A Comparison of the Ability of Classification Models to Predicting whether the Breast Cancer type is Malignant or Benign

Presented by:

Alyaqadhan Alfahdi

Arya Amarnath

Professor - Dr. Tianjian Zhou

STAT 460 - Applied Multivariate Analysis

Colorado State University

May 3, 2024

GitHub Repo

Table of Contents

Abstract

The classification of breast cancer tumors into malignant or benign types is crucial for early intervention and treatment strategies. This study compares several machine learning models based on the Breast Cancer Wisconsin (Diagnostic) Dataset to determine the most effective classifier. This project includes extensive data preprocessing, exploratory data analysis, and rigorous model evaluation including logistic regression, random forest, k-nearest neighbors, and support vector machines utilizing grid-search and cross-validation to acquire specialized hyperparameters. This allowed us to identify the logistic regression model with L1 regularization as the most proficient in predicting breast cancer malignancy with high sensitivity and balanced accuracy.

1. Introduction

Breast cancer, characterized by the uncontrolled growth of breast cells, stands as one of the most common cancers worldwide significantly impacting women's health globally. Early and accurate diagnosis of breast cancer types allows doctors to distinguish between malignant (cancerous) and benign (non-cancerous) tumors which is pivotal to addressing the treatment of their patient's cancer type. It not only enhances the treatment's success but significantly improves survival rates. Traditional diagnostic methods rely on physical examinations, mammography, and biopsy results, sometimes leading to ambiguous or delayed results. In recent years, machine learning (ML) has emerged as an amazing tool in medical diagnostics to help shape the industry, augmenting traditional methodologies with higher accuracy and efficiency. By leveraging historical data and advanced analytical techniques, ML models can learn to identify patterns and anomalies that may not be immediately apparent to human observers.

This study employs a breast cancer dataset to explore the effectiveness of various ML models in classifying breast tumors. The models tested include logistic regression, support vector machines, k-nearest neighbors, and random forests. Each model's capability to accurately predict tumor type is critically important due to the cost of misdiagnosis which can lead to unnecessary or missed treatment opportunities. Our analysis aims to evaluate each model's performance using statistical metrics such as balanced accuracy, sensitivity, and Type II error. Through this comparative study, we endeavor to identify the most reliable model for breast cancer classification, hopefully contributing to more meaningful, accurate, and timely diagnoses in medical practice.

2. Data and EDA

The primary dataset used for this research project is the Breast Cancer Wisconsin (Diagnostic) Dataset, sourced from Kaggle (H.M. Yasser, 2022). This dataset is sourced utilizing fine-needle aspiration (FNA) which utilizes cell samples from tumors in a patient's body. (Citation, Date) With the help of this technology, we are provided a dataset of 570 entries, each representing an individual case of breast cancer. Each entry includes a diagnosis of being malignant (M) or benign (B) and ten real-valued features calculated for each cell nucleus. Some of these features include the radius, perimeter, area smoothness, and concavity of the cell which all provide important information to classify cancerous or non-cancerous tumors. Each feature is then further split into three statistics: the mean, standard error, and the worst (mean of the three largest values) helping to provide a comprehensive view of each tumor's characteristics.

The first step in our EDA was to examine the distribution of the target variable, which categorizes tumors as either malignant (M) or benign (B). Understanding the balance between these categories is essential because an imbalanced dataset could lead to models that are biased toward the majority class. In our dataset, we observed that benign cases were more prevalent than malignant ones, which guided our decision to employ techniques such as stratified sampling during the train-test split to ensure that both categories are adequately represented.

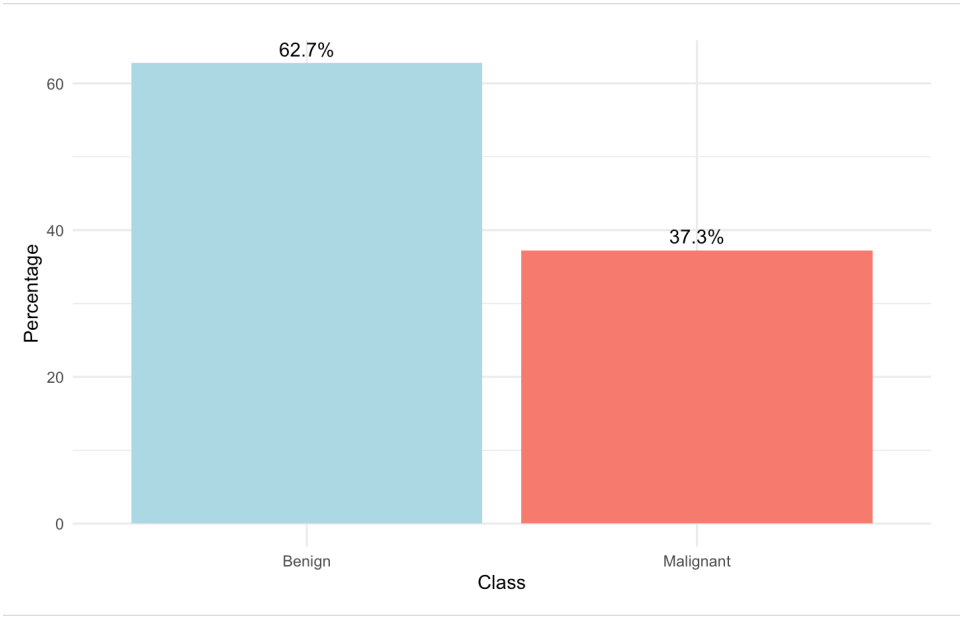


Figure 1.

A key component of our EDA was creating a correlation matrix between our features allowing us to identify any linear relationships between variables. A high correlation between variables can lead to multicollinearity which can cause our models to overfit on our data and make it harder for our model to interpret the features. We visualized these correlations using a heatmap, which provided a clear view of how features are related. For example, features related to the size of the nucleus were highly correlated which can be expected as they are mathematically related, for example, the perimeter and area of a tumor would naturally increase with its radius. After examining the matrix and identifying features with high correlations, we decided to employ Principal Component Analysis (PCA) to reduce dimensionality and mitigate multicollinearity.

3. Principal Component Analysis

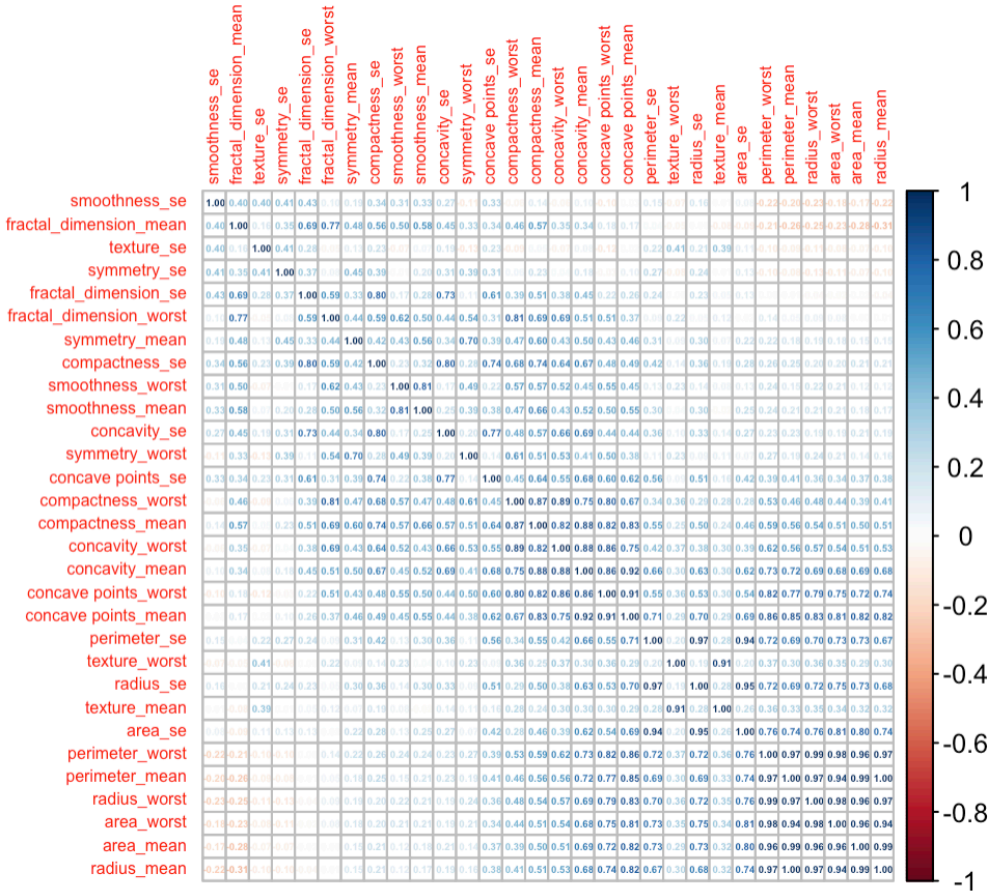


Figure 2.

Given the high dimensionality of our dataset (32 features), we employed Principal Component Analysis (PCA) to reduce the number of variables. PCA transforms the original correlated features into a new set of uncorrelated variables called principal components which arrange linear combinations of the original features. The following 'Scree plot' helps to visualize the 'explained variance' of each principle component and shows us that the first 3 components contain 70% of the explained variance and as the number of components increases, the explained variance drops.

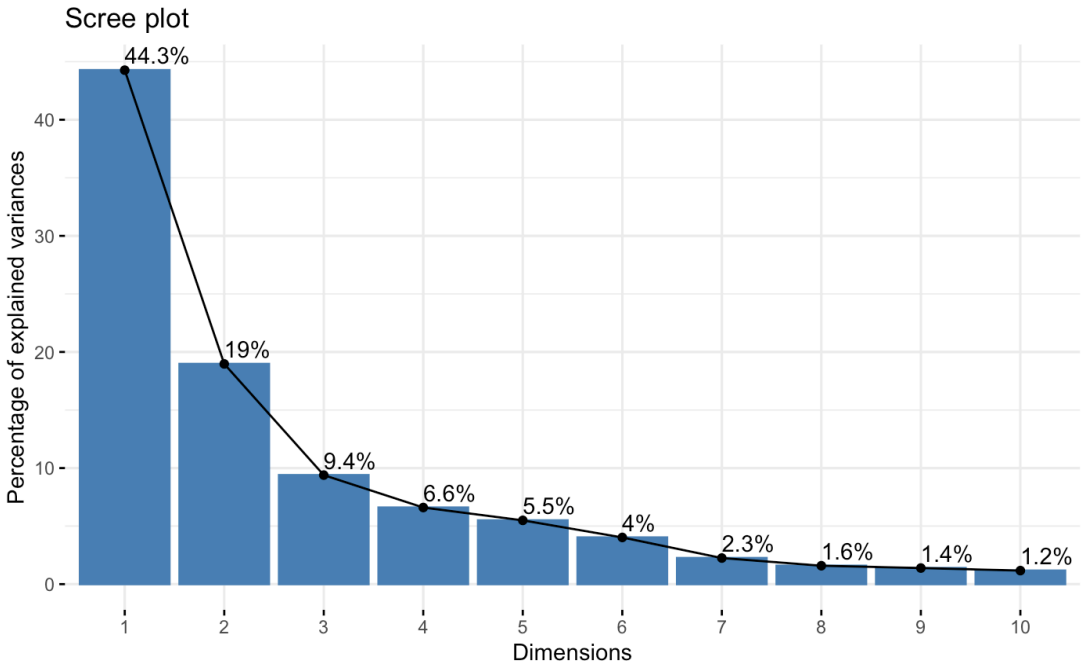


Figure 3.

To find our selection of principal components, we retained the number of PCs that explained approximately 95% of the total variance in the dataset. This reduction simplifies the modeling process and helps visualize the data in a lower-dimensional space. The results of this cumulative variance test are shown below with ten Principal Components reaching our 95% requirement, allowing us to utilize these ten PCs within our model creation stage.

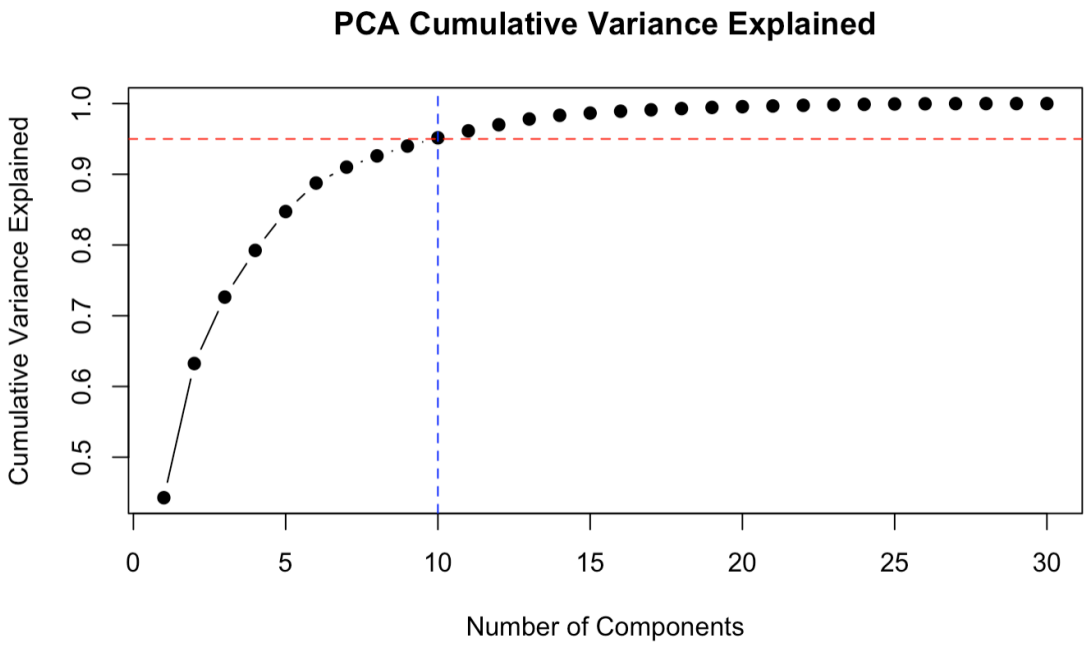


Figure 4.

4. Models

To achieve our goal of creating a predictive model for breast cancer classification, we followed a specific methodology for each algorithm we applied. Our process can be outlined in the following stages:

Data Splitting: With the dimensionality reduced to the ten most significant principal components, we divided the dataset into training and testing subsets. 30 percent of the data was the test set used to evaluate the model, while the remaining 70 percent was the training set.

Model Training and Hyperparameter Tuning: Employing cross-validation with 7 folds, we searched through various combinations of parameters using a grid search method. Our objective was to optimize the model's performance.

Evaluation Metrics: After training and optimizing our models, we evaluated their performance using the test set. We focused on several key metrics created to address the challenges posed by

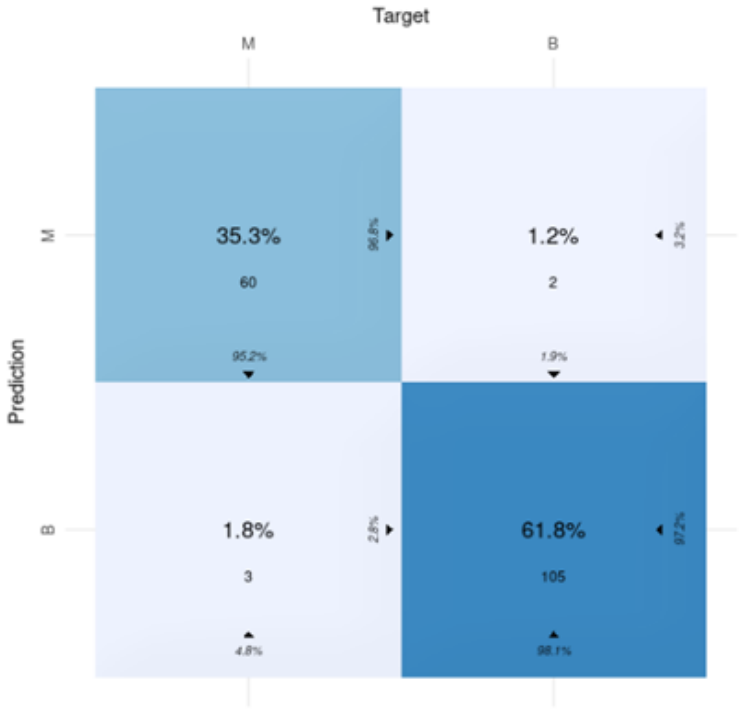
unbalanced class distributions. Sensitivity measures the model's ability to correctly identify all types of tumors within the dataset. Additionally, we examined the model's performance in minimizing Type II errors, which occur when the model incorrectly predicts a malignant tumor as benign. This metric is particularly important in medical contexts where false negatives can have significant effects on future treatments and patients. Furthermore, we evaluated the balanced accuracy, which considers the model's ability to perform well across all classes, accounting for class imbalance. Lastly, we quantified the model's classification ability using the ROC-AUC score, which provides a comprehensive measure of how well the model differs between benign and malignant cases across various decision thresholds. These metrics collectively provide a deeper analysis of our models' effectiveness in cancer diagnosis.

4.1 Logistic Regression

Logistic regression is a method used to classify data into different groups. It works under the assumption that the groups can be separated by a straight line, or are very close to being separable in this way. The main goal of logistic regression is to find a line or plane that does the best job at dividing the data into its respective categories, such as positive or negative groups. It's quite versatile since it can handle problems where we have more than two groups to classify (P, 2018). We added L1 regularization (Lasso) and L2 regularization (Ridge) to the model to help with robustness and overall accuracy.

L1 regularization adds a penalty equivalent to the absolute value of the magnitude of coefficients. This method is useful in feature selection since it effectively reduces the number of features by driving the coefficients of less important features to zero. In the context of our breast cancer dataset, where some features might be less predictive of the outcome, Lasso

regularization helps by retaining only the most significant features, thereby simplifying the model. Unlike Lasso, Ridge regularization adds a penalty equivalent to the square of the magnitude of coefficients. This approach does not reduce the coefficients to zero but instead reduces their impact allowing the handling of multicollinearity among features. By penalizing the size of the coefficients, Ridge regularization smoothens the learned weights across all features thus distributing the importance more evenly and ensuring that no single feature dominates the prediction logic excessively. By combining these two regularization methods, our logistic regression model can understand nuances of high-dimensional data, ensuring that the predictions remain consistent across different samples of the total population.

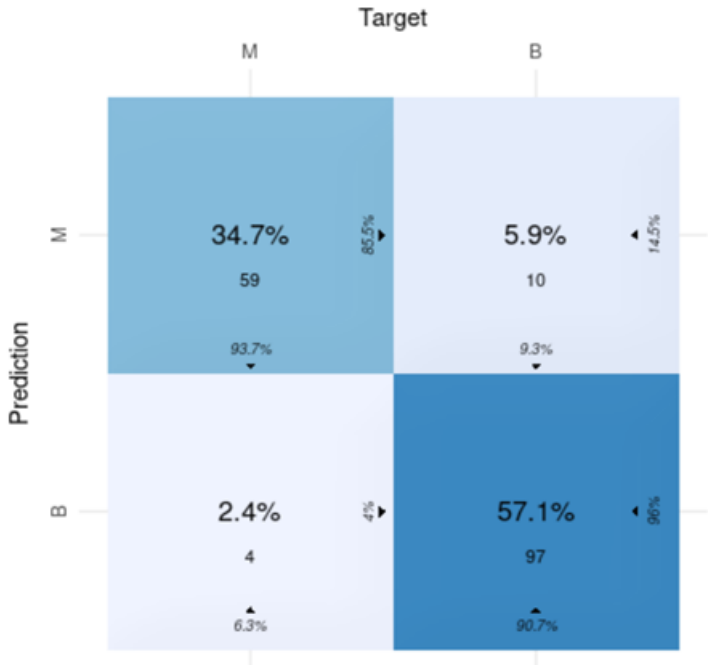


The model, which used alpha = 1 for L1 regularization and a lambda value of 0.01, achieved notable performance in breast cancer classification, as assessed by the specified evaluation metrics. With a sensitivity of 95.24%, the model demonstrated a high rate of correctly

identified malignant cases, crucial for effective medical diagnosis. The rate of type II error, at 4.8%, indicates the model's ability to minimize false negatives, particularly relevant in medical contexts. Moreover, the balanced accuracy score of 0.9668 suggests the model's consistent performance across all classes, addressing the challenge of class imbalance. Additionally, the ROC-AUC score of 0.97 (Figure 5 in the appendix) signifies the model's strong discriminative ability between benign and malignant cases across various thresholds.

4.2 Random Forest (RF)

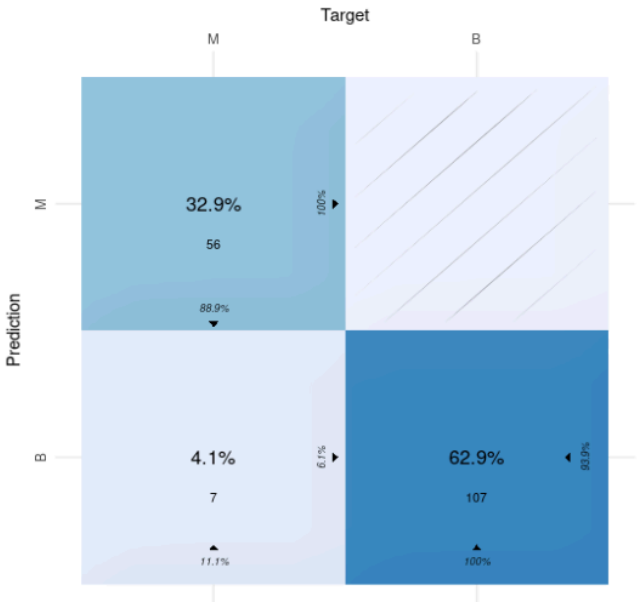
Random Forest is a strong machine learning method that creates a 'forest' filled with many decision trees. Each tree in the forest is formed using a random set of data points and features, and it provides its prediction. The overall result of the Random Forest model is derived by averaging the predictions from all these trees. This method generally results in a more accurate and reliable outcome than what we would get from just one tree (IBM, n.d).



In our Random Forest analysis with ‘mtry’ set to 10, the model showed less significant efficacy in classifying breast cancer types, as indicated by the specified evaluation metrics. With a sensitivity of 93.65%. The type error II rate, at 6.3%, underscores the model's effectiveness in minimizing false negatives. Additionally, the balanced accuracy score of 0.92 suggests the model's robust performance across all classes, considering class imbalance. Furthermore, the ROC-AUC score of 0.908 (Figure 6 in the appendix).

4.3 K-Nearest Neighbors (KNN)

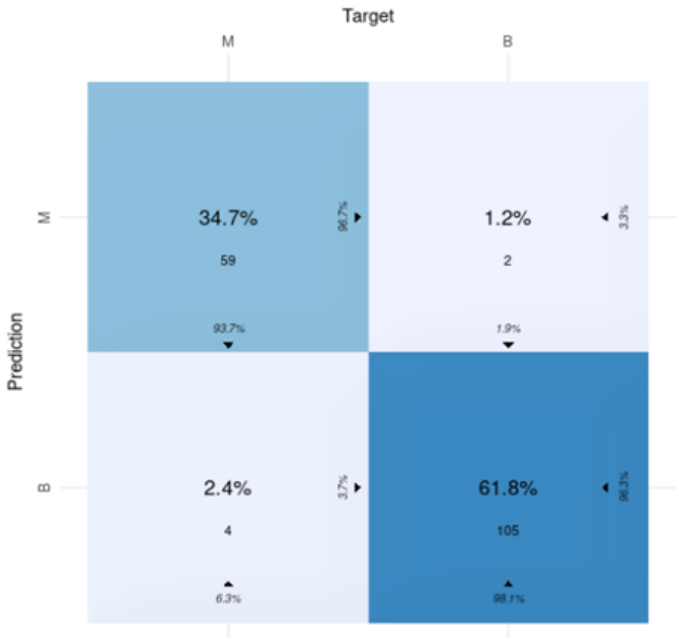
K-Nearest Neighbors (KNN) is a straightforward machine learning technique that determines the outcome for a new data point by looking at the most frequent outcomes among its closest neighbors in the dataset. To predict a new point, KNN examines the 'C' nearest neighbors from the data it has learned from. In classification tasks, the final prediction is usually the most common class among these neighbors (Singh, 2024).



In the K-Nearest Neighbors (KNN) analysis with k=17, the model showed less efficacy in classifying breast cancer types, as evidenced by the specified evaluation metrics. With a sensitivity of 88.89%. The type error II rate, at 11.1%, indicates less of the model's effectiveness in minimizing false negatives. Additionally, the balanced accuracy score of 0.94 suggests the model's robust performance across all classes, despite class imbalance. Moreover, the model achieved a precision rate of 100%, highlighting its accuracy in predicting malignant conditions and effectively reducing false positives. Lastly, the ROC-AUC score of 0.969 (Figure 7 in the appendix).

4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a type of supervised machine learning algorithm that is generally used for classification, although it can also handle regression tasks. Primarily, SVM is more effective for classification purposes. The core goal of the SVM algorithm is to identify the best possible hyperplane in an N-dimensional space that distinctly separates the data points into different classes within the feature space. This hyperplane is chosen such that it maximizes the distance, or margin, between the nearest points of the different classes, ensuring that the separation between them is as wide as possible (GfG, 2023).



Support Vector Machine (SVM) with C set to 0.1, the model showed excellent performance in classifying breast cancer types, as indicated by the specified evaluation metrics. With a sensitivity of 93.65%, the model demonstrated a robust ability to correctly identify malignant cases. The type error II rate, at 6.3%, underscores the model's effectiveness in minimizing false negatives. Moreover, the balanced accuracy score of 0.95 suggests the model's consistent performance across all classes, addressing class imbalance issues. ROC-AUC score of 0.951 (Figure 8 in the appendix)

5. Model Comparison

Table 1: Comparison of Model Performances for Breast Cancer Classification

	Performance Metrics				
Model	**Sensitivity**	**Type II Error**	**ROC-AUC**	**Accuracy**	**Balanced Accuracy**
Logistic w/ L1	0.9524	0.0476	0.970	0.9705	0.9668
Random Forest	0.9365	0.0635	0.908	0.9176	0.9215
K-NN	0.8889	0.1111	0.969	0.9588	0.9444
SVM	0.9365	0.0635	0.965	0.9647	0.9589

The comparison of models for telling apart the serious kind of breast cancer (malignant) from the not serious kind (benign) shows that some models are more accurate than others. The Logistic Model with L1 and L2 (Lasso/Ridge Regression) stands out with a 95.24% success rate in correctly identifying serious cases. It is sharp at this, likely thanks to Lasso and Ridge Regression which work together to figure out complex structures in the data without overfitting and negatively impacting accuracy. It also has a tiny chance, only 4.76%, of missing a malignant case, which is important for making sure people get the correct treatment they need. With an impressive ROC-AUC score of 0.970, it is also great at telling the difference between benign and malignant cases under various scenarios. All these strong numbers suggest that the Logistic Model with L1/L2 Regularization is the best of our attempted models for reliable breast cancer classification, making it an excellent choice for doctors to use.

6. Conclusion

Analyzing historical data from breast cancer cases provides a strong basis for understanding and predicting whether the disease is malignant or benign. Principal Component Analysis (PCA) proved useful in simplifying the data, reducing unnecessary details while preserving most of the original information within 10 main components. The logistic regression model with L1/L2 regularization was chosen as the top choice in our study due to its high accuracy, achieving a sensitivity of 95.24% and a balanced accuracy of 96.68%, effectively distinguishing between cancer types. Looking ahead, building upon the success of this model with expanded datasets and including genetic and lifestyle factors could improve the accuracy of cancer predictions and offer a deeper insight into individual risk factors. Utilizing more complex models such as RNN/CNNs which utilize multiple hidden layers and extra computational steps to uncover even more insights into larger, more complicated datasets.

References

P, A. (2018, December 12). L1 and L2 Regularization. - Aditya .P - Medium. *Medium*. https://medium.com/@aditya97p/l1-and-l2-regularization-237438a9caa6

What is Random Forest? | IBM. (n.d.). https://www.ibm.com/topics/random-forest

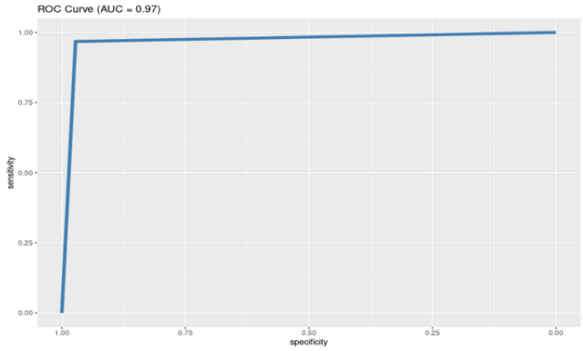
Singh, A. (2024, February 13). *KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

GfG. (2023, June 10). *Support Vector Machine (SVM) algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/

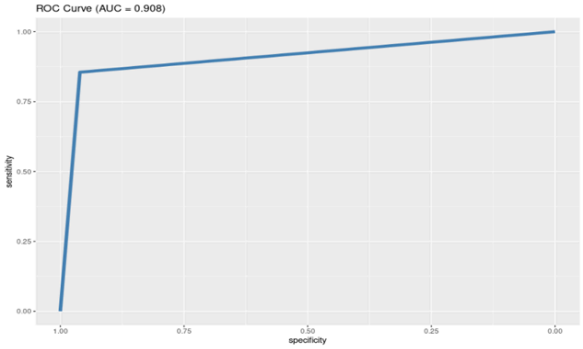
H, M Yasser. "Breast Cancer Dataset." Kaggle, updated 2 years ago, https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

Appendix

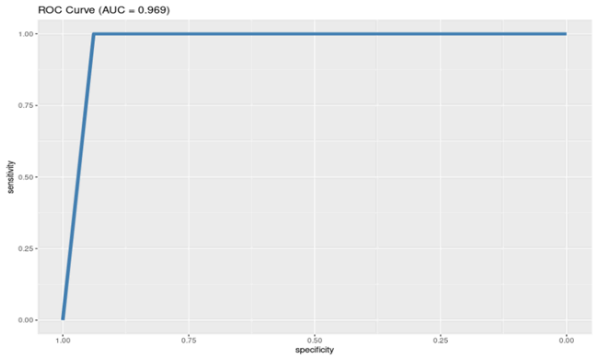
(Figure 5)



(Figure 6)



(Figure **7)**



(Figure **8)**