BACKGROUND

Breast cancer is by far the most frequent type of cancer. It is estimated that approximately two million women are affected each year. It accounts for about fifteen percent of all cancer-related fatalities among women. It is difficult to diagnose the disease in the early stages because the symptoms are not presented well. The National Breast Cancer Foundation recommends that women over the age of forty undergo mammography once a year. The survival rate for women who undergo regular mammograms is higher than that of those who do not. Unregulated cell development results in breast cancer, which must be detected as early as possible to prevent it from developing. Tumors can be classified into two categories, one benign and the other malignant, where the benign tumor is non-cancerous and the latter is cancerous. Researchers are still attempting to develop a proper diagnostic system for detecting the tumor as early as possible and also more easily so that treatment can begin earlier and there will be a higher survival rate. For developing the computerized diagnostic system, machine learning algorithms play an imperative role. Globally, data science is one of the most popular research areas. Data mining and machine learning techniques are straightforward and effective methods of understanding and predicting future data. Dealing with large amounts of data manually is nearly impossible. Data visualization is therefore a vital step to get a general idea about the given data. The application of data analysis techniques has been widely adopted in many fields and is influential in many companies. A variety of disease types is studied using data mining techniques and clustering methods so that the computer can predict current data and make sense of the data.

In this study, public data about breast cancer tumors from Dr. William H.Walberg of the University of Wisconsin Hospital were taken and used for data visualization, classification, and machine learning algorithms. All data used was downloaded from UCI Machine Learning Repository and was extracted from a .csv file. The dataset contains 32 parameters. This study aimed to establish an adequate model by revealing the predictive factors of early-stage breast cancer patients from a wider perspective and compare the strength of the model with accuracy measures.

DATA ANALYSIS AND STRUCTURE

Raw data consists of 32 meaningful features/attributes and 569 samples. I loaded the dataset from the dataset.csv file into a data frame using pandas, then I looked at the first few rows of the dataset. All the attributes have numerical values except ‘id’ and our target variable ‘diagnosis’ which has values either ‘M’ or ‘B’ representing malignant and benign breast cancer respectively. Next, I removed the column 'Unnamed: 32' from the data frame as it didn’t contain any values and was not required for further analysis. Next, I checked for the number of benign and malignant samples in our data frame to check whether the dataset is balanced or imbalanced. After that, I needed to find out the correlation between each of the variables. However, before that, I removed the 'ID' column and ‘diagnosis’ column from the data frame as we only required the independent numerical features for finding out the correlation. So now we are left with just 30 features. Then I check the correlation between different features. The correlation matrix was visualized as a heatmap using the Seaborn Python module. The heatmap had 30 rows and 30 columns, a total of 900 cells, where darker blue colors represent strong positive correlation, while lighter blue colors show a weak positive correlation, and uncorrelated features and darker red colors represent that there was a strong negative relationship between those features, while lighter red colors show weak negative to a weak positive correlation between features. After that, the data was standardized so that each feature has a mean of 0 and a standard deviation of 1. Standardization was required because some of the numerical values were high whereas some were very low. Then I converted the scaled data to a data frame for constructing grouped boxplots for each of the features, to have a look at the statistics of each feature. Using the boxplot some features were excluded based on having many outliers. Then I extracted the positively and negatively correlated features from the correlation matrix as a python dictionary. Then after feature elimination, the following features were grouped as positively correlated radius mean, concavity mean, concave points mean, perimeter worst, area worst, and compactness worst, and the following were grouped as negatively correlated smoothness mean, texture mean, radius mean, fractal dimension mean, texture worst, symmetry se, symmetry mean, area mean. Then I created two different data frames for storing positively and negatively correlated features separately. After that, I constructed scatter plots for clearly visualizing the relationship between the positively, negatively, and non-related features. After that 3 datasets were created from the original scaled dataset. The first dataset consisted of all 32 features, the second dataset consisted of highly correlated features and the third dataset consisted of low correlated features. I implemented the following machine learning algorithms for the task of classifying samples of breast cancer into malignant and benign: Logistic Regression, K Nearest Neighbors, Support Vector Machine, Naïve Bayes, and Random Forest. I have implemented the ML algorithms using the scikit learn python library. Each of the algorithms has been applied to each of the 3 datasets, and I recorded the accuracy for each of the datasets. Before applying the algorithm, I divided the data into training and testing tests. 80 percent of the data was used for training, and the remaining 20 were used for testing. This splitting of the data in the training and testing set has been performed by implementing the 'train_test_split' method from the scikit learn python library. For Logistic Regression, KNN, and Random Forest I used the random state as 42 whereas for Naïve Bayes and SVM I have used the random state as 1 for better results as setting up a random state while splitting the dataset will always split the data in the same proportion, making our results reproducible.

At last I have plotted the accuracy achieved by each of the algorithm on each of the dataset on a stacked bar plot with different algorithms on x axis and their accuracy determined the height of the bars. This bar plot will help to visualize the performance of each algorithm on each of the dataset.

CONCLUSION

In this study, we analyzed the Wisconsin breast cancer dataset, visualized it, and used machine learning algorithms to classify tumor types, whether benign or malignant. The given data was understood by supervised learning algorithms, and future predictions are created based on the given data. Under this technique, classification and regression are two distinct categories. In contrast to regression, classification is a technique for determining the label of data and is used for discrete answers. The initial stage in the categorization process is to read the data. The accuracy scores for logistic regression, k-nearest neighbor, support vector machine, random forest, decision tree, and nave Bayes classification algorithms were calculated in this work. Each algorithm was tested on three different datasets with a variety of characteristics. The first dataset contained all independent features, while the second contained highly correlated features and the third contained low correlated features. In the presented dataset, 62.7 percent of women had a benign tumor type, while 37.3 percent had a malignant tumor type. This distribution reveals that the data was imbalanced, with benign tumors being stored more frequently. The heat map shows the relationship between each feature individually.

Figure 1, a heat map with 900 relationships (30 features * 30 features) to show the relationship between all the features. Darker blue tones indicate that there was a strong and positive link between those characteristics. Lighter blue colors, on the other hand, have a negative connection and are unrelated to benign breast mass. Similarly, darker red colors indicate that those features had an obvious and significant association. Lighter red colors have a negative link with a malignant breast mass and are unrelated to it. For example, with a 1.0 coefficient value, the radius and perimeter mean showed a high and positive association.

Boxplots were designed to reveal the data's basic statistics as well as outliers. Tumor kinds were split into labels, and boxplots for each feature were created. Features that could better classify tumor types were chosen for subsequent applications based on the boxplot results. The benefits of distinguishing between the forms of the normal distribution (e.g., smoothness mean) and other distributions are illustrated graphically in Figure 2. The median was used in all of the boxplot variations examined, although summary statistics around the mean or a number near to it were rarely displayed.

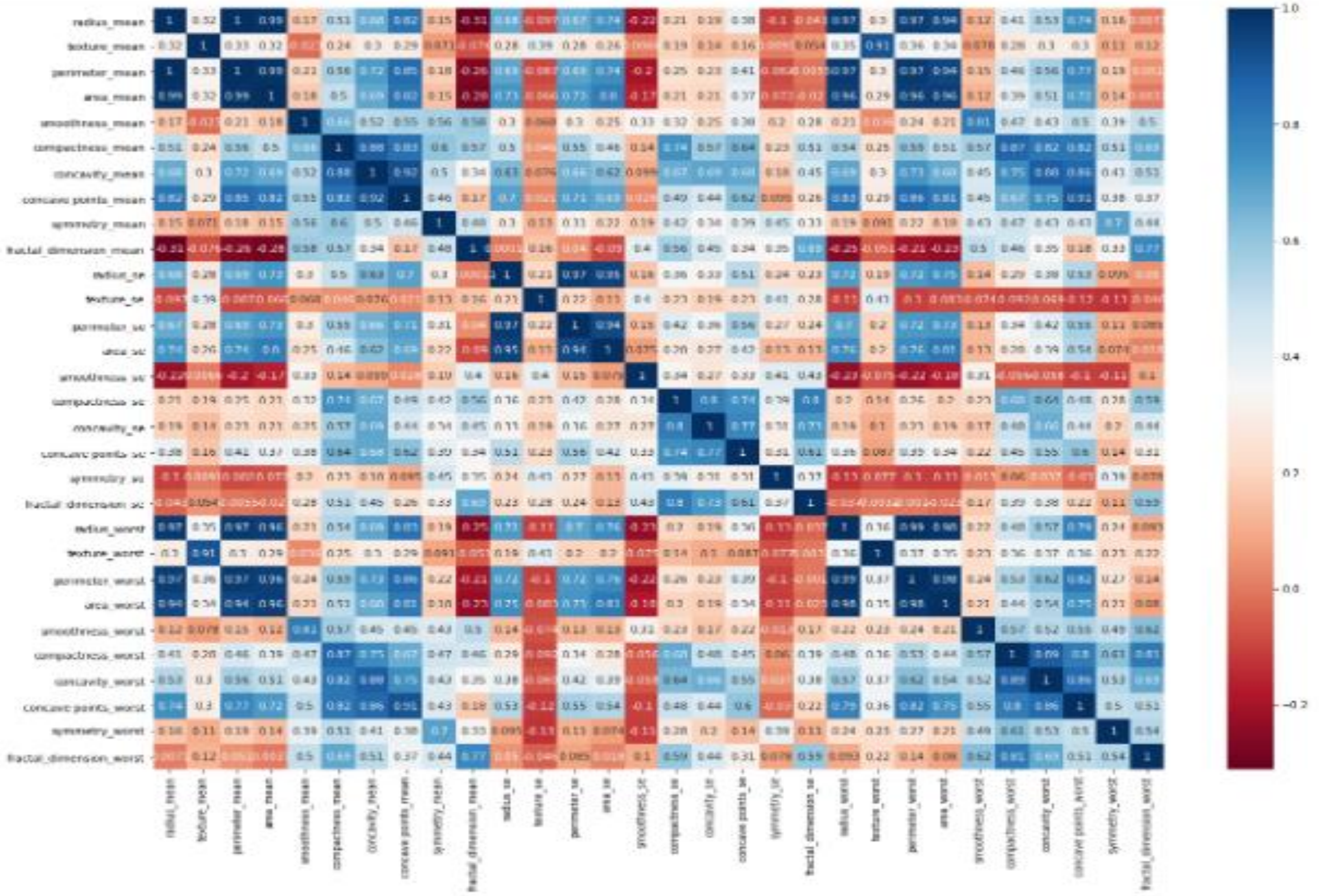

Figure 1: Heat map representative of independent variables.

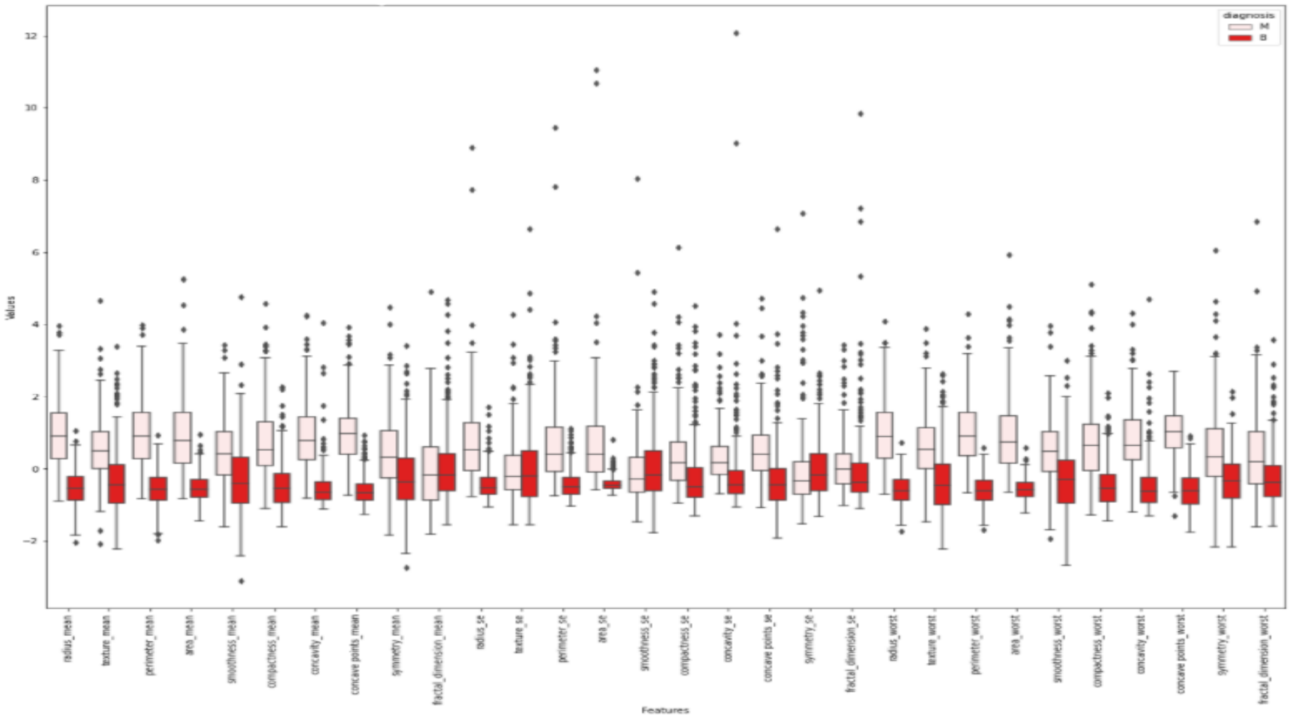

Figure 2: Box Plots

Visualizing the data helped to understand the correlation between each feature and brought out unnecessary features that were not essential to use while making predictions. After visualization was completed, three different datasets were generated. The first dataset covered all features, the second one included highly correlated features, and the last one included features with a low correlation. Machine learning algorithms, which were logistic regression, k-nearest neighbor, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest, were applied for classification of the tumor type. Accuracy results were obtained for the three different datasets. Logistic regression gave better accuracy results rather than the other methods. The main advantage of LR is that it is very efficient to train. In addition, the LR model is useful and gives more accurate results in complex algorithms.

Attributions to the cause of death in those with breast cancer may depend on numerous reasons related to the specifications of the patient. Any specific cause can reduce the risk of death as a result of breast cancer. Nonetheless, these results underline the significance of early diagnoses faced by both active and former patients with a history of breast cancer. Our study reinforces the importance of early diagnosis with high accuracy in women with breast cancer.

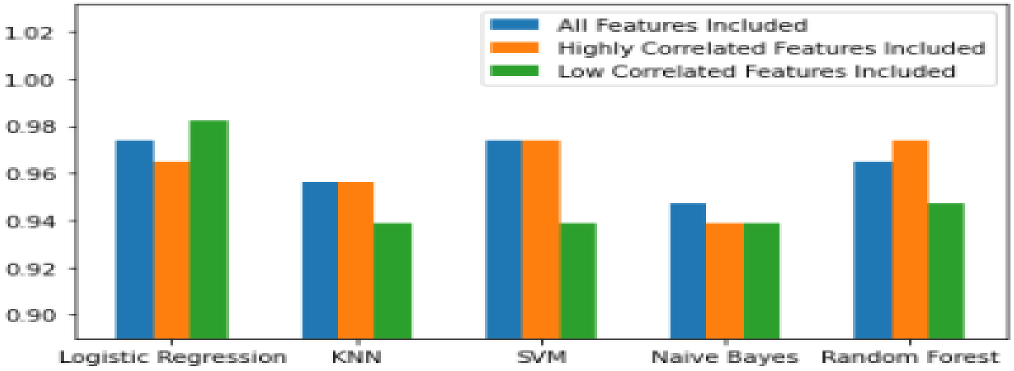

Figure 3: Applied machine learning techniques with accuracy results.

A competitive performance was demonstrated when dealing with imbalanced data. According to the experiment result shown in Figure3, the worst scenario was a decision tree algorithm with low correlated features. Nevertheless, logistic regression with all features included provided the best accuracy result, compared to all other scenarios, with 98.1%. However, it is essential that before running the algorithm, the dataset must be pre-processed, as it does not deal with missing values, and it has a better performance when learning from a dataset with discretized nominal values.

REFERENCE

Ak MF. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare. 2020; 8(2):111. https://doi.org/10.3390/healthcare8020111.