Breast Cancer Diagnosis Report

By

Team 1

Keerthi Balaji

Boukar Mahamat Mahamat

Modinat Moshood

Rachelle Patrick

Executive Summary

The goal of this project was to develop a predictive model for classifying breast tumors as malignant or benign, using data analysis and machine learning techniques. The primary objective was to improve the understanding of breast cancer diagnoses by building reliable models that accurately predict tumor types, focusing on key indicators such as tumor size, shape, and internal structure.

The dataset contains variables such as radius, texture, perimeter, area, and smoothness, which describe tumor characteristics. A classification tree and logistic regression were selected as the two primary models for this project due to their effectiveness in handling classification tasks. The models were evaluated using performance metrics such as accuracy, precision, recall, specificity, and F1 score, with a focus on minimizing false negatives to ensure the reliability of malignant case detection.

Data preprocessing included handling missing values, creating dummy variables, and normalizing the data to ensure it was suitable for machine learning algorithms. Both models were fine-tuned through parameter adjustments to improve performance. The classification tree initially demonstrated high accuracy but showed signs of overfitting, while logistic regression exhibited better generalization and higher sensitivity.

In model comparison, the classification tree outperformed logistic regression in accuracy and specificity. However, logistic regression showed superior sensitivity and F1 scores, making it the better model for identifying malignant cases. Given the high stakes of breast cancer detection, minimizing false negatives is crucial, and logistic regression achieved this more effectively.

Although both models performed well, we recommend further exploration of additional machine learning techniques to improve overall accuracy and sensitivity. Achieving a diagnosis rate higher than 92.5% is critical for ensuring reliable detection and minimizing risk for patients. Therefore, ongoing model refinement and testing are necessary for improving breast cancer diagnosis outcomes.

Business Problem

The goal of this project is to leverage data analysis and machine learning techniques to develop a highly sensitive predictive model for classifying breast tumors as malignant or benign.

Objective

The primary objective of the **Breast Cancer Data Analysis** is to explore and utilize the dataset to enhance the understanding and prediction of breast cancer diagnosis. By distinguishing between malignant and benign tumor cases, the analysis seeks to build reliable classification models using data mining and analysis techniques. These models aim to accurately predict tumor types based on the input features. By analyzing the relationships and correlations between variables, researchers can pinpoint key indicators of malignant and benign tumors. This not only improves model performance but also enhances interpretability, enabling healthcare professionals to better understand the factors influencing breast cancer outcomes. Statistical analysis and data visualization further complement this process, providing clear insights into feature distributions and relationships.

Data Description:

The dataset contains a dependent variable, Diagnosis, which classifies tumors as either Malignant (M) or Benign (B). It includes 10 independent variables describing tumor characteristics: Radius (size), Texture (internal irregularity), Perimeter (boundary length), Area (surface area), Smoothness (boundary evenness), Compactness (roundness or sprawl), Concavity (inward curvature), Concave Points (distinct inward sections), Symmetry (overall symmetry), and Fractal Dimension (boundary complexity). These variables are measured in three categories: Mean (average), Standard Error (variability), and Worst Measures (mean of the three largest values, reflecting extreme characteristics).

(For the data exploration conducted, please see Appendix Section 1.)

Methodology

Our methodology involved three steps: **data preprocessing**, where we cleaned, normalized, and encoded the data; **modeling**, where we trained and tested classification tree and logistic regression models; and **evaluation**, where we compared their performance using metrics like accuracy, precision, recall, and F1 score to determine the better model for classifying malignant and benign cases.

Models Selected and why

Classification trees and logistic regression are widely used supervised learning models for classification tasks. A classification tree is a decision-tree-based algorithm that splits the dataset into subsets based on feature values to create a tree-like structure. The model makes predictions by traversing the tree based on the input features until reaching a leaf node, which assigns a class label. Classification trees are intuitive, easy to interpret, and handle non-linear relationships between features and outcomes. On the other hand, logistic regression is a statistical model that predicts the probability of a binary outcome using a sigmoid function. Logistic regression is computationally

efficient, interpretable, and well-suited for datasets with fewer non-linear interactions. It also provides insights into the importance of features through model coefficients.

Evaluation Metrics

We evaluate the model using accuracy, precision, recall, specificity, and F1 score. Accuracy measures the proportion of correct predictions, while precision indicates the reliability of positive predictions for malignant cases. Recall (sensitivity) assesses the model's ability to detect malignant cases, and specificity measures the ability to correctly identify benign cases. The F1 score balances precision and recall, making it ideal for imbalanced datasets.

Data Preprocessing

Using the Excel Analytic Solver Add-On and the steps learned in class, we conducted a structured data preprocessing approach to ensure the dataset was ready for analysis. The process begins by handling missing values to ensure that incomplete records do not compromise the reliability of the results. Dummy variables are created to convert categorical data into numerical format, allowing machine learning algorithms to process the data seamlessly. Irrelevant or redundant features are removed to reduce complexity and enhance the interpretability of the models. The dataset is then split into training, validation, and test sets to build robust predictive models and evaluate their performance effectively. Normalization was chosen over standardization to rescale features with varying scales to a fixed range (0 to 1). This is particularly effective for algorithms like K-Nearest Neighbors and Neural Networks, which are sensitive to feature magnitude. Since the dataset does not follow a normal distribution, normalization ensures better model accuracy and consistency compared to standardization, which is suited for Gaussian-distributed data.

Models

Classification Tree

We implemented a Classification Tree model to predict Diagnosis (malignant or benign) using a partitioned dataset. The process involved initializing the model, fine-tuning its parameters, and iteratively evaluating its performance.

To begin, the Classification Tree model was initialized with its default parameters: **Levels (Depth)** = 7 and **Minimum Records in Terminal Nodes** = 29. The initial outcomes (See appendix, Fig. 6) showed training accuracy of 94.43% and validation accuracy of 92.54%, with sensitivity (recall) at 96.85% for training and 96.47% for validation. While the model performed well, further optimization was needed to improve accuracy and sensitivity, critical for minimizing false negatives in medical diagnoses. To enhance performance, we adjusted two parameters: tree depth (levels) and minimum records in terminal nodes. Iterations with levels ranging from 6 to 20 and minimum records from 12 to 35 revealed key trends: increasing depth beyond 7 led to overfitting and reduced sensitivity, while higher minimum records caused underfitting by missing critical patterns. (See appendix, Fig.7 that details the metrics for all the trials we ran to improve the model). After multiple trials, the optimal configuration was identified with Levels as 7 and Minimum Records in Terminal Nodes as 12. This combination produced the best balance between accuracy and sensitivity. (Results as shown in appendix, Fig.8)

The best model achieved training accuracy of 97.07% and validation accuracy of 93.86%, with sensitivity at 94.49% (training) and 89.41% (validation), specificity at 98.60% (training) and 96.50% (validation), precision at 97.56% (training) and 93.83% (validation), and F1 scores of 96.0 (training) and 91.57 (validation). The final binary classification tree, shown below, classifies data as benign (0) or malignant (1) with 13 nodes and a depth of 5. The root node splits on area_worst (threshold 0.17), with the left subtree splitting further on concave points_worst and other features, predicting class 0. The right subtree splits on concavity_mean (threshold 0.18), terminating in nodes predicting class 1. Key features (area_worst, concave points_worst, concavity_mean) are prioritized, and each branch annotates the record count, balancing accuracy and interpretability.

The improvement in accuracy and sensitivity can be attributed to the careful balance of model complexity. By limiting the **levels** to 7, we prevented the tree from growing excessively deep and overfitting the training data. Simultaneously, reducing the **minimum records in terminal nodes** to 12 allowed the tree to make more splits, capturing subtle patterns in the data that differentiate benign and malignant diagnoses. Sensitivity remained a focus of optimization. The final model achieved a strong sensitivity score of **89.41%** on validation data, indicating reliable detection of malignant cases. (See appendix, Fig.9)

Logistics Regression

We conducted logistic regression to predict diagnosis, initially including all variables. The first model showed overfitting, with perfect training scores and lower validation scores, indicating poor generalization to unseen data. (See appendix, Fig. 10)

To address this, we removed highly correlated "Worst Measures" variables and ran a second regression, which improved generalization and reduced overfitting. This model showed strong performance across all metrics but still had a slight gap between training and validation scores. (See appendix, Fig. 11)

We then identified and removed redundant size and shape descriptors: Perimeter Mean and Area Mean were removed in favor of Radius Mean, and Concave Points Mean was removed from shape descriptors. The third model, with these adjustments, achieved higher accuracy, sensitivity, specificity, and precision, with minimal overfitting and a smaller gap between training and validation performance, making it the best model. (See appendix, Fig. 12)

Comparison Analysis

Metric	Logistic Regression	Classification Tree
Accuracy (#correct)	211	214
Accuracy (%correct)	92.54385965	93.85964912
Specificity	0.917647059	0.965034965
Sensitivity (Recall)	0.93006993	0.894117647
Precision	0.95	0.938271605
F1 score	0.939929329	0.915662651
Success Class	1	1
Success Probability	0.5	0.5

When evaluating Logistic Regression and the Classification Tree for predicting malignant cases (class 1), priority must be given to metrics that emphasize minimizing false negatives, such as Sensitivity (Recall), Precision, and the F1 Score, as failing to identify malignancy can have severe consequences. While the Classification Tree has slightly higher accuracy (**93.86%** vs. **92.54%**) and better specificity (**0.965**), these metrics prioritize correctly identifying benign cases (class 0) rather than malignant ones. In this scenario, recall is far more critical. Logistic Regression achieves a **recall of 0.930**, outperforming the Classification Tree's **0.894**. Higher recall ensures more malignant cases are correctly identified, minimizing false negatives. Logistic Regression also has slightly higher precision (**0.95**) compared to the Classification Tree (**0.938**), meaning fewer false positives and greater reliability in malignant predictions. With a superior F1 Score (**0.9399** vs. **0.9157**), Logistic Regression balances precision and recall more effectively.

Hence, the Logistic Regression model is the better choice for detecting malignant cases due to its higher Sensitivity, Precision, and F1 Score, ensuring more malignant cases are identified with fewer false positives. While the Classification Tree performs well in overall accuracy and specificity, its lower recall makes it less suitable for this high-stakes medical application, where minimizing false negatives is essential.

Conclusion

In conclusion, both models provide a high accuracy score, which shows the model fits well with the data used. A limitation to our study is only using two methods of comparison for the data. Ethically, we think it is important to continue finding other models that increase overall accuracy and sensitivity. It is not acceptable to only accurately diagnose 92.5% of women with tumors. Our recommendation is to continue with other data analysis models to improve the overall results.

Appendix

Section 1

Data Exploration:

Descriptive Analysis

We performed a descriptive statistics analysis using Excel's "Data Analysis" function to better understand the dataset's structure and identify potential outliers. By selecting the "Descriptive Statistics" option, we input the dataset range and generated a summary of the statistics. (See appendix, Fig.1)

The descriptive analysis reveals significant variability in variables like Area, Perimeter, Texture, and Radius, as indicated by their high mean and standard deviation, suggesting a widespread in the data and potential outliers. Additionally, the skewness analysis shows that most variables, including Area, Compactness, Concavity, and Concave Points, are positively skewed, indicating that the dataset is not normally distributed.

Correlation Analysis

We conducted a correlation analysis to explore the relationships between the variables and identify potential redundancy in the dataset, aiming to enhance our model. Using Excel's "Data Analysis" function, we selected the "Correlation" option, entered the dataset, and generated the output. (See appendix, Fig.2)

Key findings include:

- Diagnosis Correlation: The dependent variable, Diagnosis, shows mostly positive correlations with the independent variables, with coefficients generally above 0.6, indicating a strong relationship. Higher values of these variables are associated with a higher likelihood of the tumor being malignant. Some variables, like Fractal Dimension Mean, Smoothness SE, and Symmetry SE, show weak negative correlations, suggesting that higher values may indicate benign tumors, though these relationships are weak and require stronger predictors for reliable conclusions.
- Mean and Worst Variable Correlation: We observed a high correlation between the mean and worst variables (e.g., Radius Mean vs. Radius Worst), indicating that higher mean values are linked to more severe or extreme cases. This redundancy could lead to overfitting, and adjusting for it would improve model performance.
- Inter-Variable Correlation: Strong correlations (greater than 0.9) were found between variables like Radius Mean, Perimeter Mean, and Area Mean, indicating redundancy in size-related information. Similarly, shape descriptors like Compactness Mean, Concavity Mean, and Concave Points Mean show strong inter-correlations (greater than 0.8), suggesting they convey similar data and could be grouped together to characterize tumor shape.

Scatter Plot Analysis

To examine the relationships between variables, we conducted a scatter plot analysis focusing on tumor characteristics like size, shape, and appearance in relation to diagnosis. This helped clarify the

high variability observed in variables like Area Mean, Perimeter Mean, Texture Mean, and Radius Mean. Using Excel's "Scatter Plot" function, we created data series for Benign and Malignant tumors, color-coded in Blue and Orange, respectively.

1. Tumor Size and Diagnosis: A strong positive linear relationship between Radius Mean and Perimeter Mean shows they increase proportionally. A non-linear positive relationship between Radius Mean and Area Mean indicates that larger radii lead to significantly larger areas. Tumors with higher values in these variables are more likely to be malignant, with variability driven by malignant tumors having higher values. The strong correlation between these size-related variables emphasizes their predictive value. (See appendix, Fig.3)
2. Tumor Shape and diagnosis: Compactness Mean and Concavity Mean exhibit a positive relationship, with malignant tumors clustering at higher values. However, the overlap with benign tumors limits their effectiveness as predictors. Similarly, Symmetry Mean and Concave Points Mean show a weak positive correlation, with benign tumors mainly at lower values and malignant tumors spread across a wider range, although the overlap at lower values reduces their distinguishing power. (See appendix, Fig.4)
3. Tumor Structure and Diagnosis: The relationship between Texture Mean and Smoothness Mean is weak, showing no clear pattern and indicating that these variables are independent. There is also no clear separation between benign and malignant tumors, suggesting these variables are not strong predictors of diagnosis. (See appendix, Fig.5)

   In conclusion, tumor size is the most significant predictor of tumor fatality, followed by shape. Tumor structure variables like texture and smoothness have limited predictive value.

Figure 1

Figure 2

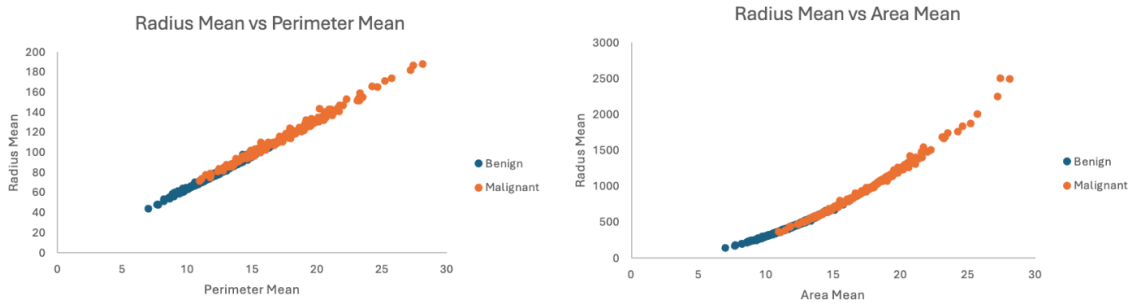

Figure 3

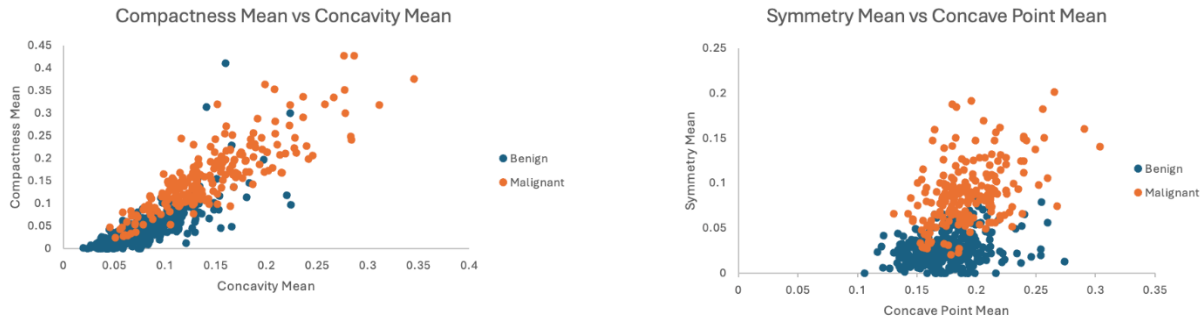

Figure 4

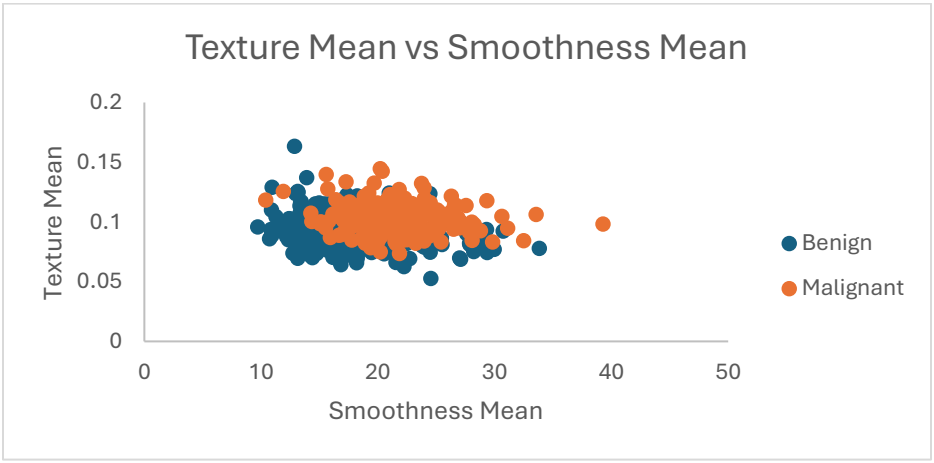

Figure 5

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	199	15
1	4	123

Error Report			
Class	# Case	# Error	% Error
0	214	15	7.009346
1	127	4	3.149606
Overall	341	19	5.571848

Metrics	
Metric	Value
Accuracy (#correct)	322
Accuracy (%correct)	94.42815
Specificity	0.929907
Sensitivity (Recall)	0.968504
Precision	0.891304
F1 score	0.928302
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	129	14
1	3	82

Error Report			
Class	# Case	# Error	% Error
0	143	14	9.79021
1	85	3	3.529412
Overall	228	17	7.45614

Metrics	
Metric	Value
Accuracy (#correct)	211
Accuracy (%correct)	92.54386
Specificity	0.902098
Sensitivity (Recall)	0.964706
Precision	0.854167
F1 score	0.906077
Success Class	1
Success Probability	0.5

Figure 6

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	199	15
1	4	123

Error Report			
Class	# Cases	# Errors	% Error
0	214	15	7.009346
1	127	4	3.149606
Overall	341	19	5.571848

Metrics	
Metric	Value
Accuracy (#correct)	322
Accuracy (%correct)	94.42815
Specificity	0.929907
Sensitivity (Recall)	0.968504
Precision	0.891304
F1 score	0.928302
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	129	14
1	3	82

Error Report			
Class	# Cases	# Errors	% Error
0	143	14	9.79021
1	85	3	3.529412
Overall	228	17	7.45614

Metrics	
Metric	Value
Accuracy (#correct)	211
Accuracy (%correct)	92.54386
Specificity	0.902098
Sensitivity (Recall)	0.964706
Precision	0.854167
F1 score	0.906077
Success Class	1
Success Probability	0.5

Figure 7

Trial	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Limit max # levels	7	7	7	7	7	7	7	7	7	7	10	10	20	20	6	6	10	10
Limit min # records in leaves	29	29	20	20	12	12	9	9	35	35	12	12	12	12	12	12	35	35
Accuracy (#correct)	322	211	324	211	331	214	331	214	317	207	331	214	331	214	331	214	317	207
Accuracy (%correct)	94.4282	92.5439	95.0147	92.5439	97.0674	93.8596	97.0674	93.8596	92.9619	90.7895	97.0674	93.8596	97.0674	93.8596	97.0674	93.8596	92.9619	90.7895
Specificity	0.92991	0.9021	0.93925	0.9021	0.98598	0.96503	0.98598	0.96503	0.98598	0.96503	0.98598	0.96503	0.98598	0.96503	0.98598	0.96503	0.98598	0.96503
Sensitivity (Recall)	0.9685	0.96471	0.9685	0.96471	0.94488	0.89412	0.94488	0.89412	0.83465	0.81176	0.94488	0.89412	0.94488	0.89412	0.94488	0.89412	0.83465	0.81176
Precision	0.8913	0.85417	0.90441	0.85417	0.97561	0.93827	0.97561	0.93827	0.97248	0.93243	0.97561	0.93827	0.97561	0.93827	0.97561	0.93827	0.97248	0.93243
F1 score	0.9283	0.90608	0.93536	0.90608	0.96	0.91566	0.96	0.91566	0.89831	0.86792	0.96	0.91566	0.96	0.91566	0.96	0.91566	0.89831	0.86792
Success Class	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Success Probability	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Figure 8

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	211	3
1	7	120

Error Report			
Class	# Cases	# Errors	% Error
0	214	3	1.401869159
1	127	7	5.511811024
Overall	341	10	2.93255132

Metrics	
Metric	Value
Accuracy (#correct)	331
Accuracy (%correct)	97.0674487
Specificity	0.98598131
Sensitivity (Recall)	0.94488189
Precision	0.97560976
F1 score	0.96
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	138	5
1	9	76

Error Report			
Class	# Cases	# Errors	% Error
0	143	5	3.496503497
1	85	9	10.58823529
Overall	228	14	6.140350877

Metrics	
Metric	Value
Accuracy (#correct)	214
Accuracy (%correct)	93.8596491
Specificity	0.96503497
Sensitivity (Recall)	0.89411765
Precision	0.9382716
F1 score	0.91566265
Success Class	1
Success Probability	0.5

Figure 9

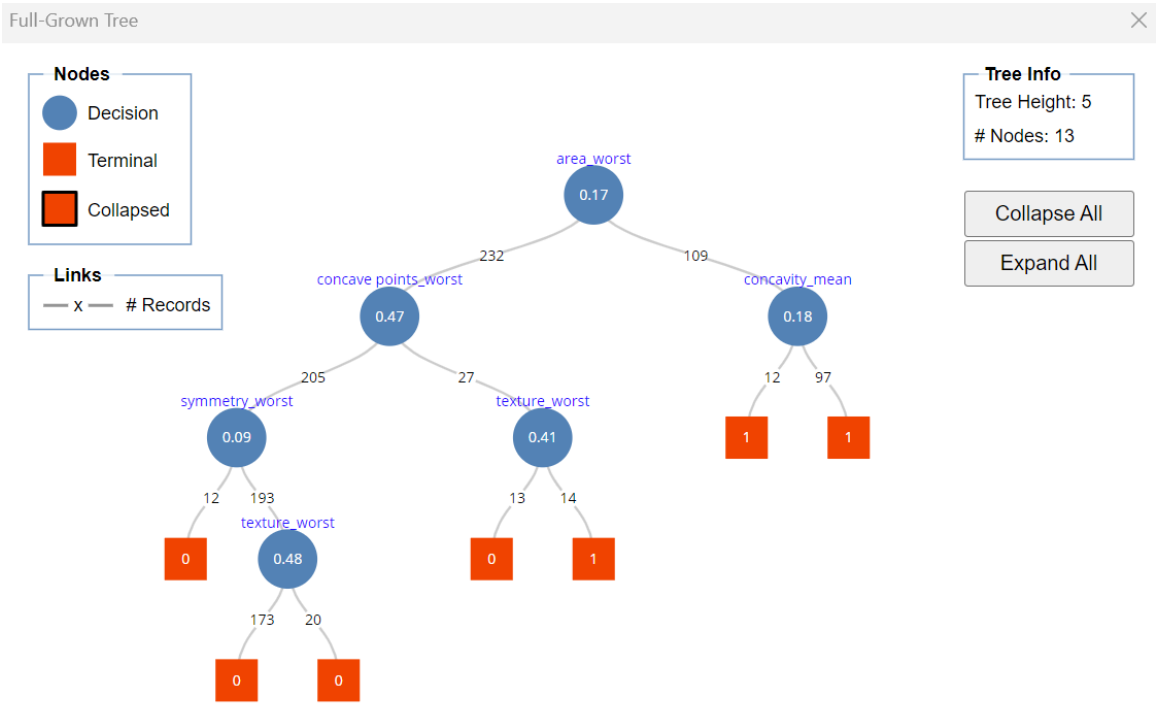

Figure 10

Metrics	Model 1		Model 2		Model 3	
Metric	Training	Validation	Training	Validation2	Training2	Validation4
Accuracy (#correct)	341	212	332	208	334	211
Accuracy (%correct)	100	92.9824561	97.36070381	91.22807018	97.94721408	92.54385965
Specificity	1	0.87058824	0.952755906	0.905882353	0.968503937	0.917647059
Sensitivity (Recall)	1	0.96503497	0.985981308	0.916083916	0.985981308	0.93006993
Precision	1	0.9261745	0.97235023	0.942446043	0.981395349	0.95
F1 score	1	0.94520548	0.979118329	0.929078014	0.983682984	0.939929329
Success Class	1	1	1	1	1	1
Success Probability	0.5	0.5	0.5	0.5	0.5	0.5