Breast Cancer Risk Detection

Isha Dighe

Executive Summary

Breast cancer is the most common cancer diagnosed in women in the United States. The cancer develops as lumps containing cells that grow exponentially out of control and spread to other tissues and organs. Therefore, there is a high demand for efficient and accurate methods that can detect an early diagnosis and prognosis of breast cancer signs to facilitate clinical treatments for patients. In this paper, based on the current features of a patient, the project aims at analyzing the methods for predicting a case as malignant or benign. For this analysis, the project examines the Diagnostic Wisconsin Breast Cancer Database from the UCI Machine Learning Repository which contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Introduction and Hypothesis

In order to lower the risk of breast cancer, this project targets the results of a breast cancer screen of female patients in Wisconsin to diagnose their case as benign or malignant and treat the patient early before the arrival of noticeable symptoms. The objective of the analysis is to delineate the features that are most significant in detecting benign or malignant cancer and try several classification models to compare their results. The application of Machine Learning methods such as Decision Tree, Random Forest, Neural Networks, Logistic Regression, and Support Vector Machines provide advanced techniques for proficient decision-making regarding cancer outcomes. Besides the overall increase in the safety of public health, this process can help scientists in several ways: (i) the prediction of cancer susceptibility (risk assessment), (ii) the prediction of cancer recurrence/local control and (iii) the prediction of cancer survival.

Data Set Description

Attributes and domains are as follows:
1) ID number
2) Diagnosis (M = malignant, B = benign)
3-32)

Ten real-valued features are computed for each cell nucleus:
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

There is a mean, standard error, and “worst” or largest record for each of the features.

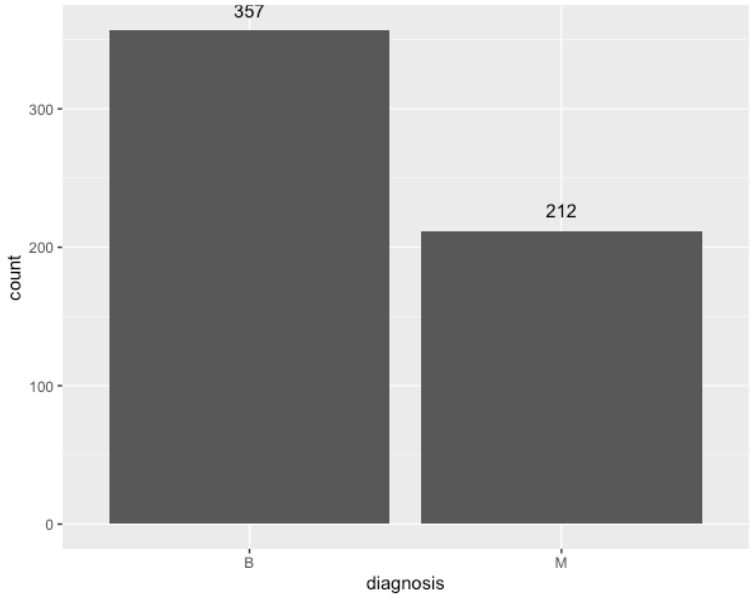
The goal of this project is to identify the appropriate classification techniques that can be used to determine if a tumor mass is benign or malignant, based on the current features and test results of the female patients. This form of data mining and analytics is very important in the healthcare field, as it strives to promote the prevention of disease and lowers the risk of late stage cancer through early detection.

Preprocessing

The first step of the preprocessing of the dataset was the removal of the ID column. It is not necessary for my analysis and therefore was removed. Next, I created a frequency distribution table to display the number of patients that are characterized as benign or malignant cases. Lastly, I explored the data for any missing or NULL values and adjusted the data accordingly.

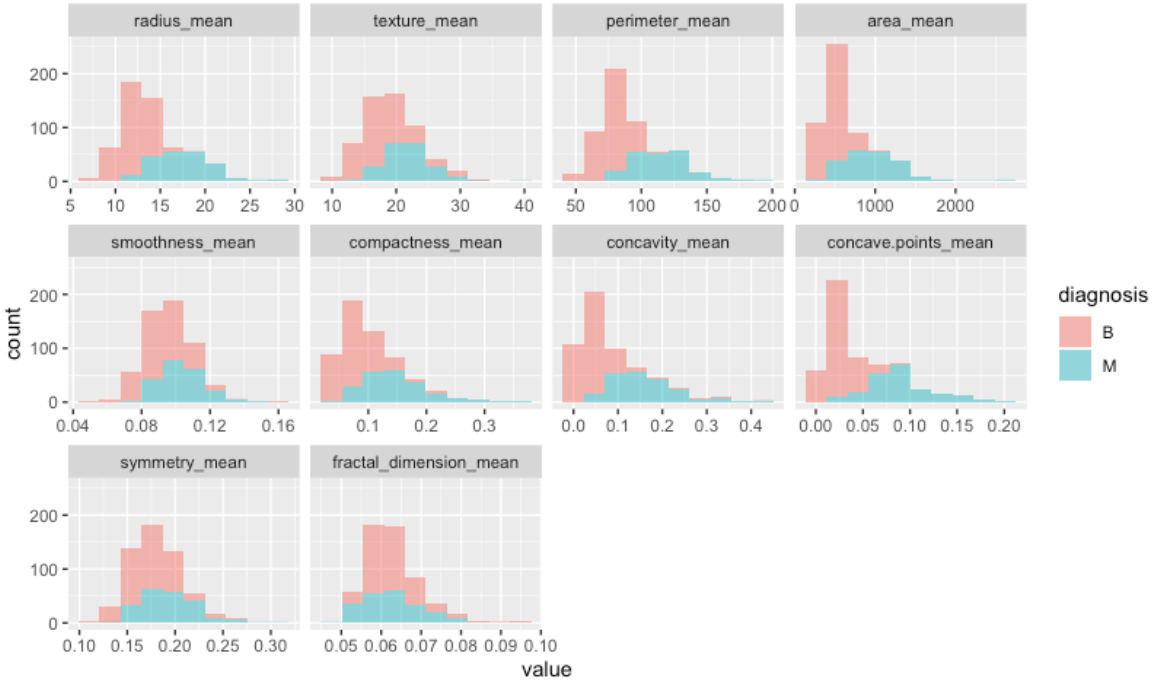
Exploratory Data Analysis

I began the EDA by displaying the distribution of Benign and Malignant cases in the dataset to determine if there was a visible class imbalance that would significantly affect my results though misleading accuracy metrics. Fortunately, although the amount of cases is not precisely equal, there isn’t a visible class imbalance in this dataset. Next, since there is a mean, standard

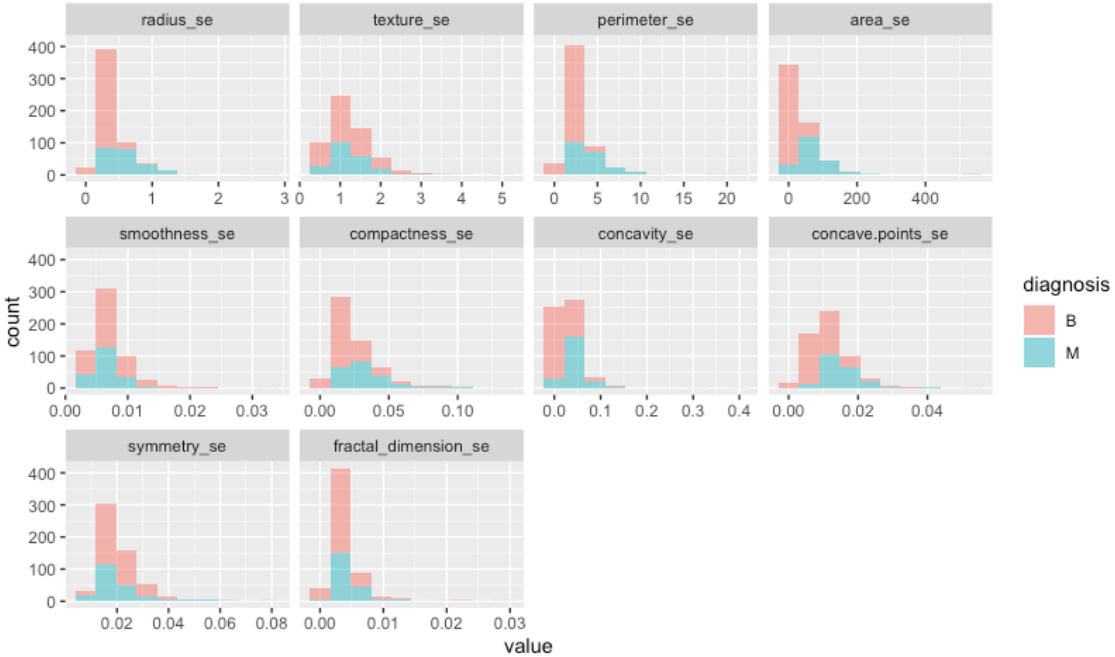


error and “worst” record for each feature, I plotted histograms to visualize the distribution of the data for each group.

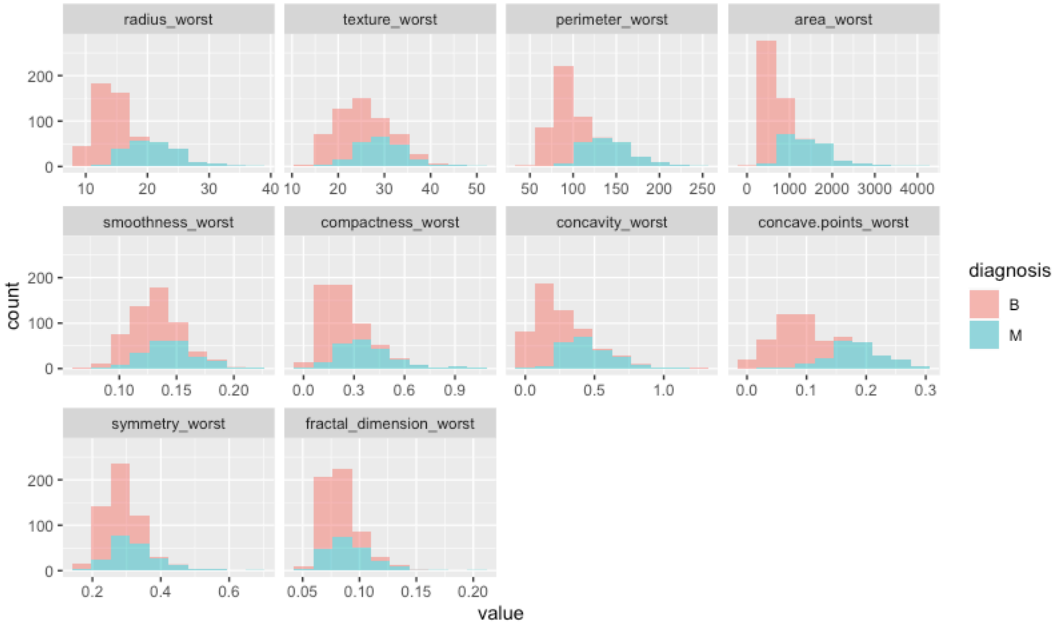
Mean Plot



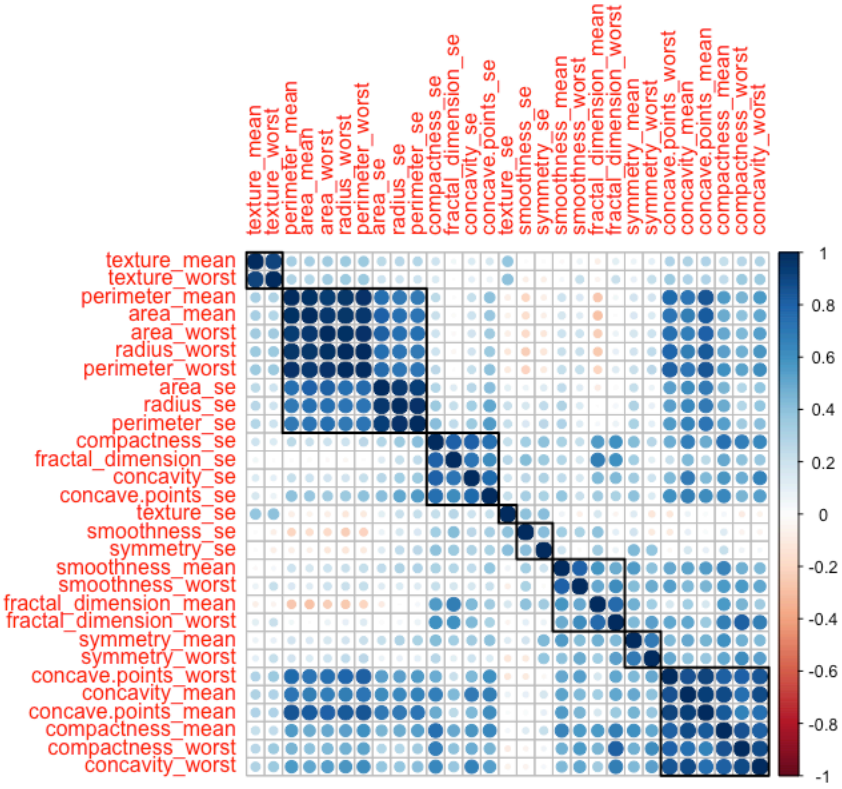
SE Plot



Worst Plot

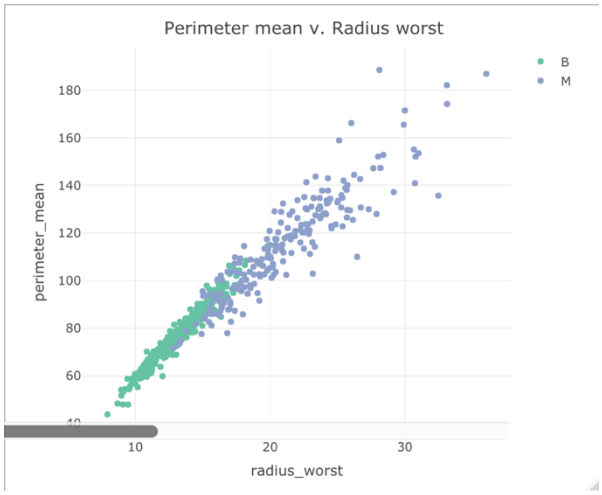


These histograms displayed that most of the features seem to be normally distributed. In particular, there seems to be a distinct difference in the benign and malignant values in the area and perimeter variables for mean, and the concave points, concavity, and perimeter for worst. These might prove to be significant variables for the determination of a case. Then, I conducted a correlation analysis and found that there is a great correlation between some variables.

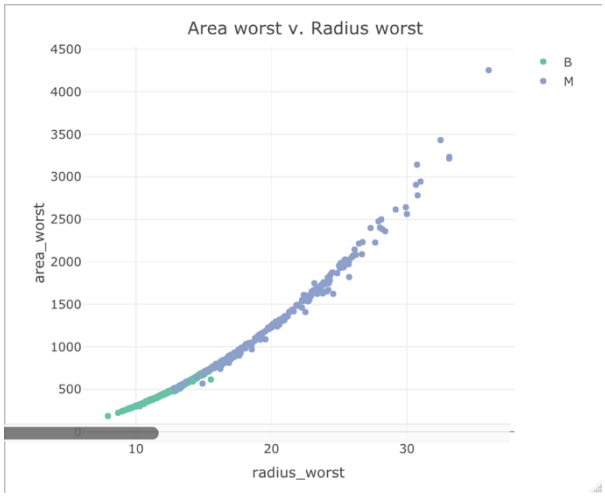


Preliminary Analysis revealed that there is high correlation between the variables:

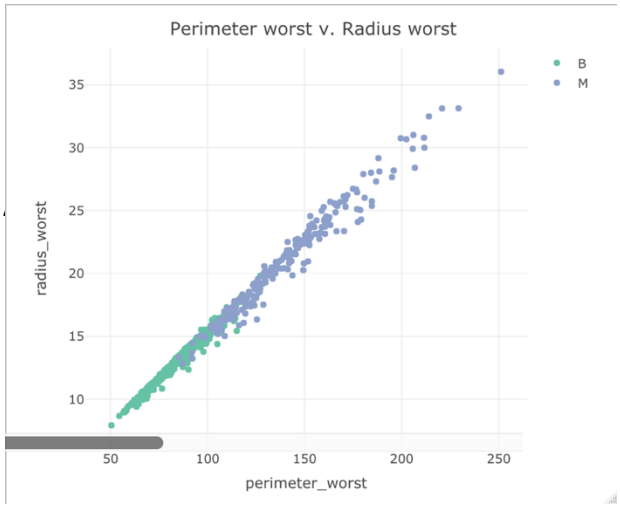
perimeter mean and radius worst



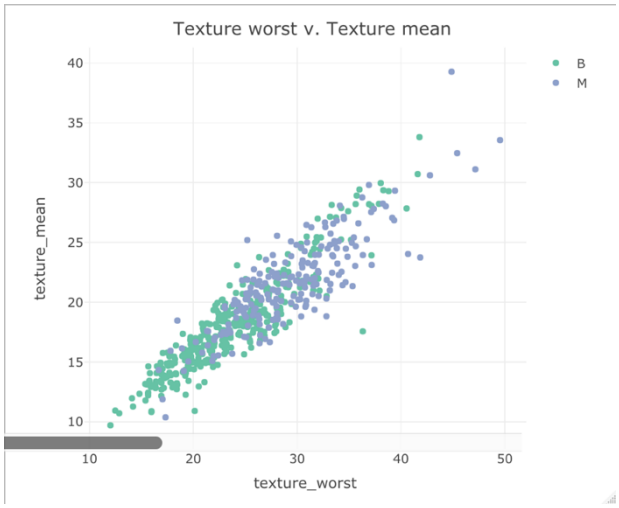
area worst and radius worst



perimeter worst and radius worst



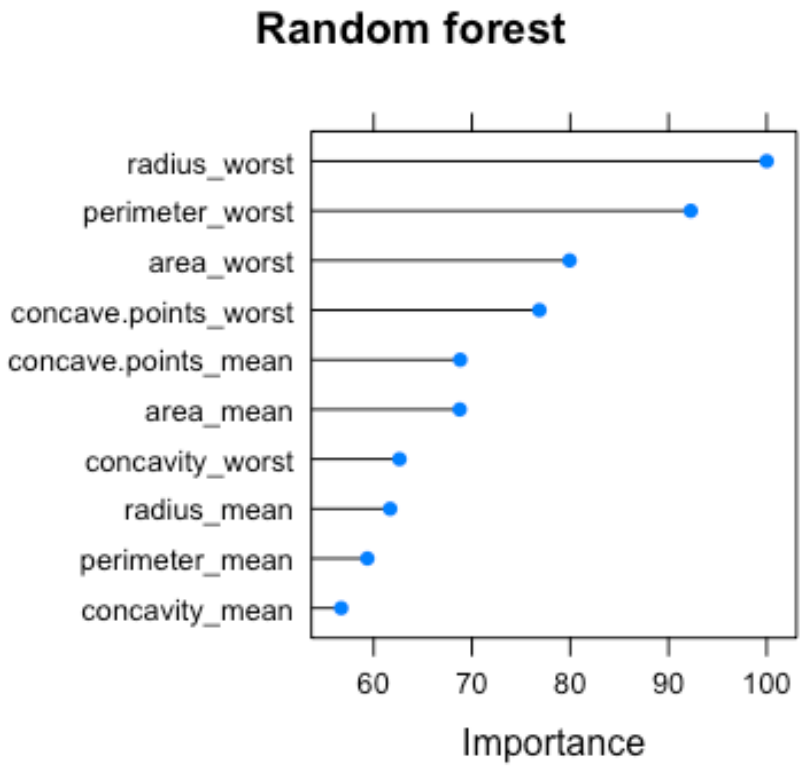
texture worst and texture mean



I chose to complete a Principal Component Analysis (PCA) as this is a high dimensional dataset with a large number of variables. The PCA method worked to reduce dimensionality and yield a more simple, workable version of the dataset. This made the dataset easier to explore and visualize. Using the PCA analysis, I observed the results for variables and conducted a correlation plot to display their relationships.

Empirical Analysis

I conducted several Machine Learning models to find the best fit. I used Naïve Bayes, Support Vector Machine, Support Vector Machine with Radial Kernel, Decision Tree, Classification Tree, Random Forest, Knn Classification, and Neural Network. The Random Forest was selected as the best predicted model amongst the rest. I tested several different sizes of Random Forests and achieved a final model with an accuracy of 98.2 %. Additionally, I used variable selection to find the most significant features for the classification. The plot and table confirm that radius_worst, perimeter_worst, and area_worst where the highest importance. The radius_mean, perimeter_mean, and concavity_mean had overall lower importance scores. It might be beneficial in the future to remove these features from the dataset prior to model fitting.



	Overall
radius_worst	18.765362
perimeter_worst	17.411560
area_worst	15.253013
concave.points_worst	14.714426
concave.points_mean	13.303263
area_mean	13.294743
concavity_worst	12.220881
radius_mean	12.053989
perimeter_mean	11.648319
concavity_mean	11.181897
area_se	10.982358
compactness_worst	6.470929
perimeter_se	5.782718
texture_worst	5.675148
radius_se	5.592951
compactness_mean	5.031151
texture_mean	5.017365
concave.points_se	3.657742
smoothness_worst	3.555415
symmetry_worst	3.393342
concavity_se	2.869040
fractal_dimension_worst	2.625872
smoothness_mean	2.314966
compactness_se	1.915146
fractal_dimension_mean	1.750169
symmetry_mean	1.542384
symmetry_se	1.383528
fractal_dimension_se	1.359212
texture_se	1.296499
smoothness_se	1.253737

Conclusion

Breast cancer research has made great advancements in the past decade through the application of data science and machine learning. The time and money that can be saved through efficient algorithms correctly identifying cancerous properties could revolutionize the medical field. For example, using my results, the Random Forest model yields approximately a 1.8% error rate for determining whether a new patient has a benign or malignant tumor. In the future however, this program could benefit from some fine-tuning regarding the elimination of redundant variables to further decrease the error rate for incorrectly classifying patients. Furthermore, the implementation of feature engineering could significantly benefit the model selection and optimization of the best parameters.

Sources

Dataset Webpage
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Journal Article
T. Ayer, O. Alagoz, J. Chhatwal, J.W. Shavlik, C.E. Kahn, E.S. Burnside
Breast cancer risk estimation with artificial neural networks revisited
Cancer, 116 (2010), pp. 3310-3321

R Studio Resource Webpage
Kuhn, Max. “Caret v6.0-86.” Caret Package | R Documentation, www.rdocumentation.org/packages/caret/versions/6.0-86.

Journal Article
Ming, C., Viassolo, V., Probst-Hensch, N. et al. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res 21, 75 (2019). https://doi.org/10.1186/s13058-019-1158-4

Journal Article
Kourou, Konstantina, et al. “Machine Learning Applications in Cancer Prognosis and Prediction.” Computational and Structural Biotechnology Journal, Elsevier, 15 Nov. 2014, www.sciencedirect.com/science/article/pii/S2001037014000464.