89.3%	93.8%
Fold 2	98.6%	91.2%	84.6%	92.3%
Fold 3	97.3%	91.5%	84.9%	91.2%
Fold 4	98.8%	91.2%	86.2%	90.1%
Fold 5	96.5%	90.3%	86.3%	89.1%
Average of all Folds	98.5%	91.1%	85.3%	90.5%

IX. Comparative Study

We have already discussed the previous research performed on this dataset. We compared our result with some of the previous studies and found that those papers have not implemented the baseline model for the analysis. On the other hand, in this paper we have implemented the baseline model and then compared it with four machine learning models.It was not only quick but also helped to understand dataset and provided point of comparison with other machine learning model.In this paper we also tried to achieve models accuracy above 90% and successfully achieved that.

X. Conclusion

Breast Cancer is one of the common diseases in women worldwide. There are two types of breast cancer: benign and malignant as shown in this dataset as well. It is very important and essential to predict breast cancer at an early stage to save the lives of millions of women. For preparing the data we used python libraries for data cleaning , visualization, encoding categorical to numerical data and feature scaling. All these are the basic requirements for data preprocessing. The goal of this project is to predict breast cancer using four machine learning techniques: Logistic Regression, K-Nearest Neighbor, Decision Tree and Random Forest. We compared the accuracy of all these four models and the Logistic Regression came out as the best model amongst all with the accuracy of 98.5%.

References

[1] Madhu Sanjeevi (Mady). Chapter 4: Decision trees algorithms, 2017.
[2] Abien Fred M Agarap. On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In *Proceedings of the 2nd international conference on machine learning and soft computing*, pages 5–9, 2018.
[3] Muhammet Fatih Ak. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In *Healthcare*, volume 8, page 111. Multidisciplinary Digital Publishing Institute, 2020.
[4] CitizenNet. A gentle introduction to random forests, ensembles, and performance metrics in a commercial system.
[5] Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.
[6] Niklas Donges. A complete guide to the random forest algorithm, 2020.
[7] Onel Harrison. Machine learning basics with the k-nearest neighbors algorithm, 2018.
[8] Import.io. What is data visualization and why is it important? 2019.
[9] Krishni. K-fold cross validation, 2018.
[10] Sanjay.M. Why and how to cross validate a model?, 2018.
[11] Statistics Solutions. What is logistic regression?