Course Number: DATA601 *Instructor: Dr. Osei Tweneboah*

Analyzing Performances of Different Binary Classification Algorithms on Breast Cancer Data

Christopher Hainzl, Muzamal Sheikh, and Moustafa Ayoub*

*** Student Emails:**

1. chainzl@ramapo.edu
2. msheikh1@ramapo.edu
3. mayoub4@ramapo.edu

Abstract:

The question we wanted to address is which of these ten predictor variables appear to have the largest impact on whether a breast cancer ends up being classified as malignant or benign: the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and/or fractal dimension of the cancer. The purpose of this study was to analyze the performances of several different supervised learning algorithms that are typically used for binary classification. The supervised learning algorithms chosen for this study were: logistic regression, support vector machine, decision tree, random forest, and K-nearest neighbor (KNN). Bagging and voting classifiers were also utilized. The three classification methods that were most accurate based on our calculations in decreasing order were the random forest, voting classifier, and bagging classifier. The three variables that were most significant, in descending order, were radius, perimeter, and area. This means that out of all of these algorithms, data scientists should consider using a random forest the most. And out of all of these predictor variables, the ones that data scientists should really pay attention to are the radius, perimeter, and area of the cancer.

Keywords: Malignant, benign, radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension

1. Introduction

We wanted to determine the most significant variables for predicting whether a breast cancer gets classified as malignant, meaning that it is life-threatening, or benign, meaning that it is not life-threatening [1]. This can help data scientists decide which attributes of the cancer they should look at when they want to classify future cancers into one of these classes. Ten predictor variables were taken into consideration: the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of the cancer. We also wanted to compare the performance of several different binary classification algorithms on these predictor variables. This can help data scientists decide which model is best for classifying cancers, so they can put it to use in the future. In the end, we concluded that the most important features to keep in mind are the radius, perimeter, and area.

2. Materials and Methods

Regarding the data that was used for this project, we worked with the breast cancer dataset located on the UCI Machine Learning Repository website, which can be accessed with the following link: Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository. To conduct the data analysis, all of the work was done using Google Colab. This enabled us to easily collaborate with one another on the Python code we wrote and discuss the results of our analysis. Several different methods were implemented for this project. Regarding the algorithms that were analyzed, the following were chosen because they are primarily used for binary classification: logistic regression, support vector machine, decision tree, random forest, and K-nearest neighbor (KNN). To help assess the performance of each of these algorithms on the data, we calculated F1 scores instead of accuracy since the data's labels were not evenly distributed. In addition to this, ensemble accuracy tools such as voting and bagging classifiers were used to help further assess the performance of some of these algorithms.

But before we could create any models or conduct any calculations, we had to figure out an effective way to work with the variables in the dataset. The original dataset contains three columns corresponding to different measurements for each variable (i.e. 'radius1', 'radius2', 'radius3'). To work around this, we implemented feature engineering by only taking the columns corresponding to the first measurement for each variable into consideration. Those were the columns that we used while generating training and testing data for each of our models. The values in the 'Diagnosis' column were either 'M' or 'B', so we converted those to numerical values (1 for 'M', and 0 for 'B') before we progressed any further.

3. Results

3.1 - Description of Results

3.1.1 - Experimental Results

- **We chose to conduct the following experiments:**
    - **Distribution of Diagnoses:** Figure 1 shows the distribution of the diagnoses from the dataset. Since the labels in our dataset are not evenly distributed, this means we will have to work with the F1 score instead of accuracy. The targets, ‘M’ (meaning malignant) and ‘B’ (meaning benign), are mildly imbalanced.
    - **Histogram Distribution of First 10 Features:** Figure 2a shows a collection of histograms, each representing the distribution of a different variable, such as radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension.
    - **Boxplot Feature Distribution for Target Diagnoses:** Figure 2b displays a collection of box plots for two categories labeled ‘M’ and ‘B’, across various measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension. These box plots compare the distributions between the two target variables for each measurement, indicating differences in central tendency, variability, and potential outliers. The box plot represents the IQR which is the middle 50% of the data set. The whiskers extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles. The points outside the range are considered to be outliers and are plotted separately from the box plot.
    - **Violinplot Plot Feature Distribution For Target Diagnoses:** The violin plot in Figure 2c shows a collection of violin plots, each comparing two categories labeled ‘M’ and ‘B’ across various variables, including radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension. Violin plots are useful for displaying the probability density of the data at different values, often revealing peaks in the data distribution and showing the full range of data, indicating the median and interquartile range within each category, which could suggest a comparison of these variables between two different conditions or classifications in a dataset, such as medical diagnoses. Similar to the box plot, the violin plots showed wider distributions of the target value ‘M’ for each radius, texture, and perimeter. In general, the values on average for each respective feature were also higher for the target value ‘M’.

- **Correlation Matrix:** Figure 3 shows a heat map of a correlation matrix, with various variables such as radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension. The colors range from black to peach, representing correlation coefficients from -1 to 1, where black indicates a strong negative correlation, and peach indicates a strong positive correlation; this type of visualization is commonly used in statistics to assess the degree of a linear relationship between variables.
- **General Linear Model Regression Results:** Figure 4 is the output of a statistical analysis, specifically the results of a Generalized Linear Model (GLM) regression with a binomial family and a logit link function, likely used to model binary outcomes such as medical diagnoses. The model uses 569 observations and includes 3 predictors. The coefficients (coef), standard errors (std err), z-values, p-values (P>|z|), and 95% confidence intervals ([0.025, 0.975]) for each predictor are provided. The significance of predictors can be inferred from the p-values; for instance, 'radius1' and 'area1' are statistically significant ($p < 0.05$), implying they have a meaningful association with the diagnosis. The overall fit of the model is assessed by the Pseudo R-squared value, which is 0.5923 in this case, indicating a moderate to strong relationship between the predictors and the dependent variable.
- **Feature Importances In Random Forest Using MDI:** Figure 5 is a bar chart representing the feature importances derived from a Random Forest model, as measured by Mean Decrease in Impurity (MDI). The chart displays radius, perimeter, and area, along with their corresponding importance values. The length of the bars indicates the importance of each feature in the model, with longer bars signifying greater importance. The error bars indicate variability in the importance metric across the trees within the forest. Features like radius show relatively high importance, while features like area have lower importance in this particular model.

3.1.2 - Experimental Interpretation

- **Distribution of Diagnoses:** Figure 1 is a great way to visualize the distribution of the labels, as it shows how the labels are not evenly distributed.
- **Histogram Distribution of First 10 Features:** The series of distributions for the features shown in Figure 2a followed either a right-skewed or normal distribution, which is common in most medical datasets that showcase characteristics in a cancer study.
- **Boxplot Feature Distribution for Target Diagnoses:** Let’s take a look at Figure 2b. By analyzing the spread of radius1, texture1, and perimeter1,

they are all shown to have a higher median for ‘M’ when compared to ‘B’, indicating that malignant cases tend to have a larger radius, greater visibility in texture, and a larger perimeter in texture. The area of the tumor is generally represented to be larger for malignant cases as well. A higher median value for compactness and smoothness of a tumor also seems to indicate a higher compactness of the tumor in malignant cases. Concavity, the concave points and the fractal dimensions also all show a higher median value for the malignant cases. In general, the boxplot demonstrates that a malignant tumor generally has higher values on average for each respective feature that we studied.

- **Violinplot Plot Feature Distribution For Target Diagnoses:** Figure 2c shows the similarities and differences. Much like the boxplot, the violin plot also demonstrated that a malignant tumor generally has higher values on average for each respective feature that we studied. Features like radius, area, and concave points appear to have a broader distribution and higher median values in malignant cases than benign ones.
- **Correlation Matrix:** Figure 3 shows the results of correlation between any two given features. Our results showed that radius and area had a 0.99 correlation coefficient. Similarly, area and perimeter had a 0.99 correlation coefficient as well. As mentioned before, a value close to 1 is indicative of a significant positive correlation. As you can imagine, any two pairs with correlation coefficient of over 0.8 may indicate how well they can predict whether a tumor is benign or malignant. Because of this, and for the sake of simplicity, we used feature engineering again by only selecting radius, area, and perimeter for usage in our models.
- **General Linear Model Regression Results:** In Figure 4, radius, area, and perimeter all have p-values less than 0.05 which means that we reject the null hypotheses associated with each of those variables. Radius, for example, is a significant predictor as to whether a tumor gets classified as benign or malignant.
- **Feature Importances In Random Forest (MDI):** As visualized in Figure 5, a recurring theme emerges, showing how, in decreasing order, the radius, perimeter, and area are good predictors of determining if a tumor is malignant or benign.
- **Ensemble of Machine Learning Model with Performance Plotting:** Figure 6 depicts a bar chart showing the F1 scores of some of the machine learning models used, including the decision tree, bagging classifier, support vector machine (SVM), K-nearest neighbors (KNN), voting classifier, and random forest. The F1 score is a measure of a model's accuracy, considering both precision and recall, and is particularly useful

when the class distribution is uneven. Scores are out of 1, with higher values indicating better performance. According to the chart, the random forest model has the highest F1 score, suggesting it performs the best among the models evaluated, followed closely by the voting classifier. The decision tree model has the lowest F1 score, indicating it is the least effective model in this comparison.

3.1.3 - Experimental Conclusions

- Based on our findings, we can conclude that the algorithm which is most useful in predicting whether a cancer gets classified as benign or malignant is the random forest model. And out of all the variables that we decided to look at once analyzing the correlation matrix, the radius is the most significant predictor, while the area is least significant. The cancer's perimeter is the second most significant out of all three of these predictor variables.
- The predictive power of these variables suggests that the physical features of the tumor, such as the shape and size, are critical in the diagnosis process, which also aligns with the medical understanding that larger, irregular tumors are, at times, malignant. The prominence of these variables confirms the hypothesis that malignant tumors often present themselves with more noticeable physical features. Additionally, the physical features' high importance highlights the need for detailed imaging and measurements in early cancer detection and classification.

3.2 - Figures

Figure 1. Bar Graph For Distribution of Diagnoses

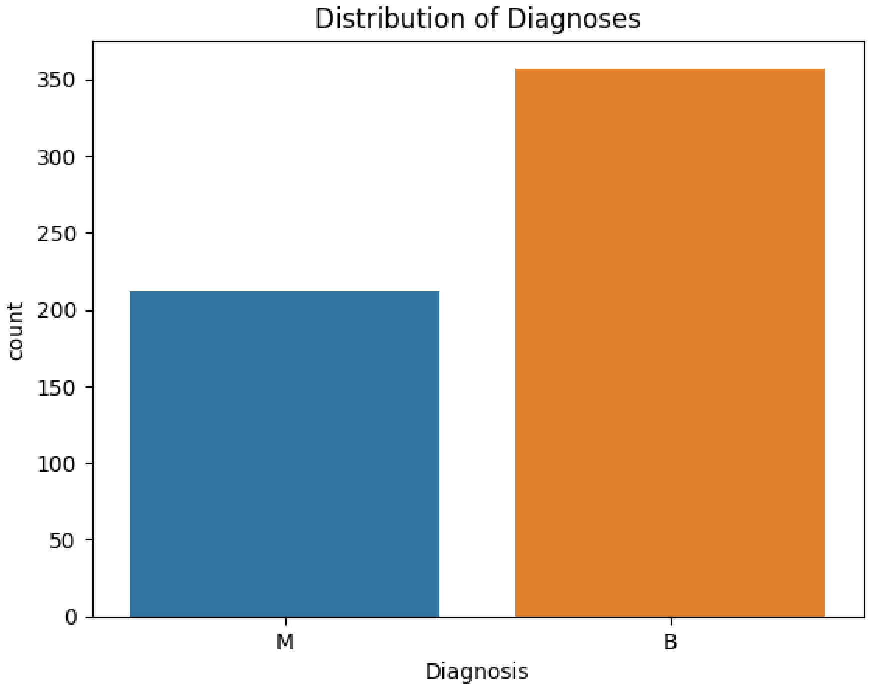


Figure 2a. Histograms of Each Feature Of Importance

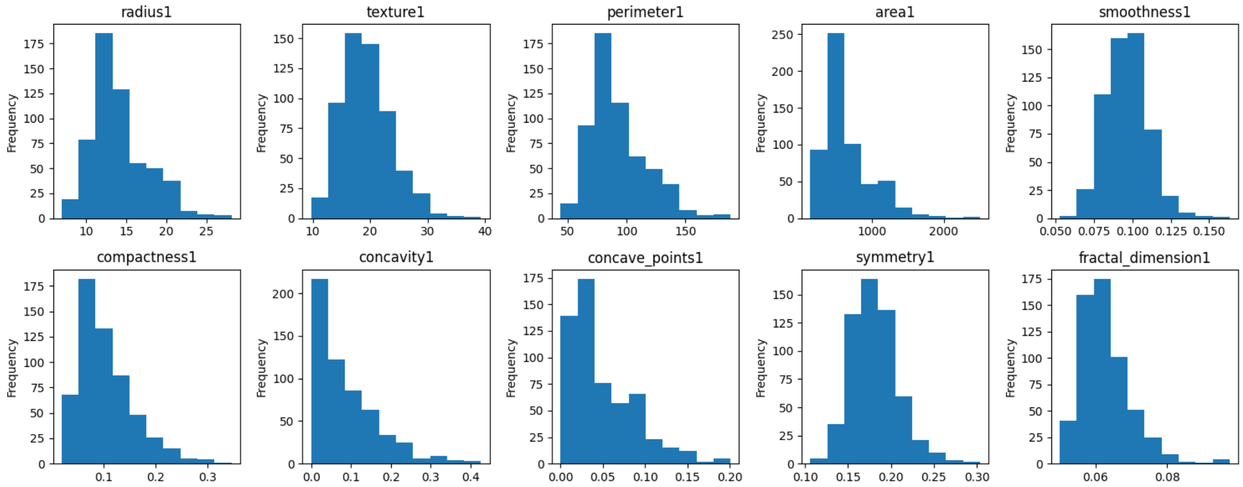


Figure 2b. Boxplots of Feature Distribution For Respective Target Diagnosis

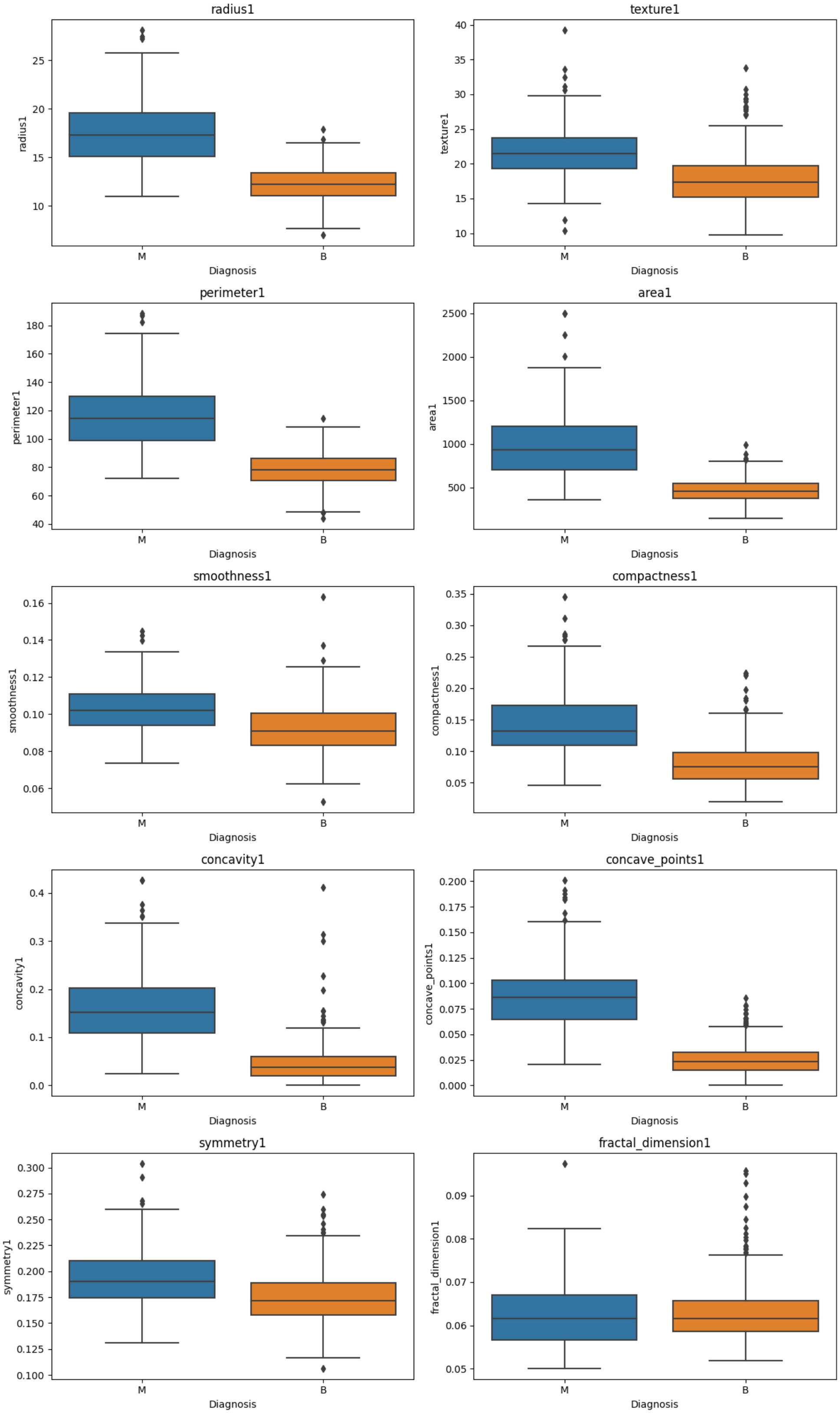


Figure 2c. Violin Plots of Feature Distribution For Respective Target Diagnosis

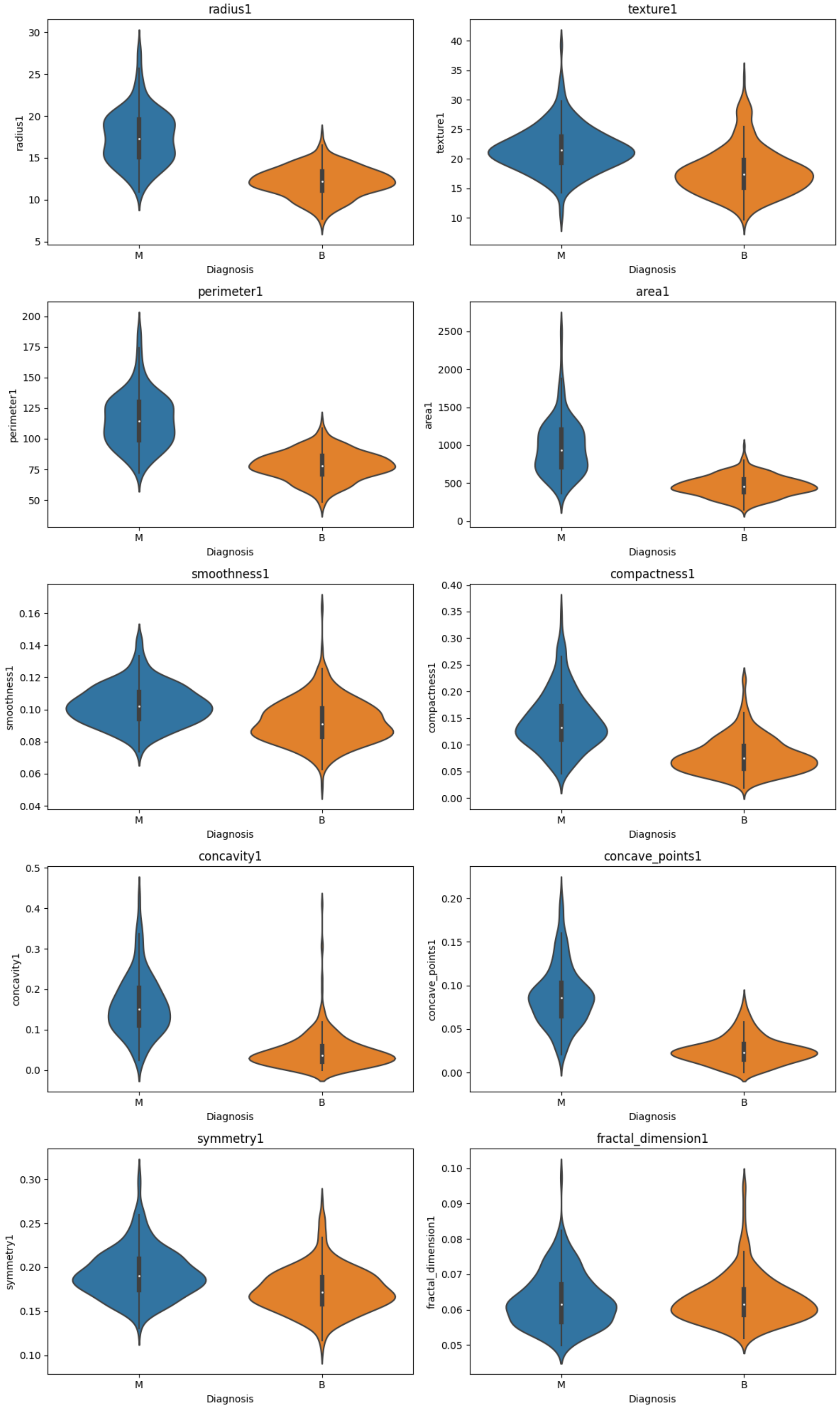


Figure 3. Correlation Matrix

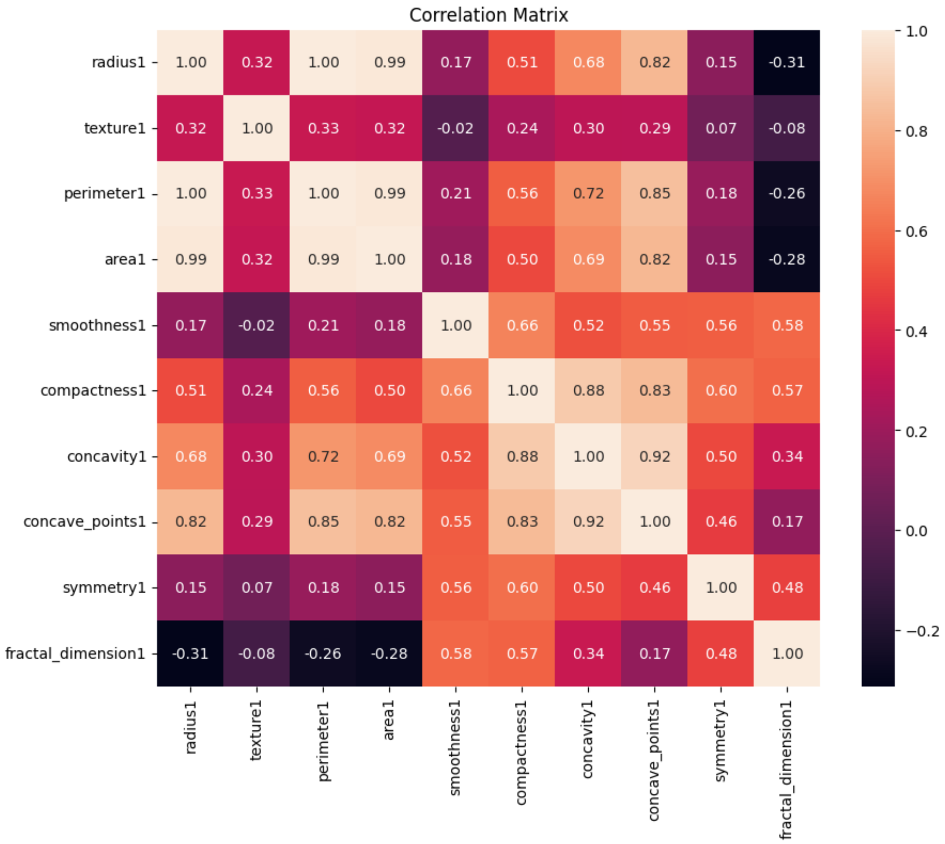
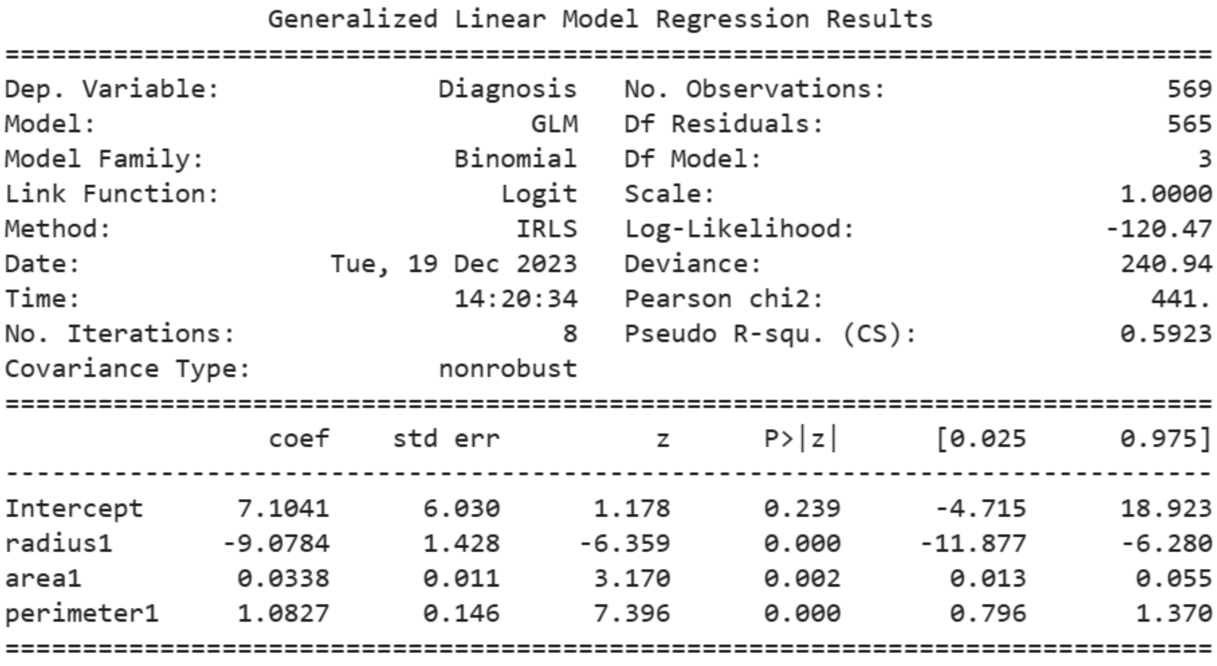


Figure 4. Generalized Linear Model Regression Results



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Diagnosis   No. Observations:                  569
Model:                            GLM   Df Residuals:                      565
Model Family:                Binomial   Df Model:                            3
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -120.47
Date:                Tue, 19 Dec 2023   Deviance:                       240.94
Time:                        14:20:34   Pearson chi2:                     441.
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5923
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.1041      6.030      1.178      0.239      -4.715      18.923
radius1       -9.0784      1.428     -6.359      0.000     -11.877      -6.280
area1          0.0338      0.011      3.170      0.002       0.013       0.055
perimeter1     1.0827      0.146      7.396      0.000       0.796       1.370
==============================================================================
```

Figure 5. Feature Importances in Random Forest (MDI)

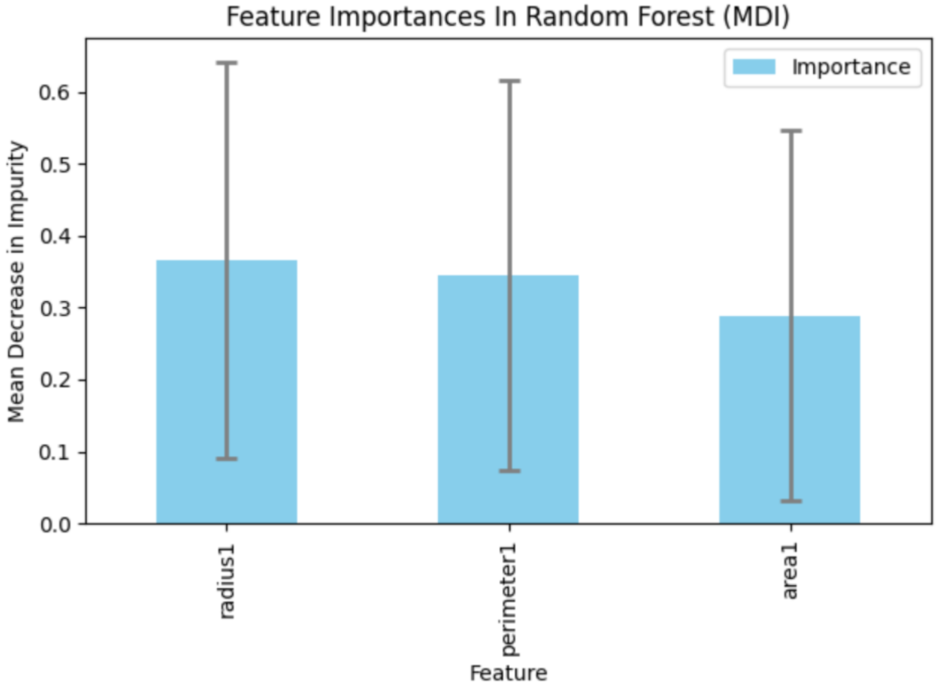


Figure 6. F1 Scores Of Different Machine Learning Models

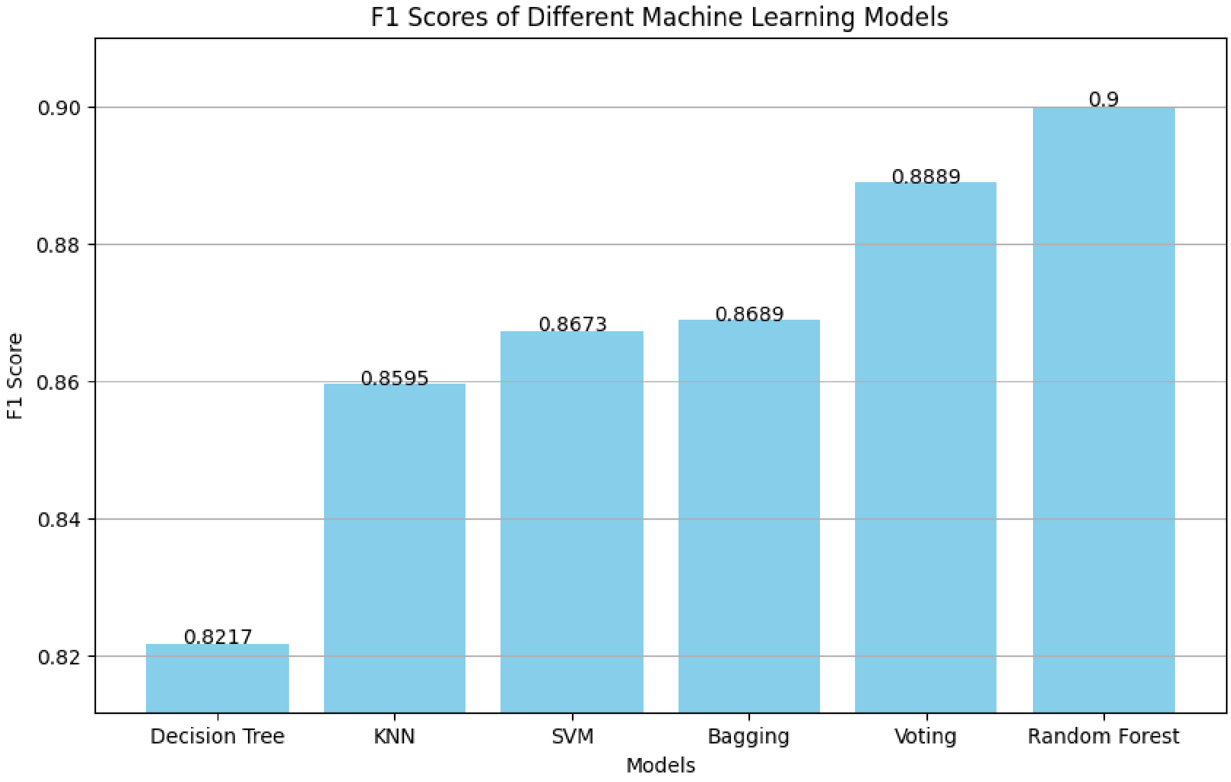
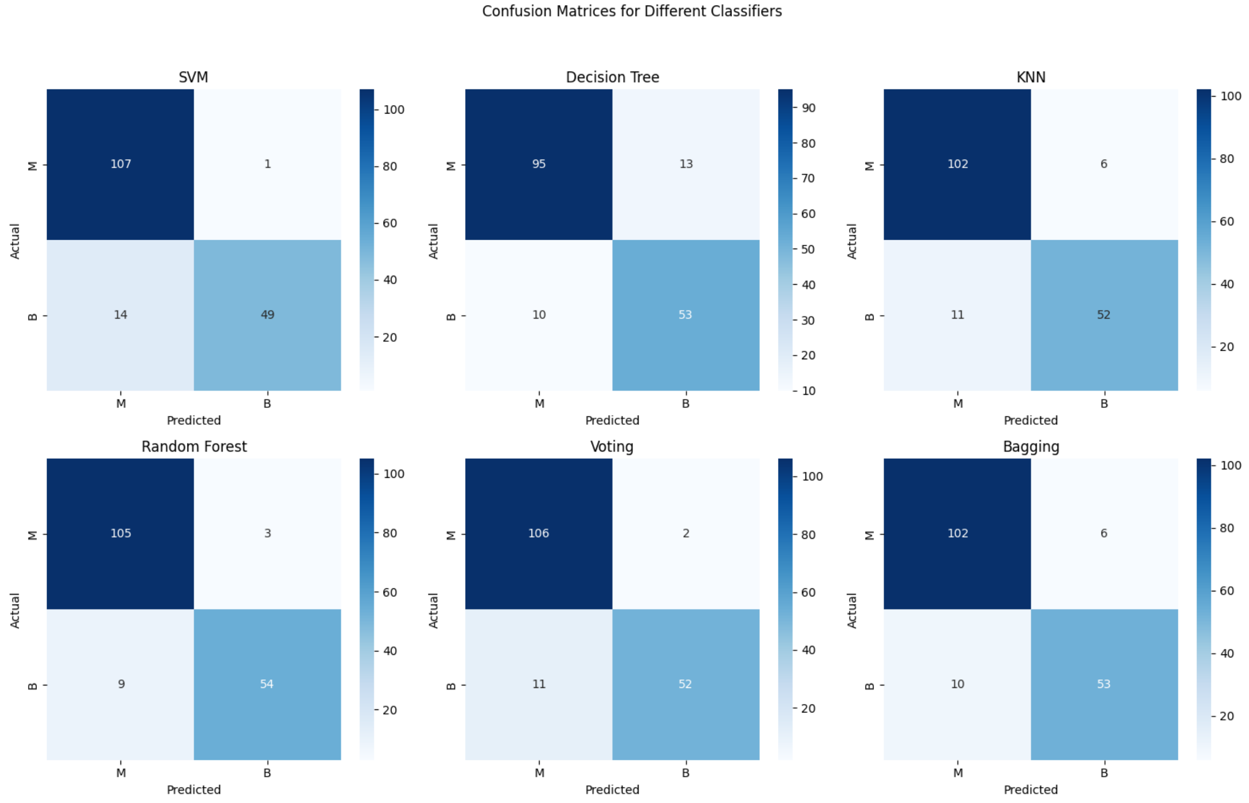


Figure 7. Confusion Matrix



4. Discussion

- When using a random forest to predict whether a cancer gets classified as malignant or benign, the most important variable to keep in mind is the radius, while the second most important is the perimeter. And its F1 score of 0.9 indicates that it is an extremely reliable algorithm to use for this scenario. When using a logistic regression to predict what a cancer will get classified as, the radius, perimeter, and area are all significant. But based on its pseudo R-squared value of 0.5923, this indicates that when compared to the random forest, a logistic regression is not as reliable an algorithm to use for this type of situation.
- Some of the other classification methods that were utilized, such as the voting classifier, KNN, and support vector machine are also not as reliable as a random forest for classifying cancers (based on their F1 scores being less than that of the random forest), but are still very good algorithms to consider. In addition, the decision tree algorithm is not the best choice for classifying cancers as malignant or benign, as indicated by how it had the lowest F1 score out of all the algorithms whose performances were measured using this calculation.
- When thinking about how the correlations between variables may influence cancer classifications, the ones that appear to be most significant are the ones between the radius and area of the cancer, as well as the perimeter and area of the cancer. Those variable combinations had a correlation coefficient of 0.99, indicating a strong positive correlation.

5. Conclusions

In this study, we aimed to identify which predictor variable was the most significant in the classification of malignant and benign breast cancers. We analyzed ten main predictor variables that included radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension of the cancer. We applied logistic regression, support vector machine (SVM), decision tree, random forest, and K-nearest neighbor (KNN), the five most commonly used supervised learning algorithms for binary classification. Furthermore, bagging and voting classifiers were applied.

Our results have shown that random forest, voting classifiers and bagging classifiers are the most effective in a descending order of classification models. Random forest has appeared to be the most effective in predicting if a cancer is benign or malignant. In addition to that, through our analysis comes the realization of the importance of some physical features that describe the tumors' shape and size in the process of diagnosis. Features of high relevance like the radius and perimeter emphasize a need for details of imaging and measurements necessary in early cancer detection and classifying.

The implications of this study are enormous, potentially guiding future diagnostic strategies and advancements of automated tools in early cancer detection. Future research may involve integrating these models into clinical decision support systems and their effectiveness research in operational healthcare data. Additional research on deep learning could also improve predictive accuracy and disrupt information generation systems from raw medical images that perceive subtle patterns.

6. Appendix

1. **Classifier Implementation:** We employed various classifiers, including SVM, decision tree, random forest, KNN, bagging, and voting classifiers. These were chosen for their prevalent use in binary classification tasks.
2. **Dataset:** The breast cancer dataset from the UCI Machine Learning Repository was used for this study.
3. **Feature Selection:** The first measurement for each variable in the dataset was considered for analysis to simplify the feature set.
4. **Performance Metrics:** F1 scores were used to evaluate the performance of each algorithm, considering the uneven distribution of labels in our dataset.
5. **Statistical Analysis:** We employed statistical models using libraries such as statsmodels in Python to analyze the data and draw conclusions.
6. **Data Preprocessing:** StandardScaler from scikit-learn was used for feature scaling, crucial for algorithms like SVM. Data was split into training and testing sets for model validation.

Author Contributions: Conceptualization, Data Preprocessing, Logistic Regression, Random Forest, Feature Importances In Random Forest Using MDI, F1 Score Plot: Christopher Hainzl; Boxplots, Correlation Matrix, Voting Classifier, Bagging Classifier, KNN: Muzamal Sheikh; Decision Tree, Support Vector Machine, and Violinplot: Moustafa Ayoub

References

1. Benign & Malignant Tumors: Orthopedics & Sports Medicine. Available online: https://health.uconn.edu/orthopedics-sports-medicine/conditions-and-treatments/a-z-index/benign-malignant-tumors/ (accessed on 12 Dec 2023).