Predictive Breast Cancer Classification in the U.S.

Eli Padilla, Praharshith Jamalapuram, Zhaoqi Liang
Industrial & Systems Engineering, University of Wisconsin - Madison
December 2024

1. Introduction

Breast cancer is a critical health concern, with the American Cancer Society projecting approximately 310,000 new diagnoses in women across the United States in 2024 (“Breast Cancer Statistics: How Common Is Breast Cancer?”). Unfortunately, about 42,000 women are expected to succumb to the disease this year. Early and accurate diagnosis is very important in determining the most effective treatment plans and improving patient outcomes. Critical to this process is the classification of breast masses as benign or malignant, which guides the urgency and type of treatments required. In this project, we look to leverage machine learning to develop a predictive model for breast cancer classification, with the goal of improving diagnostic accuracy and supporting clinical decision-making in the fight against this common disease.

2. Data

The dataset we decided to use to build our model is the “Breast Cancer Wisconsin (Diagnostic) Data Set” from Kaggle. The data set contains 30 diagnostic features from 569 patients with breast cancer, including 212 malignant cases and 357 benign cases. These 30 features are based on the cell nuclei present in breast cancer biopsies, along with the target class (malignant or benign). These features describe various characteristics of the cell nuclei, including radius, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. We can use these features to identify whether the patient's breast cancer is malignant or benign.

Dataset Link: Kaggle Link

3. Methods

3.1 Exploratory Data Analysis (EDA)

We conducted an Exploratory Data Analysis (EDA) to understand the characteristics of the dataset and identify key patterns and relationships between the features and the target variable. The EDA process focused on three main steps: looking at the distribution of the target variable, checking feature correlations, and exploring the relationship between individual features and the target.

For target variable distribution, we plotted a bar chart to visualize the distribution of the diagnosis, where 1 represents malignant tumors and 0 represents benign tumors. This step helped us see how the two classes are distributed and whether there is a class imbalance, which is important because imbalanced data can affect the performance of classification models. The second aspect of the EDA is feature correlation. To understand how features relate to each other and to the target variable, we calculated the Pearson correlation coefficients. We used a heatmap to show these correlations for a subset of the mean feature values, such as radius_mean, perimeter_mean, and area_mean. This helped us identify which features are highly correlated with the target and which ones might be redundant. For example, radius_mean and perimeter_mean showed very high correlations with the target variable, suggesting they are important for prediction. In the final part of the EDA, we also examined the relationship between individual features and the target variable. For example, we made perimeter_mean against diagnosis using a swarm plot. This helps us to see how the values of this feature are distributed for benign and malignant cases. Features like perimeter_mean clearly showed differences between the two classes, which means they might be useful for classification.

3.2 Binary Logistic Regression

Upon completion of the EDA, we saw 8 key, uncorrelated features that we knew we wanted to include in the formulation of our logistic regression: radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, and symmetry_mean. Using these features, we utilized Python to complete a binary logistic regression. The first step of this analysis was to set the target column as a binary variable (1 for Malignant (M), and 0 for Benign (B)). Then, in order to prepare the features for regression analysis and create a standard scale of reference, we used the StandardScalar() function. Once the features were scaled to the unit variance, we created a test and training set split with 20% of the observations being used for testing. Finally, the LogisticRegression() function was implemented on the training data and L2 regularization was used in order to prevent overfitting. From there, the logistic regression was complete and the results were ready to be analyzed.

3.3 K-Nearest Neighbors (KNN)

We chose KNN model as an accuracy priority model with due credits to its ease of implementation and scalability. Since the chosen dataset is not huge in terms of observations, KNN makes one of the ideal choices for such non-linear data. KNN method helped identify the key features which are closely related to the target. Therefore, in extension, this enabled more crosscheck ability for feature selection among other models.

The dataset was first run through a K-Means Clustering model with cluster number ranging from 1 to 40. This was done to determine the optimal number of clusters for further study. An Elbow plot was plotted against the euclidean distances & number of clusters to determine the

appropriate complexity and cluster count for the KNN model. This was then implemented into a 10-Fold cross validated KNN model, cross-validation was performed in order to minimize overfitting and promote better generalization. A graph was plotted with RMSE against the K number in order to determine the least error-ridden K number. Utilizing GridSearchCV function, we were able to determine & display the best parameters, best cross-validated accuracy & the test accuracy of the best model of the KNN Model. Upon the determination of the best model and the respective parameters, the Permutation Importance function was utilized to determine the key features that have a significant effect on the model's performance. This helped in fine tuning parameters for future studies.

3.4 Random Forest (RF)

We utilized a Random Forest Classifier in this project due to its strong predictive performance and ability to handle high-dimensional data while minimizing the risk of overfitting. Random Forest was particularly well-suited for this dataset because of its robustness in modeling non-linear relationships and its capability to provide feature importance scores, which helped in identifying the most relevant features for the classification task. The dataset was split into training and testing sets, with 80% of the data used for training and 20% reserved for testing. A 10-fold cross-validation was applied to the training set to ensure the model generalized well and avoided overfitting. During training, key hyperparameters of the Random Forest Classifier were tuned to optimize its performance. These hyperparameters included the number of estimators (n_estimators), the maximum depth of the tree (max_depth), the minimum samples required to split an internal node (min_samples_split), and the minimum samples required for a leaf node (min_samples_leaf). Tuning these parameters could prevent overfitting or underfitting and was essential to balance the trade-off between model complexity, computational efficiency, and prediction accuracy. The best combination of hyperparameters was determined using grid search with cross-validation, optimizing for accuracy as the primary metric.

After training the model, feature importance was assessed using the impurity-based importance metric provided by the Random Forest algorithm. This metric evaluates the reduction in impurity contributed by each feature across all decision trees in the forest. The resulting feature importance scores were used to rank features, with concave points_mean, radius_mean, and perimeter_mean emerging as the most significant predictors. These features provided valuable insights into the key characteristics associated with malignant and benign tumors.
The performance of the final Random Forest model was evaluated on the test set using both accuracy and recall. Accuracy measured the overall correctness of the predictions, while recall focused on the model's ability to identify malignant tumors, which was critical for minimizing false negatives in this medical application. Visualizations, including a bar plot of feature importances and a confusion matrix of test predictions, were included to support the analysis and interpretation of the results.

4. Results

Exploratory Data Analysis

Figure 1 (a) illustrates the distribution of the "diagnosis" variable in the breast cancer dataset, where "0" represents benign tumors and "1" represents malignant tumors. The chart reveals a class imbalance, with significantly more benign cases (around 350) compared to malignant cases (approximately 200). This imbalance suggests that benign tumors are more prevalent in the dataset, with benign cases accounting for roughly 65% of the total. This imbalance has important implications for model development. If not addressed, a machine learning model might become biased toward predicting the majority class (benign), potentially leading to high accuracy but poor performance in identifying malignant cases. To mitigate this, techniques like oversampling the minority class, undersampling the majority class, or using class weights in the model can be employed. Additionally, recall and precision metrics should be emphasized during evaluation to ensure the model effectively detects malignant cases, which are of higher clinical importance.

The heatmap in figure 1 (b) visualizes the correlation matrix of the mean features in the breast cancer dataset, showing the strength and direction of relationships between features and with the target variable, "diagnosis". The correlation values range from -1 to 1, with positive values (red hues) indicating that two variables tend to increase together, and negative values (blue hues) suggesting that as one variable increases, the other decreases. The stronger the correlation, the more closely the variables are related. Diagnosis exhibits strong positive correlations with features such as "concave points_mean" (0.78), "perimeter_mean" (0.74), and "radius_mean" (0.73), highlighting their potential as key predictors for distinguishing between malignant and benign tumors. In contrast, features like "smoothness_mean" (0.36) and "fractal_dimension_mean" (-0.013) show weak correlations with the target variable, indicating their limited importance in classification. Additionally, some features demonstrate very high inter-correlations, such as "radius_mean" and "perimeter_mean" (almost 1.0), as well as `concave points_mean` and "concavity_mean" (0.92). This suggests that certain features provide overlapping information, which might lead to redundancy in the dataset. Such redundancy could be reduced through dimensionality reduction or feature selection techniques, focusing on those variables most strongly correlated with "diagnosis". The heatmap thus serves as a crucial tool for understanding the relationships within the data and identifying the most relevant features for predictive modeling.

This scatter plot (Figure 2) illustrates the relationship between the mean concave points ("concave points_mean") and the "diagnosis" variable, where "0" represents benign tumors and "1" represents malignant tumors. The distribution of points shows a clear separation between the two diagnostic categories based on `concave points_mean`. Malignant cases (1) generally have significantly higher "concave points_mean" values compared to benign cases (0). For benign

tumors (0), most values of “concave points_mean” are concentrated near 0.025, with relatively little variation. In contrast, malignant tumors (1) exhibit a wider range of values, with many cases exceeding 0.10. This suggests that the mean number of concave points is a key distinguishing feature between the two tumor types, with malignant tumors exhibiting more pronounced concavity on their contours. The plot highlights the strong predictive potential of “concave points_mean” for tumor classification. The distinct separation between the two groups supports its use as an important feature in classification models aimed at distinguishing between benign and malignant tumors.

Binary Logistic Regression

The results in Figure 3 show a promising model with an accuracy of 95.6% and a recall of 93%. These metrics indicate that the model correctly identifies a large majority of cases, particularly succeeding at identifying malignant cases, as reflected in the high recall. Recall, also known as sensitivity, is important in medical diagnostics because it measures the regression model’s ability to correctly identify true positive cases of malignant breast cancer, minimizing the likelihood of false negatives. The confusion matrix further highlights this by showing that while false negatives (misclassifications of malignant cases as benign) are present, their occurrence is relatively low, highlighting the reliability of the model in prioritizing accurate medical diagnostics.

Upon training the model, we also analyzed its coefficients to gain insights into feature importance. Among the 8 key features evaluated, area_mean was found to be the most significant predictor, with the highest positive coefficient value. This finding suggests that the average size of the tumor area plays a key role in distinguishing malignant from benign cases in this dataset. This aligns with both intuition and existing medical knowledge, where larger tumor sizes are often associated with malignancy. The emphasis on area_mean reinforces the model's interpretability, offering a meaningful link between machine learning outputs and medical knowledge. This not only validates the model's predictive power, but also provides doctors with the insights to prioritize feature-based analysis in diagnostic processes.

K-Nearest Neighbors (KNN)

Starting with the model’s accuracy and recall rate, the scores (interpreted from Figure 8) stand at a strong 95.61% & a moderate 89.36% respectively. KNN model’s high accuracy promises a high True Positive prediction rate, however the low recall rate signifies that certain cases can be falsely identified as Negative. In a sensitive field such as medical study this could prove fatal. Though the False Negative count is relatively low when compared to the True Positive rate, it shows that the model is mostly reliable through a majority of cases. However, cross validation is highly recommended with the help of other models.

Using the Permutation Importance function, we were able to determine the factors influencing the model, both positive and negative alike. The top influencing features include radius_worst, concave_point_worst, concavity_worst, texture_mean & smoothness_mean. Some of these features (Shown in Figure 9) are affecting the model negatively as well. Therefore, these features signify a significant power of effect on the prediction power of the model between Benign and Malignant.

Random Forest (RF)

After training the Random Forest model, we obtained the feature importance chart as shown in Figure 4. From the chart, it is clear that "concave points_mean" is the most important feature, which contributes about 30% of the result. Additionally, "perimeter_mean", "concavity_mean", and "area_mean" also play important roles in the model. These features are known to have strong discriminative power in distinguishing between benign and malignant tumors.

The confusion matrix on the test set shown in Figure 5 further validates the model's performance. Among the 67 benign cases, the model correctly classified 63, with only 4 false positives. For the 47 malignant cases, the model identified 45 correctly, with just 2 false negatives. The accuracy is 94.74% and the recall is 95.74%. This shows the RF model is particularly effective at identifying malignant cases since it has a high recall.

5. Discussion

5.1 Which features are most indicative of whether a tumor is malignant or benign?

- As shown in Figure 11, 4 Features, namely Compactness_Mean, Smoothness_Mean, Radius_Mean & Texture_Mean have shown the highest consensus between the 3 models, making them key to indicating whether a tumor is malignant or benign.

5.2 How can we accurately classify breast cancer tumors using diagnostic features?

- The accuracy & recall rates of the models could be further improved by fine tuning the features to the most critical ones. In addition to that, an Ensemble model with the best parts of the three models could be built to retain the highest accuracy and recall rates. In furtherance of this, collaboration with oncologists could vastly improve the performance of the models in accurately classifying the cancer tumors through clinical understanding of the diagnostic features.

5.3 Are there any patterns or groupings among the features that distinguish malignant from benign tumors?

- Figures 12 (a), (b) & (c) signify the groupings observed between important features which helped distinguish malignant tumors from benign.

6. Conclusion & Future Directions

It has been noted that all models exhibit strong accuracies, which signify a strong relationship between the features. However, the varying recallability suggests that the test values need to be cross validated across the three models for stronger confidence in the diagnosis. In an optimistic perspective, the models share up to 50% of the significant features with each other. Which opens a possibility that an ensemble model could perform relatively better retaining the best of the three models. However, before conducting an ensemble model analysis, it is important to fine tune the features with high significance in the models.

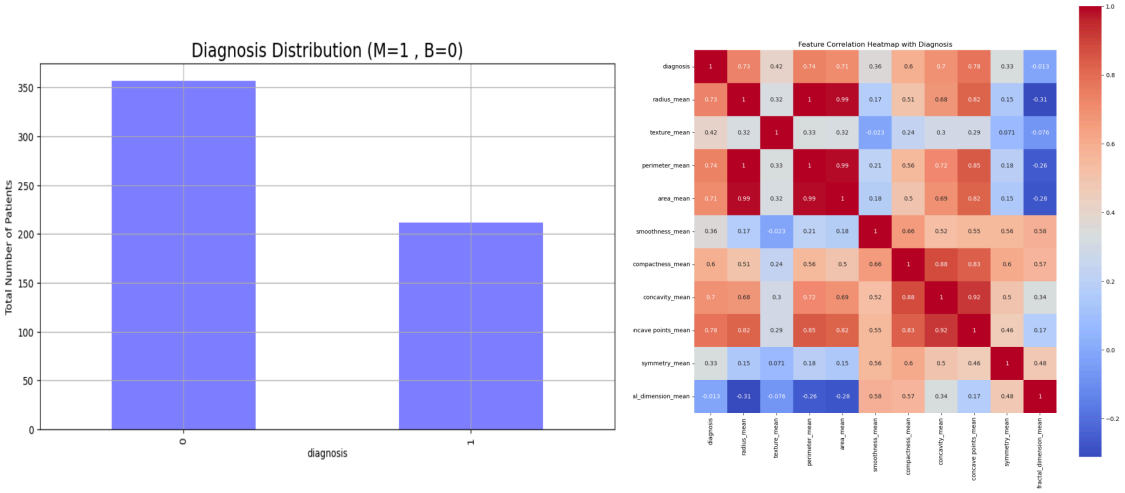
The significant features (represented as common features between models in Figure 11), have been categorised into 3 separate bins for future analysis:

Category	Features
High Priority Consensus Features	Compactness_Mean, Smoothness_Mean, Radius_Mean & Texture_Mean.
Accuracy-Boosting Features	Area_Mean & Concavity_Mean
Recall-Boosting Features	Concave_Points_Mean, Perimeter_Mean, Concavity_Mean & Radius_Worst

Appendix: References

"Breast Cancer Statistics: How Common Is Breast Cancer?" American Cancer Society, 17 Jan. 2024, www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html.

Appendix: Reference Figures



(a) Diagnosis Distribution (b) Correlation Heatmap

Figure 1: Diagnosis Distribution and Correlation Heatmap in EDA

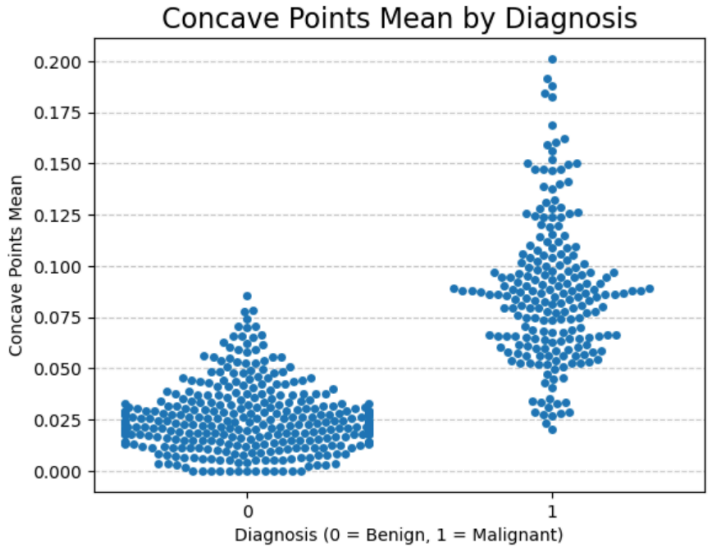


Figure 2: Concave Points Mean by Diagnosis

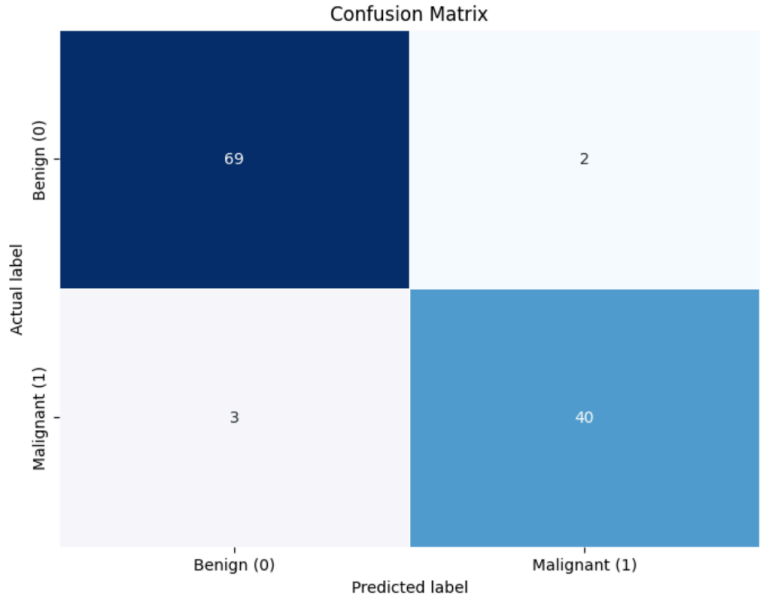


Figure 3: Confusion Matrix for the binary logistic regression

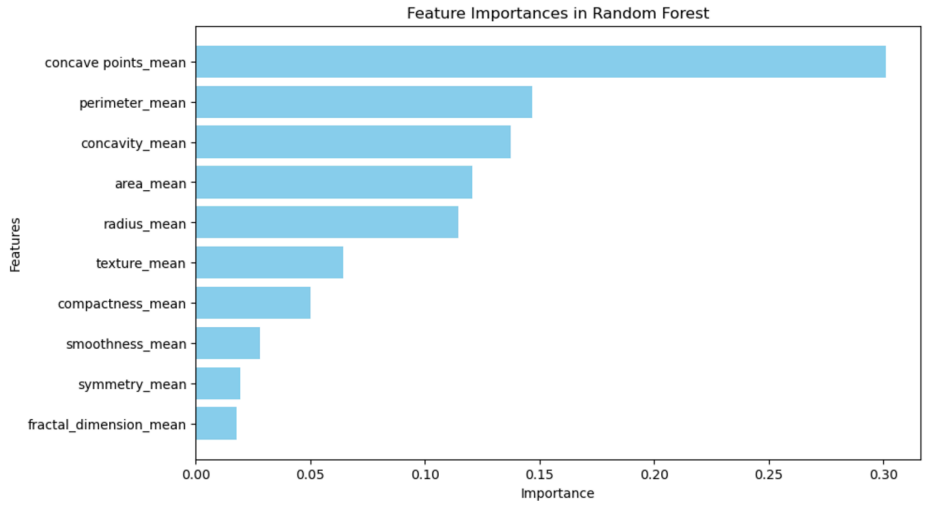


Figure 4: Feature Importance in Random Forest

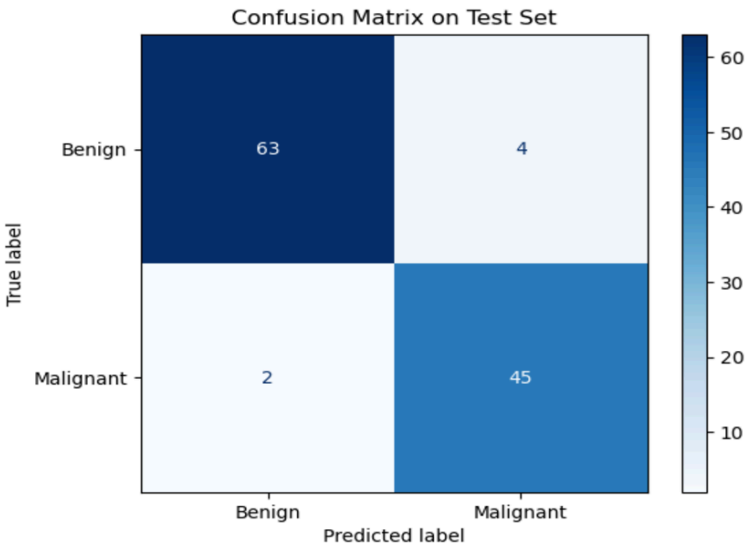


Figure 5: Confusion Matrix for the Random Forest

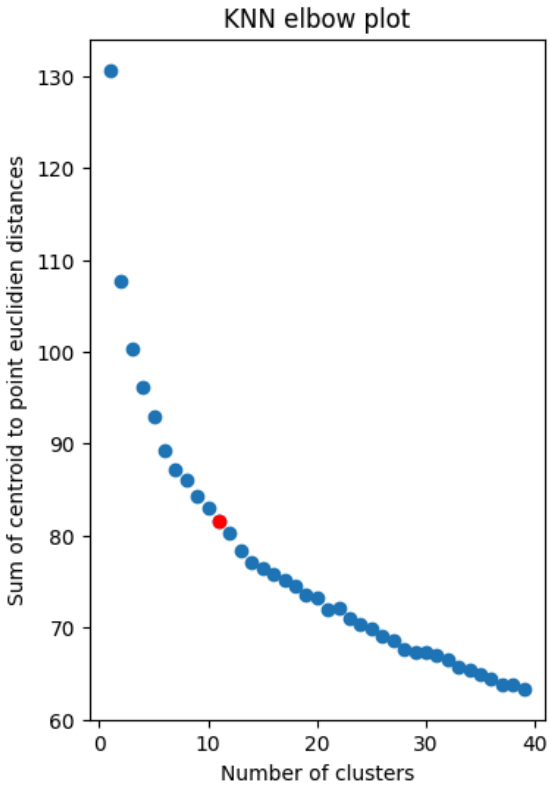


Figure 6: KNN Elbow Plot for cluster number selection

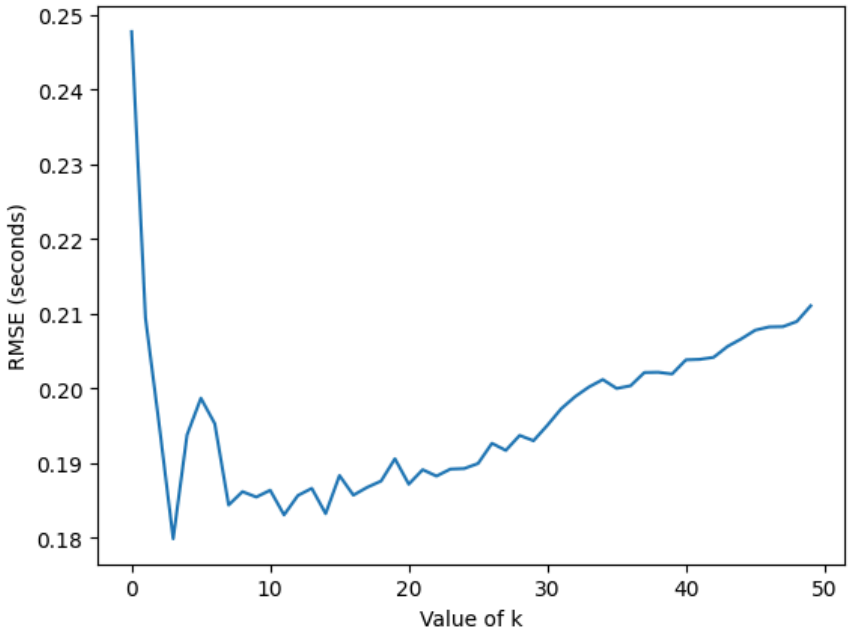


Figure 7: RMSE Scores for best model selection

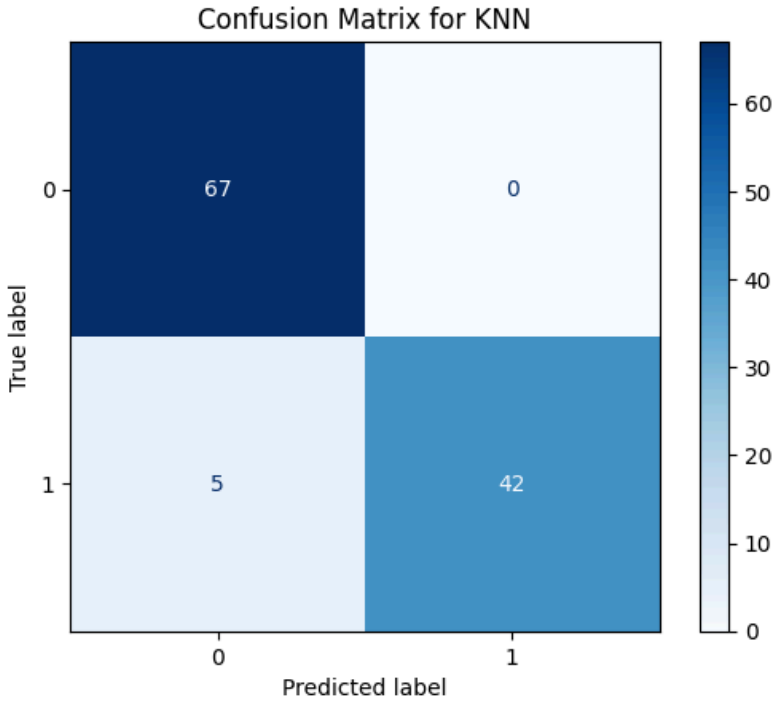


Figure 8: Confusion Matrix for KNN Model

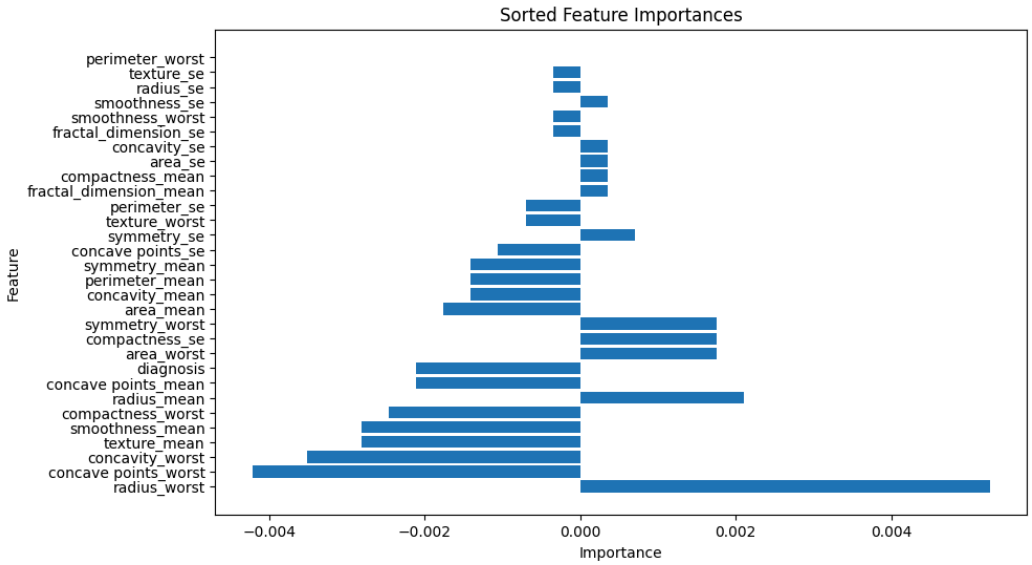


Figure 9: Feature Importance chart for KNN Model

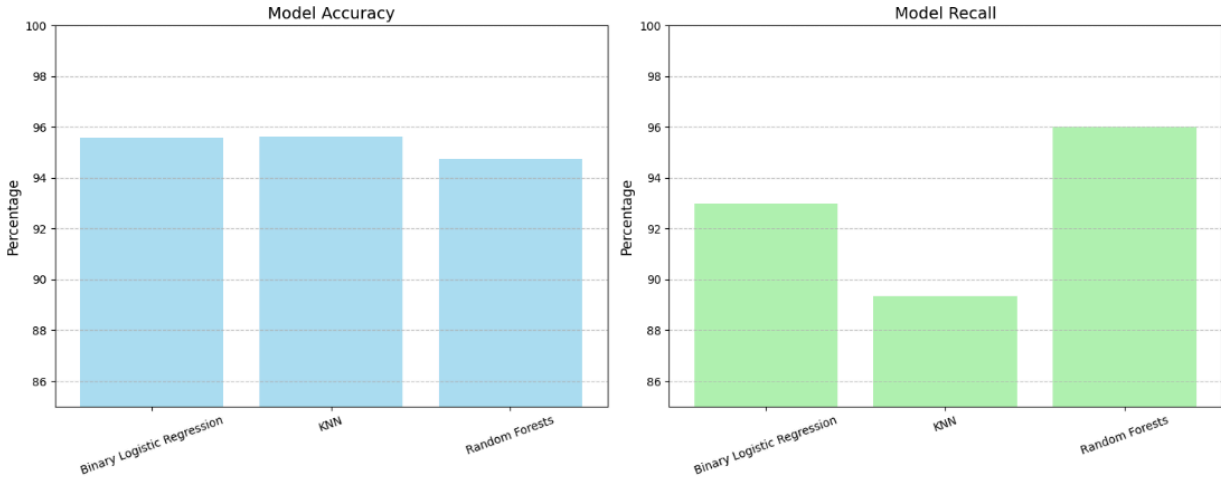


Figure 10: Comparison Charts for Model Accuracy & Recall Rate

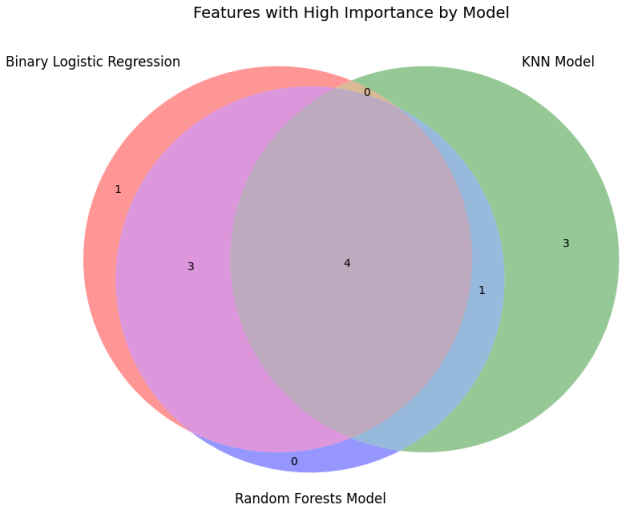


Figure 11: Common High Priority Feature Map between models

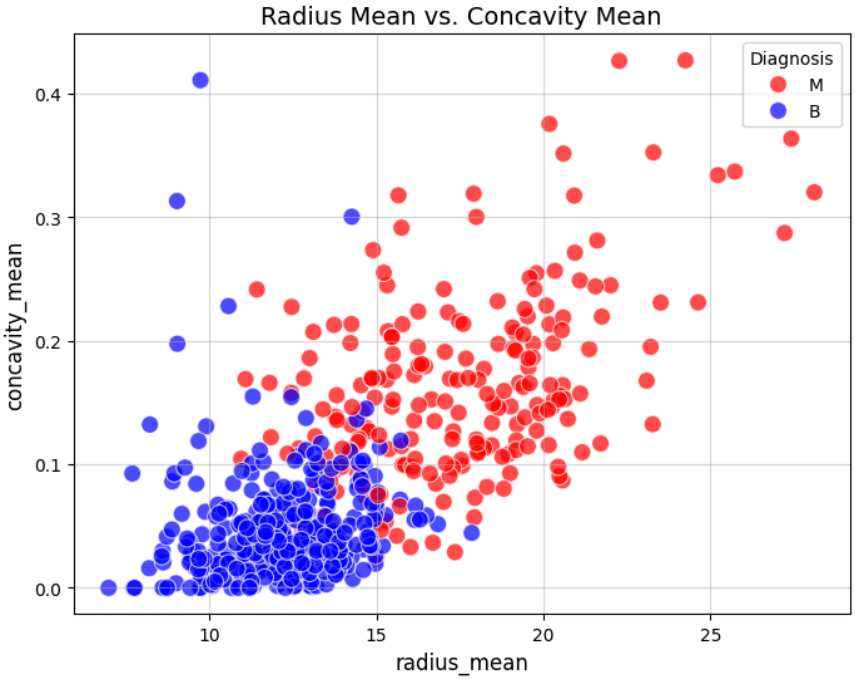


Figure 12(a): Groupings/Patterns exhibit by Features: Radius Mean vs Concavity Mean

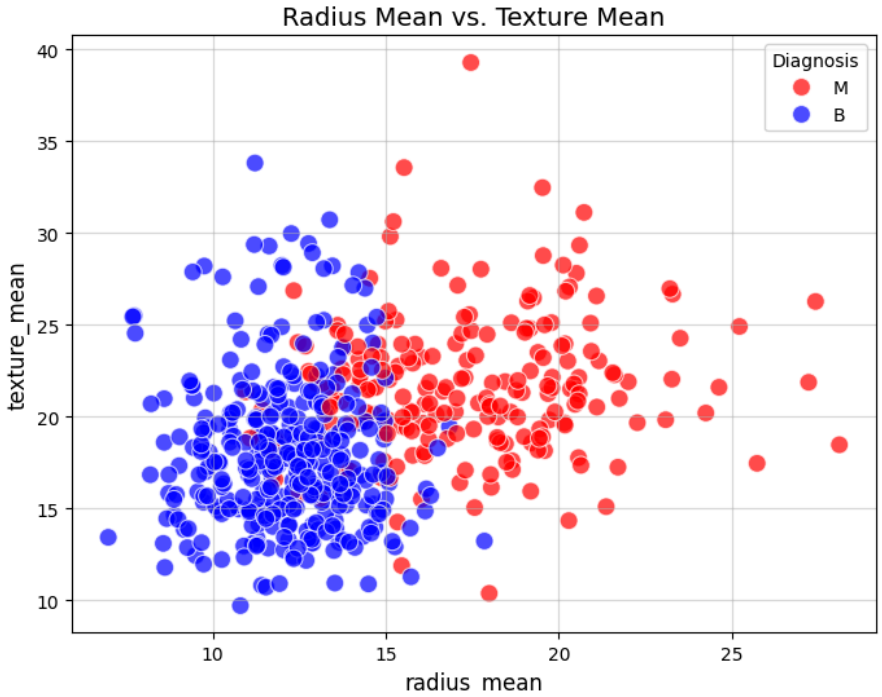


Figure 12(b): Groupings/Patterns exhibit by Features: Radius Mean vs Texture Mean

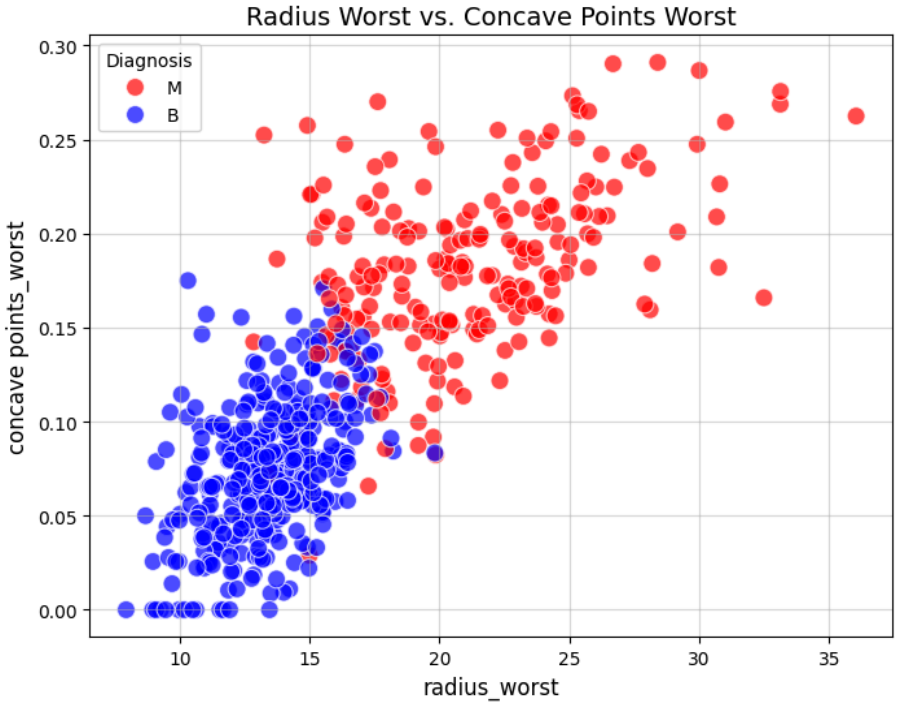


Figure 12(c): Groupings/Patterns exhibit by Features: Radius Worst vs Concave Points Worst