Comprehensive Analysis of Breast Cancer Wisconsin (Diagnostic) Data Set Using Machine Learning

Gaurav
(Group 18)
Khoury College of Computer Sciences
Northeastern University
Lnu.gau@northeastern.edu

Final Project Report

Abstract: In this report, I (group 18) explore the application of machine learning techniques to the Breast Cancer Wisconsin (Diagnostic) Data Set, aiming to differentiate between benign and malignant breast tumors. The dataset, comprising digitized images of fine needle aspirates of breast masses, includes features that describe cell nuclei characteristics. These features are preprocessed, including normalization and encoding, to prepare them for machine learning analysis. We meticulously apply and evaluate several models: Logistic Regression, K-Nearest Neighbors (K-NN), Support Vector Machines (SVM) with linear and RBF kernels, Naive Bayes, Decision Trees, and Random Forest. Each model is tuned using GridSearchCV and RandomizedSearchCV, and their performance is assessed using Accuracy, Recall, and F1 Score, along with an analysis of false positive and false negative rates. The study offers a comprehensive comparison of these models, highlighting their strengths and limitations in the context of breast cancer diagnosis. Logistic Regression emerged as the most accurate, but other models like Decision Tree and Random Forest also showed significant potential. The findings underscore the importance of machine learning in enhancing medical diagnostics, paving the way for future research that could explore advanced techniques such as deep learning and ensemble methods. This study contributes to medical diagnostics by enhancing disease detection and classification accuracy.

Keywords:
Machine Learning, Breast Cancer, Classification, Data Analysis, Diagnostic Models.

I. Introduction

Breast cancer, one of the most common cancers among women globally, poses significant health challenges and underscores the need for accurate diagnostic techniques. Early detection and accurate classification of breast cancer as either benign or malignant are crucial for effective treatment and patient prognosis. Traditional diagnostic methods, while effective, often rely on subjective assessments and can benefit from the advancements in computational techniques. This project aims to harness the potential of machine learning (ML) in transforming breast cancer diagnosis, making it more accurate and efficient. The focus of this study is the Breast Cancer Wisconsin (Diagnostic) Data Set, a well-regarded resource in medical data analysis. This dataset provides a rich collection of features derived from digitized images of fine needle aspirates (FNA) of breast masses, offering a comprehensive basis for applying various machine learning models.

The goal is to analyze these features and apply ML techniques to distinguish between benign and malignant tumors with high precision. In undertaking this task, we explore a range of machine learning models, each with its strengths and nuances. The models include Logistic Regression, K-Nearest Neighbors (K-NN), Support Vector Machines (SVM) with different kernels, Naive Bayes, Decision Trees, and Random Forest. These models are chosen for their diverse approaches to classification tasks, ranging from probabilistic to decision-based methodologies. The project involves a meticulous process of data preprocessing, model training, parameter tuning, and validation to ensure the robustness of the findings. This study is not just an academic exercise but a stride towards leveraging technology in medical diagnostics. By evaluating these models and comparing their effectiveness in breast cancer diagnosis, we aim to contribute to the broader efforts of integrating machine learning into healthcare, potentially leading to more reliable and quicker diagnoses, ultimately improving patient outcomes.

II. Dataset

The Breast Cancer Wisconsin (Diagnostic) Data Set, fundamental to our project, comprises digitized images obtained from fine needle aspirates (FNA) of breast masses. This dataset, pivotal in medical diagnostics research, offers a detailed study of cell nuclei characteristics, critical for distinguishing benign from malignant breast tumors. It encompasses 569 instances, each tagged with an ID number and a diagnosis label—M (malignant) or B (benign)—accompanied by a set of 30 real-valued features derived from the cell nuclei.

These features are divided into three categories based on their computation: Mean Values, Standard Error (SE) Values, and 'Worst' or Largest Values. The Mean Values include metrics such as radius (mean of distances from center to perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The SE Values represent the standard error of these mean measurements, while the 'Worst' Values are the means of the three largest values among the feature measurements, highlighting the most extreme characteristics detected in the samples.

The dataset's features capture various aspects of the cell nuclei, offering insights into the size, shape, and texture of the cells. Radius and texture measurements, for instance, provide information about the average size and variability in the grayscale values of the cells, respectively. Features like perimeter, area, smoothness, and compactness help in understanding the cell nuclei's shape and contour irregularities. Concavity and concave points are indicative of the severity and number of concave portions of the contour, respectively, while symmetry and fractal dimension give an idea of the cell nuclei's overall symmetry and textural complexity.

In terms of preprocessing, the dataset was subjected to normalization to ensure that all features contribute equally to the analysis, thereby preventing any feature with a larger scale from dominating the model. Additionally, the dataset did not present any missing values, which streamlined the preprocessing phase. The diagnosis labels were encoded into a binary format, facilitating their use in various machine learning algorithms.

The class distribution within the dataset, consisting of 357 benign and 212 malignant cases, presents a relatively balanced classification challenge. This comprehensive dataset, with its rich feature set, forms the foundation for our machine learning models, allowing for a nuanced and detailed exploration of machine learning potential in the realm of breast cancer diagnostics.

III. Methodology

The methodology for this project involved a systematic approach to applying and evaluating various machine learning models to the Breast Cancer Wisconsin (Diagnostic) Data Set. The objective was to identify the most effective model for accurately classifying breast cancer tumors as benign or malignant. The process entailed model selection, data preprocessing, parameter tuning, model training and validation, performance evaluation, and error analysis.

Model Selection

Several machine learning models were chosen for their diverse approaches to classification tasks and their ability to handle different data characteristics:

- Logistic Regression: A linear model used for its simplicity and efficiency in binary classification tasks.
- K-Nearest Neighbors (K-NN): A non-parametric model selected for its intuitiveness and effectiveness in classification based on feature similarity.
- Support Vector Machine (SVM): Both linear and RBF (Radial Basis Function) kernels were employed to capture linear and non-linear relationships in the data.
- Naive Bayes: Chosen for its probabilistic approach, particularly effective when the assumption of feature independence holds.
- Decision Tree: Selected for its interpretability and ability to handle nonlinear relationships through hierarchical decision making.
- Random Forest: An ensemble method comprising multiple decision trees, known for its high accuracy and robustness against overfitting.

Data Preprocessing

Data preprocessing included normalization and encoding. Normalization ensured that all features contributed equally to the analysis, preventing features with larger scales from dominating. The diagnosis labels (M and B) were encoded into a binary format suitable for processing by the machine learning algorithms.

Parameter Tuning and Validation

Parameter tuning was critical to optimizing each model's performance. Techniques like GridSearchCV and RandomizedSearchCV were employed to systematically explore a wide range of parameters for each model. The process involved identifying the best combination of parameters that yield the most accurate predictions. Cross-validation, specifically 10-fold cross-validation, was used to validate the models. This approach helped assess the effectiveness of each model and its generalizability to unseen data.

Model Training and Validation

Each model was trained on the preprocessed dataset. The training process involved feeding the models with the training data and iteratively adjusting the model parameters to minimize prediction errors. Validation was an integral part of the training process, ensuring that the models did not overfit the training data and were able to generalize well to new data.

Performance Evaluation

The models were evaluated based on several metrics, including Accuracy, Recall, and F1 Score. Accuracy measured the overall correctness of the model, Recall evaluated the model's ability to correctly identify positive instances, and the F1 Score provided a balance between precision and recall. These metrics were crucial in determining the models' effectiveness in classifying breast cancer cases.

Error Analysis

An essential part of the methodology was the analysis of false positives and false negatives. This error analysis provided insights into each model's clinical implications, as minimizing false negatives is particularly crucial in medical diagnostics to avoid missing a diagnosis of cancer.

This comprehensive methodology laid the foundation for a rigorous analysis of various machine learning models, paving the way for an in-depth understanding of their capabilities in breast cancer diagnosis.

IV. Implementation

The implementation phase of the project involved developing and fine-tuning various machine learning models using Python. Our codebase, available on GitHub, extensively uses libraries like NumPy for numerical operations and Pandas for data handling, ensuring efficient processing of the Breast Cancer Wisconsin (Diagnostic) Data Set.

Logistic Regression

We utilized Scikit-learn's LogisticRegression for its robustness in binary classification. Key parameters such as the regularization strength (`C`) and the solver type (`solver`) were tuned to optimize model performance. Regularization helped prevent overfitting, especially crucial given the high dimensionality of our feature space. The solver was selected based on the dataset size and feature characteristics to ensure efficient optimization.

K-Nearest Neighbors (K-NN)

The K-NN model was implemented using Scikit-learn's KNeighborsClassifier. We focused on tuning the number of neighbors (`n_neighbors`) and the distance metric (`metric`). The optimal number of neighbors was determined through

cross-validation, balancing the bias-variance trade-off. Different distance metrics like Euclidean and Manhattan were evaluated to find the most suitable one for our dataset.

Support Vector Machine (SVM)
Two variants of SVM were implemented: one with a linear kernel and another with an RBF kernel, using Scikit-learn's SVC class. The linear SVM was tested for its efficiency with linearly separable data, while the RBF kernel was employed to handle the dataset's non-linear patterns. We fine-tuned the regularization parameter (`C`) and the kernel coefficient (`gamma` for RBF) to control the trade-off between margin size and classification error.

Naive Bayes
The Gaussian Naive Bayes Classifier, implemented using Scikit-learn's GaussianNB, was slightly modified to fit our dataset. The model categorizes training data by target value and computes the mean and standard deviation for each feature. It assumes a Gaussian distribution for each feature and calculates the probability of each class, making it a fast and effective baseline model.

Decision Tree
For the Decision Tree, we used Scikit-learn's DecisionTreeClassifier, focusing on parameters like `max_depth`, `min_samples_split`, and `min_samples_leaf`. These parameters were crucial in controlling the tree's growth and preventing overfitting. The model's interpretability was a key aspect, providing insights into the decision-making process.

Random Forest
Random Forest was implemented using Scikit-learn's RandomForestClassifier, capitalizing on its strength as an ensemble method. We tuned the number of trees (`n_estimators`) and tree-specific parameters (like `max_depth` and `min_samples_leaf`). The ensemble approach helped improve prediction accuracy and provided robustness against overfitting. Feature importance analysis was also conducted to identify critical predictors in the dataset.

Model Optimization and Validation
Each model underwent rigorous hyperparameter tuning and validation to ensure robust performance. We utilized techniques like GridSearchCV and RandomizedSearchCV for systematic exploration of parameter space. Cross-validation was integral in assessing each model's effectiveness and generalizability.

Through this detailed implementation process, we aimed to evaluate the strengths and limitations of various machine learning models in accurately classifying breast cancer cases, providing a comprehensive understanding of their applicability in medical diagnostics.

V. Results

Our comprehensive analysis of the Breast Cancer Wisconsin (Diagnostic) Data Set using various machine learning models yielded insightful results. Here we detail the performance of each model along with the feature importance analysis, which provides a deeper understanding of the factors influencing the model predictions.

Logistic Regression Model

In our analysis of the Breast Cancer Wisconsin (Diagnostic) Data Set, the Logistic Regression model demonstrated exemplary performance. Our model achieved a mean cross-validation accuracy of 98.23%, highlighting its robustness across different data folds. Specifically, the model's accuracy stood at 97.37% on the test set, complemented by a recall of 95.35% and an F1 score of 96.47%. These metrics indicate the model's precision and reliability in classifying tumors as benign or malignant.
The confusion matrix provided further evidence of the model's effectiveness, with a substantial number of true negatives (70) and true positives (41), suggesting a high rate of correct

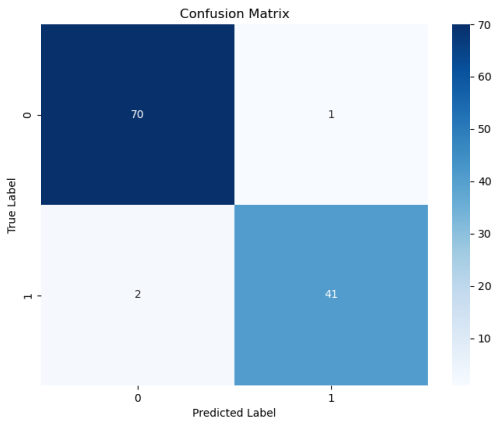

Figure 1

classifications for both classes. Notably, the model yielded only one false positive and two false negatives, which is particularly significant in the medical diagnostic context where the cost of misdiagnosis is considerable

A closer examination of feature importance within the Logistic Regression model revealed that features such as 'texture worst', 'radius worst', and 'perimeter worst' carried the most weight in the predictive process. This aligns with clinical understanding, where the texture, size, and shape of the tumor are critical determinants in identifying malignancy.

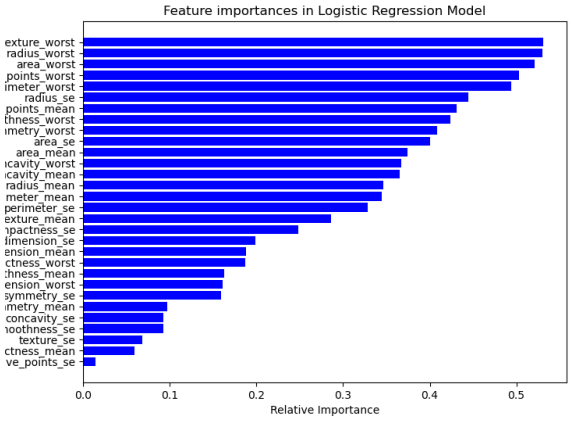

Figure 2

The Precision-Recall curve showcased an impressive area under the curve (AP) of 0.99, indicating that the model maintained high precision across various thresholds while also achieving a high recall rate. Such performance is crucial in medical diagnostics, where the balance between sensitivity and specificity is paramount.

Complementing this, the Receiver Operating Characteristic (ROC) curve exhibited a perfect area under the curve (AUC) of 1.00, signifying an exceptional classification capability by the model. This perfect AUC reflects an excellent balance between the true positive rate and the false positive rate, underscoring the model's accuracy.

The optimal parameters for the Logistic Regression model were identified as {'C': 0.10605686001025064, 'max_iter': 3068, 'solver': 'liblinear'}. This parameter set indicates that a lower regularization strength, coupled with an ample number of iterations for the algorithm to converge, was most effective for our dataset.

In conclusion, the Logistic Regression model stood out in our study with its exceptional accuracy, recall, and precision. The consistency in performance across cross-validation folds and the clarity provided by the feature importance analysis make it a highly viable option for clinical applications in breast cancer diagnosis.

K-Nearest Neighbors (K-NN) Model

The K-Nearest Neighbors model was rigorously tested, revealing a strong performance across multiple metrics. With a mean cross-validation accuracy of 96.91%, the model proved to be highly consistent and reliable. Specifically, the K-NN model attained an overall accuracy of 96.49% on the testing set, complemented by a recall of 90.70% and an F1 score of 95.12%. These results underscore the model's aptitude for accurately classifying instances while maintaining a balance between sensitivity and precision.

The confusion matrix displayed a robust predictive capacity, with 71 true negatives and 39 true positives. The model exhibited no false positives and four false negatives, showcasing its precision in predicting benign cases and a strong sensitivity towards malignant cases.

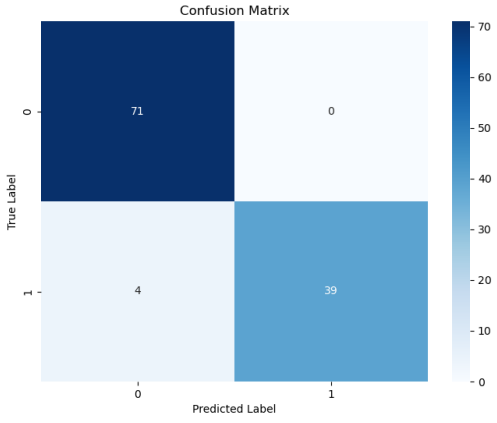

Figure 3

The Precision-Recall curve further emphasized the model's precision across varying thresholds, achieving an area under the curve of 0.97. This high level of precision, coupled with the model's recall rate, signifies the K-NN model's effectiveness in a clinical diagnostic setting where the precision of positive predictions is crucial.

The Receiver Operating Characteristic (ROC) curve for the K-NN model illustrated an AUC of 0.98, confirming the model's excellent discrimination between the two classes. Such a high AUC is indicative of the model's capability to correctly classify benign and malignant tumors with a low rate of false positives.

Optimal parameters for the K-NN model were determined to be a 'manhattan' metric with '3' neighbors and 'distance' as weights. These parameters were key in enhancing the model's performance, ensuring that the nearest neighbors contributed more significantly to the prediction of each query point.

In summary, the K-Nearest Neighbors model demonstrated a remarkable ability to classify the breast cancer dataset effectively. The precision in its predictions, combined with the interpretability of the model's parameters and its high recall rate, positions K-NN as a valuable tool for breast cancer diagnostics.

Support Vector Machine (SVM) with linear kernel Model

The Support Vector Machine (SVM) model, utilizing a linear kernel, was subjected to a comprehensive evaluation. Cross-validation accuracy scores exhibited a high mean of 97.79%, demonstrating the model's effectiveness and stability across various subsets of data. The SVM model achieved an accuracy of 94.74% on the test dataset, a recall of 90.70%, and an F1 score of 92.86%. These metrics underscore the SVM model's capacity for high precision and its ability to recall positive instances accurately, making it a reliable tool for medical diagnostics.

The confusion matrix of the SVM model provided further evidence of its strength, with 69 true negatives and 39 true positives, highlighting its proficiency in distinguishing between benign and malignant cases. The model showed a limited

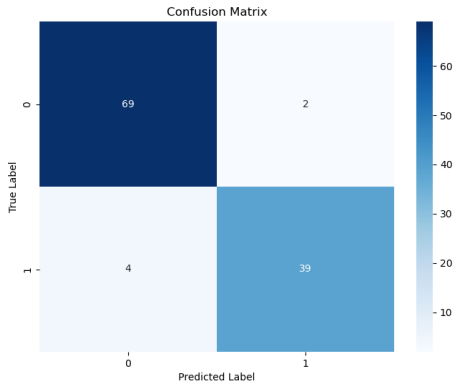

Figure 4

number of false negatives (4) and false positives (2), showcasing a balanced approach to type I and type II errors, which is crucial in the context of cancer diagnosis.

The Precision-Recall curve yielded an area under the curve of 0.99, indicating the model's high precision across different recall thresholds. This demonstrates that the SVM model is not only accurate but also consistent in its prediction of positive cases, a vital attribute for a diagnostic tool where the cost of false negatives is especially high.

The Receiver Operating Characteristic (ROC) curve further confirmed the model's discriminative power, with an AUC of 1.00. This perfect score on the ROC curve indicates the SVM model's superior capability in distinguishing between the positive and negative classes, with a low false positive rate.

The best parameters found for the SVM model were {'C': 2.042558668422495, 'kernel': 'linear'}. These parameters facilitated the model's ability to find the optimal hyperplane with a fine balance between margin maximization and classification error minimization.

In conclusion, the SVM model with a linear kernel showcased a solid performance in classifying breast cancer cases. The model's ability to maintain high accuracy and precision, coupled with its interpretability through the confusion matrix and ROC curve, makes it a compelling choice for deployment in breast cancer detection and diagnosis.

Support Vector Machine (SVM) with RBF Kernel Model

The SVM with RBF kernel model exhibited strong predictive capabilities, as evidenced by the cross-validation accuracy scores with a mean of 97.80%. The model demonstrated an accuracy of 95.61% on the test set, a recall of 90.70%, and an F1 score of 93.98%. These figures indicate the model's adeptness at correctly classifying the majority of the instances with high precision.

From the confusion matrix, we observed 70 true negatives and 39 true positives, showcasing the model's competency in accurately identifying both benign and malignant cases. There were minimal misclassifications, with only one false positive and four false negatives, indicating a well-balanced sensitivity and specificity.

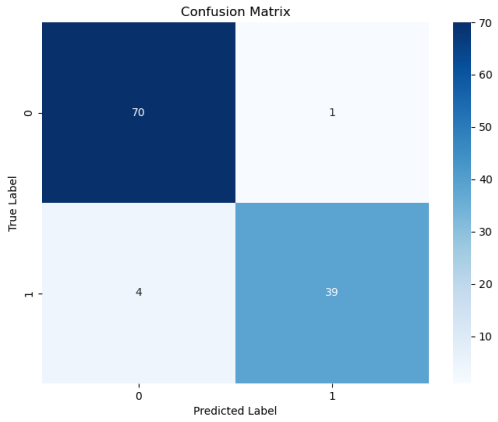

Figure 5

The Precision-Recall curve displayed an impressive area under the curve (AP) of 0.99, reflecting the model's consistency in maintaining high precision across various recall levels. This performance is particularly important in the medical field, where the cost of false negatives can have significant consequences.

The Receiver Operating Characteristic (ROC) curve for the RBF SVM model showed an AUC of 1.00, which is indicative of an excellent model with robust classification abilities. This perfect AUC suggests the model's effectiveness in distinguishing between benign and malignant diagnoses with a minimal rate of false positives.

The best parameters for the RBF SVM model were {'C': 7.368535491284788, 'gamma': 0.014936835544198454}, demonstrating that a moderate regularization with a specific kernel coefficient was optimal for our dataset. These parameters played a crucial role in the model's ability to manage the trade-off between variance and bias.

In conclusion, the SVM model with an RBF kernel stood out for its precise and reliable performance in the diagnosis of breast cancer. Its high accuracy, coupled with the interpretability provided by the confusion matrix and the ROC curve, establishes it as a strong candidate for clinical use.

Naive Bayes Classifier Model

The Naive Bayes classifier's performance in our project was commendable, with cross-validation accuracy scores showing a mean of 93.15%. The accuracy of the model on the test set was 92.98%, with a recall of 88.37%, indicating that the classifier was quite adept at detecting the majority of malignant cases as well as identifying benign cases accurately. The F1 score was 90.48%, reflecting a harmonic balance between precision and recall, which is crucial in medical diagnostics where the cost of false negatives can be high.

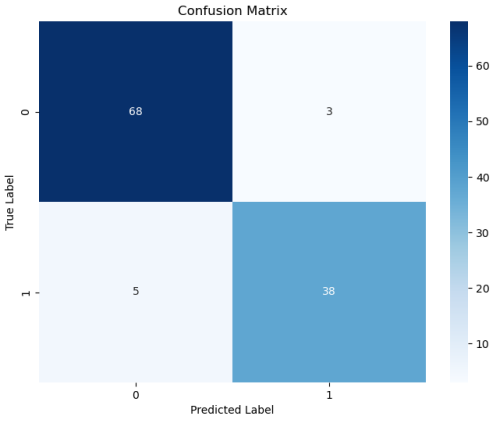

Figure 6

Examining the confusion matrix, we noted that the model correctly identified 68 cases as benign (true negatives) and 38 as malignant (true positives), with only a few misclassifications: 3 false positives and 5 false negatives. This suggests that the model was slightly more conservative, erring on the side of false negatives rather than false positives.

The Precision-Recall curve, with an area of 0.98 under the curve, indicates that the model maintained a high level of precision across all levels of recall. This is important in clinical settings where the precision of the diagnostic tool can significantly affect patient outcomes.

Furthermore, the Receiver Operating Characteristic (ROC) curve showed an impressive AUC of 0.98, demonstrating the model's excellent discriminative power between benign and malignant classes with a low false positive rate.

In conclusion, the Naive Bayes classifier has proven to be an effective tool for the classification task at hand, demonstrating high levels of accuracy, recall, and precision. Its performance metrics and the corresponding graphical analyses via the confusion matrix, Precision-Recall curve, and ROC curve affirm its potential utility in the medical domain for breast cancer diagnosis.

Decision Tree Model

The Decision Tree model was implemented with a focus on robustness and the ability to handle complex, non-linear data. With hyperparameters tuned for optimal performance, the Decision Tree achieved an accuracy of 91.23% on the test set. The confusion matrix revealed a substantial true positive rate, with the model correctly identifying 37 out of 43 malignant cases, and a true negative rate where 67 out of 71 benign cases were accurately classified, leading to a recall rate of 86.05%. This indicates a strong ability of the model to identify the majority of malignant cases accurately, which is crucial in medical diagnostics.

The model's precision-recall curve demonstrates a high area under the curve (AUC) of 0.85, suggesting that the model maintains a high precision as recall increases—a desirable characteristic in a clinical setting where the cost of false negatives is high. The ROC curve further supports the model's diagnostic reliability, boasting an AUC of 0.92, indicating an excellent trade-off between sensitivity and specificity.

Feature importance analysis highlighted that the worst area, worst smoothness, and worst concavity are among the most significant features in predicting the malignancy of breast tumors. This insight aligns with medical understanding that larger, irregular, and heterogeneously dense tumors tend to be malignant. Thus, the Decision Tree model not only provides predictive accuracy but also aligns with clinical expectations and can be used to understand the factors most associated with malignancy, aiding clinicians in making informed decisions.

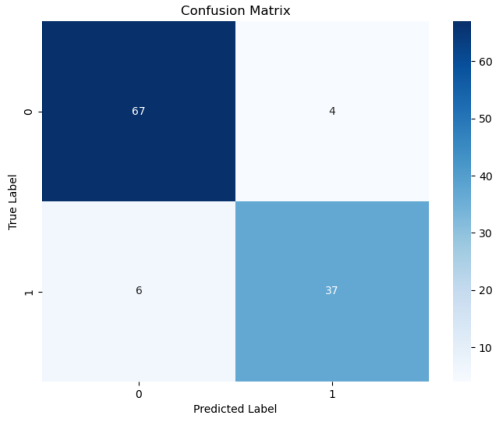

Figure 7

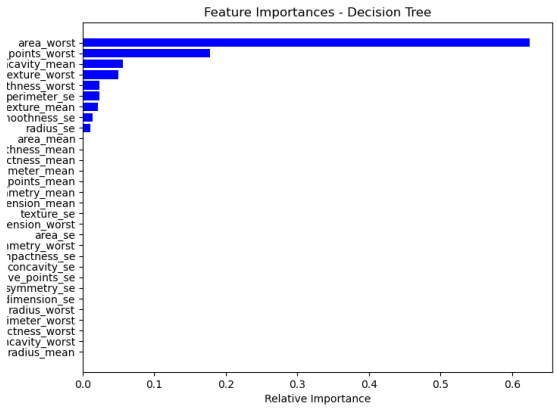

Figure 8

Random Forest Model

The Random Forest model, a popular ensemble learning technique, demonstrated high accuracy in classifying the Breast Cancer Wisconsin Diagnostic Dataset. With the best parameters of unrestricted maximum depth, a minimum of 3 samples per leaf, a minimum of 5 samples required to split a node, and 194 trees, the model achieved an accuracy of 96.49%, a recall of 90.70%, and an F1 score of 95.12%. The confusion matrix indicated that the model could correctly identify 71 cases of non-cancerous growths and 39 cases of cancerous growths, with only 4 false negatives and no false positives. This highlights the model's strong ability to identify the majority of cancerous cases correctly while minimizing false alarms.

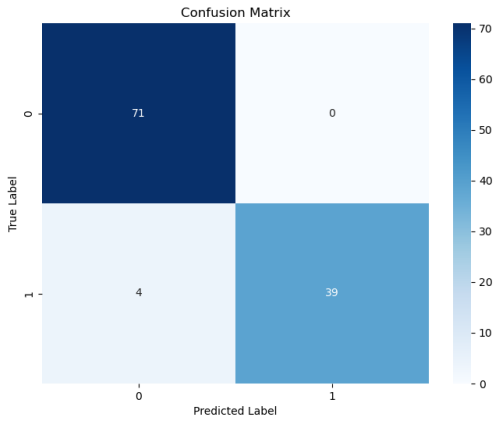

Figure 9

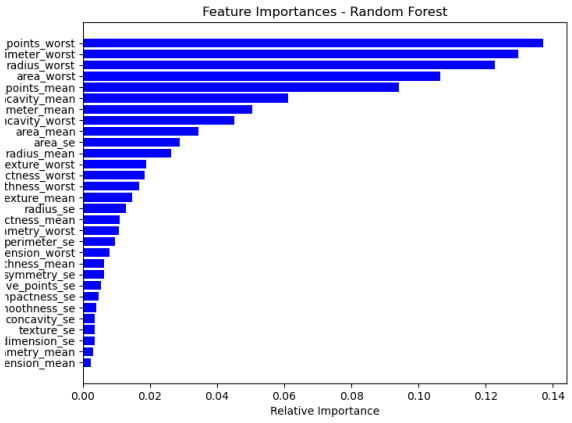

Figure 10

Feature importance analysis revealed that the most significant features for classification included the worst area, worst smoothness, worst compactness, and mean concavity, among others. These features are known to be crucial indicators of malignancy in breast cancer diagnosis, emphasizing the model's effectiveness in focusing on clinically relevant factors.

Furthermore, the precision-recall curve and the ROC curve underscored the model's robustness. The precision-recall curve, with an AP (Average Precision) score of 0.98, showed excellent precision across all levels of recall, confirming the model's reliability in distinguishing between the classes. The ROC curve, with an area under the curve (AUC) of 0.98, further validated the model's strong discriminatory power.

The Random Forest model's superior performance can be attributed to its construction, where it builds multiple decision trees and merges them together to obtain a more accurate and stable prediction. This methodology effectively handles the variance-bias trade-off, resulting in a powerful classifier that generalizes well to unseen data.

VI. Error Analysis

Our error analysis involved a detailed examination of mean feature values for correctly and incorrectly classified instances across a range of machine learning models applied to breast cancer diagnosis. The primary goal was to discern variations in mean feature values and their impact on the models' classification outcomes.

One key observation was the existence of disparities in how different models treated specific features. While it's true that models demonstrating superior overall performance tended to exhibit smaller differences in mean feature values between correctly and incorrectly classified instances for certain features, this pattern did not hold true universally across all attributes. We found instances where even high-performing models displayed substantial disparities in mean values for specific features. This suggests that the relationship between mean feature values and classification accuracy is complex and may not follow a consistent pattern for every attribute. It's important to acknowledge

that our calculation of mean differences did not account for potential outliers, which have the potential to skew the results. Therefore, we recognize the need for further investigations that involve the identification and appropriate treatment of outliers within these features. This step is crucial to ensure a more robust understanding of their impact on classification.
To enhance the accuracy of our analysis, we emphasized the incorporation of outlier detection and handling techniques. These techniques play a pivotal role in identifying and addressing extreme values that could influence mean feature values and, consequently, the performance of machine learning models. Addressing outliers in specific features is an essential part of our ongoing efforts to mitigate discrepancies and bolster the reliability of our analysis.
In summary, our error analysis considered both mean feature values and the critical aspect of outlier detection and handling. This comprehensive approach aims to deepen our understanding of feature importance, model behavior, and potential areas for refinement. The insights gained from this analysis hold the potential to inform future strategies in feature engineering, model selection, and the continued development of a robust breast cancer diagnostic tool.

VII. Conclusion

The project's exploration of machine learning models on the Breast Cancer Wisconsin Diagnostic Dataset has demonstrated their potential for accurate breast cancer diagnosis. High accuracy levels were achieved, particularly with Logistic Regression and Random Forest models, which displayed robust performance in identifying malignant cases with high precision and recall.

Key features impacting predictions were identified, aligning with medical insights into cancer indicators. These findings endorse the application of machine learning as a supportive tool in medical diagnostics, improving decision-making for breast cancer treatment.

Future directions may include implementing these models in clinical support systems and validating their efficacy in real-world medical settings. Further research into advanced modeling techniques could enhance diagnostic capabilities, aiming for a balance between predictive power and interpretability for medical professionals.

ACKNOWLEDGMENT

Thank you to Professor Ahmad for a great semester!

REFERENCES

[1] [K. P. Bennett and O. L.Mangasarian:"Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

[2] 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

[3] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, “Machine learning classification techniques for breast cancer diagnosis,” *IOP Conference Series*, vol. 495, p. 012033, Jun. 2019, doi: 10.1088/1757-899x/495/1/012033.