Analysis of Wisconsin's female breast cancer data set

Machine Learning and Data analysis report

Software utilized: R programming language

Prepared by Abdinur Khalif

February 20, 2023

University of Sunderland

Table of Contents

Section 1: Introduction

Cancer has become a major problem in the world. It has been identified as the second most common cause of mortality in the globe, claiming about 10 million people annually. Additionally, breast cancer disproportionately affects women, with 30% of cases becoming fatal (World Health Organization). However, due to increased awareness and improved access to medical treatments, cancer the mortality rate has been steadily declining in recent years.

In addition to the increasingly enhanced medical treatments available to cancer patients, modern Machine Learning algorithms play a major role in detecting breast cancer in its early stages by contributing significantly to the increase in survival rate. In this report, we will train the Naïve Bayes and Decision tree on Wisconsin's breast cancer data set to predict our target class which classifies whether a diagnostic test of patients is benign or malignant (cancerous). Furthermore, we will measure and evaluate the models' performance using the confusion matrix.

Section 2: Data used.

The data set, which is now in the public domain, was obtained from the UCI ML Repository. The data set had been originally collected by Dr. William H. Wolberg. It is composed of 9 variables with a total of 699 records. All variables have integer values except class label which has a factor data type. The class variable contains two values that is "Benign" and "Malignant". Moreover, 66% of cases are benign, while malignant cases make up the other 34%.

In literature, since Machine Learning modeling techniques emerged in early 2000, multiple research papers have been produced to address the accuracy and precision diagnosis of Wisconsin's breast cancer data set. A broad range of ML algorithms are utilized, and the most commonly used are Random Forest, Support Vector Machines, Naïve Bayes, and Neural Networks, among other ML algorithms. In previous studies, You and Rumbe (2010) analyzed the data set by utilizing a range of ML prediction models to classify breast cancer tumors from diagnostic tests recorded in the dataset.

Section 3: Machine learning methods used

The data set is a binary classification problem, thus, to classify the class labels into categories, binary classification models will be applied. By definition, classification models attempt to classify labeled or unlabeled training data into two or more categories for the purpose of providing an insightful picture (Garge, 2018).

After performing ML prediction modeling, rigorous performance measurement will be applied. To measure the model's success, it is imperative to recognize the proportion of cases that are true positive and true negative and vice versa. The criteria by which the success of the model will be measured is how high the rate of classification of True Positive and True Negative cases are. When it comes to which model evaluation method is suitable, the nature of the data always determines the method. According to Suresh (2020), when handling a data set with a class imbalance like the one we are using, Confusion Matrix produces the desired result. Therefore, the Confusion Matrix will be a perfectly ideal solution for measuring the success of the models.

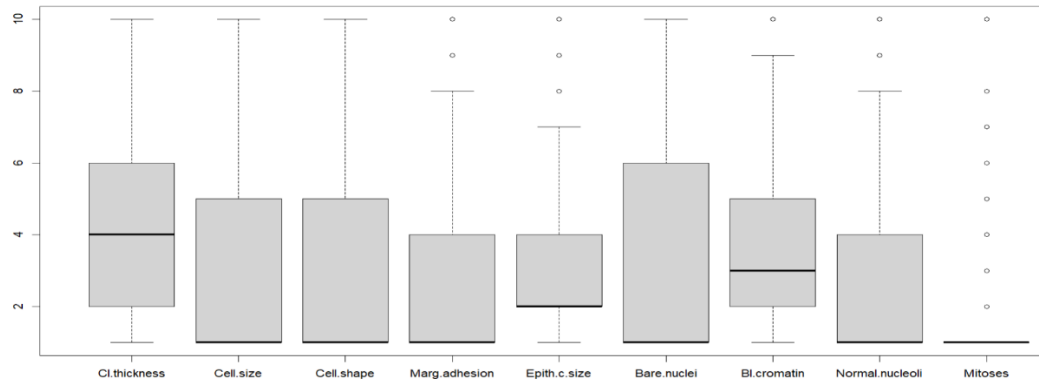
Section 4: Data exploration and preprocessing

After defining criteria for performance measurement of models and ML algorithms that will be utilized, another important step is preprocessing and preparation of data for training and testing on the proposed model. Due to the bulkiness and uncoordinated collection process, data is susceptible to errors, thus preprocessing is a crucial step prior to using it for model building or analysis (Shah and Jivani, 2013). Therefore, we applied a range of data exploration steps including an overview of data structure, correlation, distribution, dimensions, descriptive statistics, and a basic summary of the data.

The following diagrams present the main issues in the data. Further details were provided in the captions attacked with the figures.

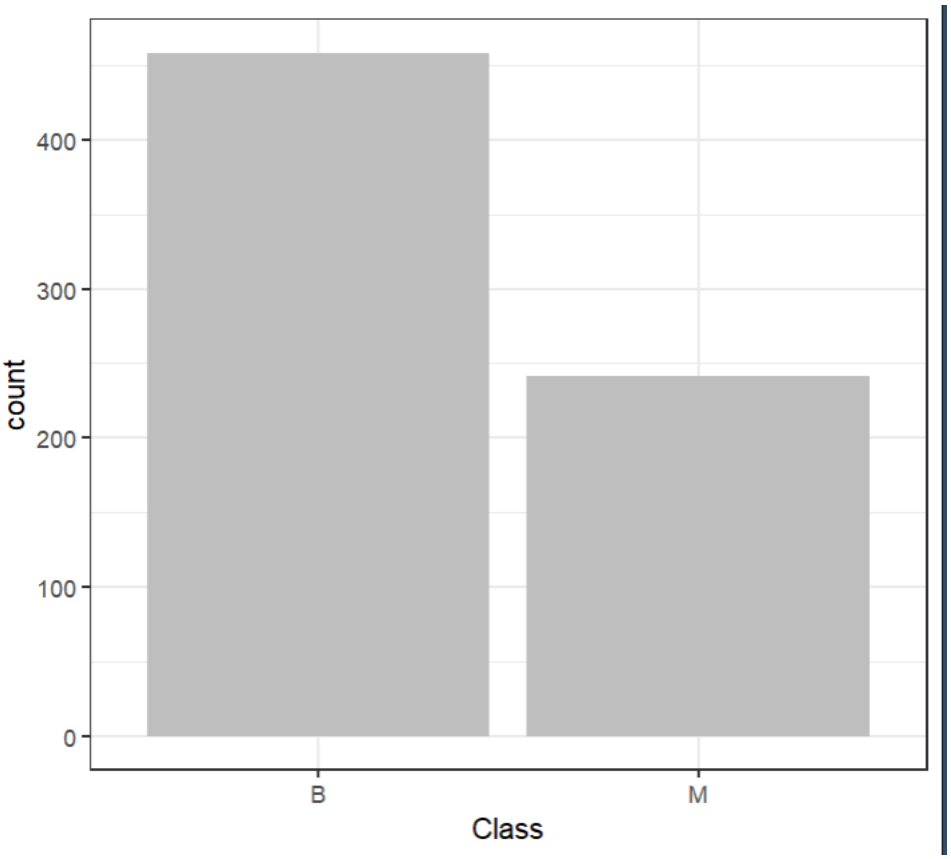
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Cl.thickness	1	699	4.42	2.82	4	4.15	2.97	1	10	9	0.59	-0.63	0.11
Cell.size	2	699	3.13	3.05	1	2.56	0.00	1	10	9	1.23	0.08	0.12
Cell.shape	3	699	3.21	2.97	1	2.67	0.00	1	10	9	1.16	-0.01	0.11
Marg.adhesion	4	699	2.81	2.86	1	2.19	0.00	1	10	9	1.52	0.96	0.11
Epith.c.size	5	699	3.22	2.21	2	2.78	0.00	1	10	9	1.70	2.13	0.08
Bare.nuclei	6	683	3.54	3.64	1	3.06	0.00	1	10	9	0.99	-0.81	0.14
Bl.cromatin	7	699	3.44	2.44	3	3.10	1.48	1	10	9	1.10	0.17	0.09
Normal.nucleoli	8	699	2.87	3.05	1	2.23	0.00	1	10	9	1.42	0.45	0.12
Mitoses	9	699	1.59	1.72	1	1.12	0.00	1	10	9	3.55	12.51	0.06
Class*	10	699	1.34	0.48	1	1.31	0.00	1	2	1	0.65	-1.58	0.02

Descriptive statistics of variable



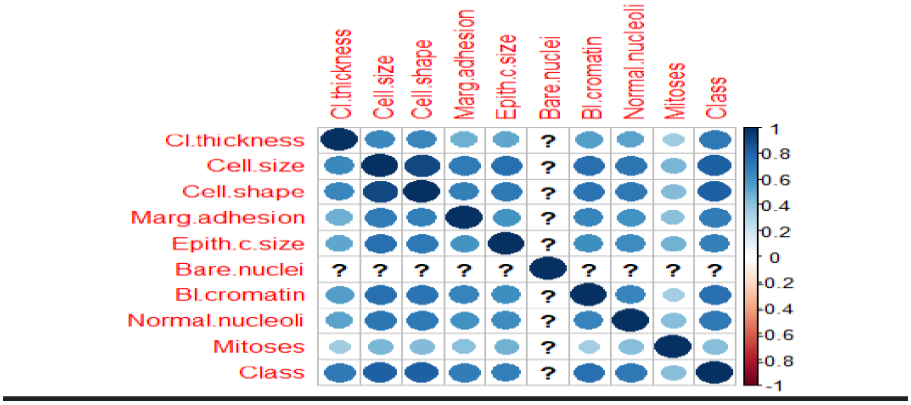
Boxplots

The figure indicates outliers exist in the data. Mitoses has the highest number of outliers which can influence our model result.




```
+             is.numeric))], use = 'complete.obs')  # to run correlation matrix, missing values were completed
               Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli   Mitoses
Cl.thickness       1.0000000 0.6424815  0.6534700     0.4878287    0.5235960   0.5930914   0.5537424       0.5340659 0.3509572
Cell.size          0.6424815 1.0000000  0.9072282     0.7069770    0.7535440   0.6917088   0.7555592       0.7193460 0.4607547
Cell.shape         0.6534700 0.9072282  1.0000000     0.6859481    0.7224624   0.7138775   0.7353435       0.7179634 0.4412576
Marg.adhesion      0.4878287 0.7069770  0.6859481     1.0000000    0.5945478   0.6706483   0.6685671       0.6031211 0.4188983
Epith.c.size       0.5235960 0.7535440  0.7224624     0.5945478    1.0000000   0.5857161   0.6181279       0.6289264 0.4805833
Bare.nuclei        0.5930914 0.6917088  0.7138775     0.6706483    0.5857161   1.0000000   0.6806149       0.5842802 0.3392104
Bl.cromatin        0.5537424 0.7555592  0.7353435     0.6685671    0.6181279   0.6806149   1.0000000       0.6656015 0.3460109
Normal.nucleoli    0.5340659 0.7193460  0.7179634     0.6031211    0.6289264   0.5842802   0.6656015       1.0000000 0.4337573
Mitoses            0.3509572 0.4607547  0.4412576     0.4188983    0.4805833   0.3392104   0.3460109       0.4337573 1.0000000
```

Correlation matrix



correlation heat amp

There is strong correlation among most of the variables. Cell.size variable has the highest correlation with many other variables. The darker the spot, the higher the correlation of predictors. Feature extraction will import consideration.

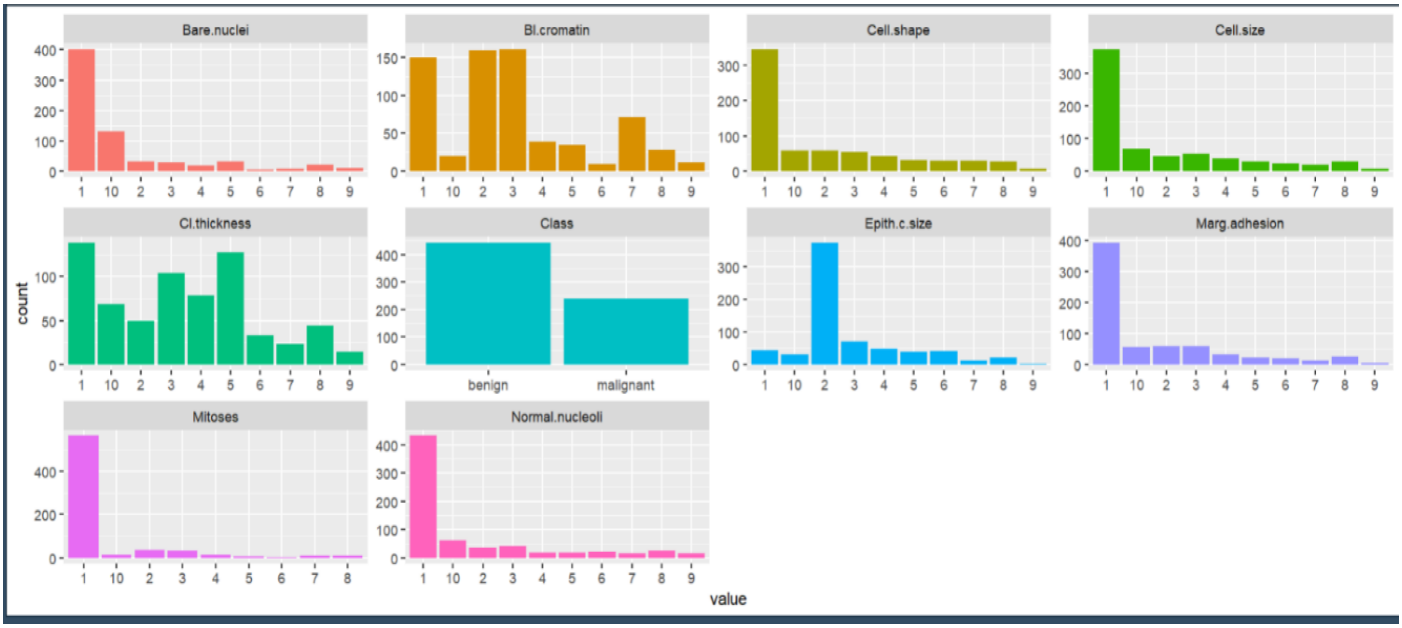


Figure 1: bar graphs indicating variable distributions.

As the figure indicates, not only do outliers exist, but many variables are also skewed to the left. There is also class imbalance that creates bias in the result of the classification model.

After doing data exploration, we took the following steps to clean up the data. First, we found that there is a variable that has a total of 16 missing values, so removed records with missing values to avoid it affecting our model result. We also discovered that the correlation coefficients of some variables are statistically insignificant, thus we checked whether insignificant p values were caused by multicollinearity.

```
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)   0.076860   0.116839   0.658              0.51087
Cl.thickness  0.059247   0.023068   2.568              0.01043 *
Cell.size     0.619975   0.031512  19.674 < 0.0000000000000002 ***
Marg.adhesion 0.009568   0.024638   0.388              0.69789
Epith.c.size  0.051910   0.032250   1.610              0.10795
Bare.nuclei   0.063099   0.022419   2.815              0.00503 **
Bl.cromatin   0.024716   0.031351   0.788              0.43075
Normal.nucleoli 0.072514   0.023136   3.134              0.00180 **
Mitoses       0.001707   0.030560   0.056              0.95547
Classmalignant 0.591985   0.236056   2.508              0.01238 *
```

Section 5: Model training with R programming language

After preparing data for analysis and training, Naïve Bayes and the Decision tree Machine Learning classification algorithms were built using Wisconsin's breast cancer data set. As the screenshots of the code shown, several steps were taken to build the models. The first step was to spilt data into two groups: training and testing data for model prediction. For both classification models, 65% of the data were used as training data and the rest for testing. We employed the Rpart package to visualize the tree model. We also performed a performance evaluation of the models to determine which provides the best prediction of diagnosis.

Section 6: Analysis results and visualization

In the result section, the report presents the finding of the classification models. During the classification process, as the report mentioned in the model building phase, we first trained the models with the training set that makes up 65% of the data set and tested with the rest to assess the effectiveness of the former set. subsequently, we used the confusion matrix characterization models to evaluate and compare the performance of trained models. The models yielded slightly different results during testing and training. The decision tree model accurately classified benign cases 286 times as benign tumors, and 12 times inaccurately identified benign cases as malignant cases. On the other hand, it has accurately identified 152 times the malignant cases as malignant cases and misclassified 5 cases of malignant tumors as benign. Overall, the decision tree has yielded a high accuracy rate (84%) in both training and testing data classification.

```
prediction  benign malignant
  benign       286         5
  malignant     12       152
```

Performance evaluation of dt training and testing results

```
prediction  benign malignant
  benign       149         5
  malignant     11        79
```

However, the Naïve Bayes model performed better in both cycles of training than the Decision tree. Additionally, as the screenshot below indicates, the model yielded a slightly better result in testing phase when compared to training. It has a 97% of accuracy rate, 97 specificities, and 98% balanced accuracy. Overall, the Naïve Bayes returned the highest accuracy rate in prediction when compared to the Decision tree model.

```
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       151         4
  malignant      4        79

               Accuracy : 0.9664
                 95% CI : (0.9348, 0.9854)
    No Information Rate : 0.6513
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.926

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9742
            Specificity : 0.9518
         Pos Pred Value : 0.9742
         Neg Pred Value : 0.9518
             Prevalence : 0.6513
         Detection Rate : 0.6345
   Detection Prevalence : 0.6513
      Balanced Accuracy : 0.9630
```

```
Confusion Matrix and Statistics

            Reference
Prediction   benign malignant
  benign        282         4
  malignant       7       152

                 Accuracy : 0.9753
                   95% CI : (0.9562, 0.9876)
      No Information Rate : 0.6494
      P-Value [Acc > NIR] : <2e-16

                    Kappa : 0.946

 Mcnemar's Test P-Value : 0.5465

              Sensitivity : 0.9758
              Specificity : 0.9744
           Pos Pred Value : 0.9860
           Neg Pred Value : 0.9560
               Prevalence : 0.6494
           Detection Rate : 0.6337
     Detection Prevalence : 0.6427
        Balanced Accuracy : 0.9751
```

Section 7: Conclusion

Breast cancer is associated with a high fatality rate and is considered the second deadliest disease on the earth. Early detection of this disease helps identify breast cancer tumors before they become malignant and life-threatening. Modern Machine Learning algorithms can be used to classify and cluster the current status of breast cancer. The aim of this report was to build ML classifier models by using Wisconsin's breast cancer data set to predict breast cancer diagnostic tests into malignant or benign tumors. To perform the classification task, the Naïve Bayes and Decision tree ML models were employed. Moreover, to test the effectiveness of classification accuracy, we used the confusion matrix table. During model evaluation, the Naïve Bayes proved to have a superior classification accuracy rate. Finally, we discovered the importance of adjustment of model parameters such as removing and reintroducing the variables that possess statistically insignificant P values to assess the overall relative importance of predictors. Future research will need to employ more advanced ML classification models such as Radom Forest and Neural Networks to achieve superior classification accuracy.

Section 9: References list

Garge, R., (2018). *7 Types of Classification Algorithms.* [Online]
Available at: https://analyticsindiamag.com/7-types-classification-algorithms/#:~:text=Classification%20model%3A%20A%20classification%20model,of%20a%20phenomenon%20being%20observed.
[Accessed 1 Jan 2023].

McGarry, K. (2020) 'Performing Classification', CETM72: Data Science Principles. University of Sunderland. (N.d). Available at: https://study.online.sunderland.ac.uk/courses/659/pages/5-dot-6-activity-rstudio-perform-classification?module_item_id=55835(Accessed: 4 January 2023)

Shah, C. and Jivani, A.G., 2013, July. Comparison of data mining classification algorithms for breast cancer prediction. In *2013 Fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1-4). IEEE.

Suresh, A., 2020. *What is the Confusion Matrix.* [Online]
Available at: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5#:~:text=A%20good%20model%20is%20one,for%20your%20machine%20learning%20model.
[Accessed 5 January 2023].

Word Health Organization[WHO], (2020). *Cancer.* [Online]
Available at: https://www.who.int/news-room/fact-sheets/detail/cancer
[Accessed 3 Jan 2022].

You, H. and Rumbe, G., 2010. Comparative study of classification techniques on breast cancer FNA biopsy data.