BREAST CANCER

WEKA

AUGUST 6, 2024

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
Bidholi, Dehradun

Detecting Breast Cancer using WEKA

Abstract –

Breast cancer is one of the serious global health problems and hence requires the diagnostic tools that can enhance its early detection. This abstract describes the use of WEKA, a machine learning toolkit in analysing and predicting breast cancer. WEKA is free and open-source software, a suite of machine learning, offering a wide array of algorithms and tools to perform classification, regression, and clustering tasks. With respect to breast cancer detection, WEKA mines clinical and diagnostic data—mammographic images and patient demographics—establishing the existence or risk of suffering from breast cancer. This approach aims at improving the accuracy and efficiency of diagnosis in cases where breast cancer is suspected, applying different machine learning algorithms in WEKA incorporating decision trees, support vector machines, and neural networks. WEKA integration into breast cancer research underlines its potentials with regard to supporting data-driven decision-making and increased diagnostic precision, consequently assuring more appropriate screening strategies and better patient outcomes.

Key Words –

Classification, Regression, Clustering, Mammographic images..

Introduction –

A breast cancer is a malignant NEOPLASM originating from the cells of the BREAST. It is the most frequently diagnosed cancer in women worldwide and it can also occur in men. These risk factors are due to age, genetic mutations, family history and lifestyle including alcohol consumption or obesity. Breast cancer signs and symptoms may include lump in the breast or armpit or sickened skin around it. Early detection is key and usually accomplished through clinical breast exams or self-exam. The course of action for your care will depend on what type of cancer you have, but may include surgery radiotherapy, chemotherapy hormone therapy or treatments. In recent years, the application of machine learning has emerged as a promising approach to enhance breast cancer detection and diagnosis. Through the use of WEKA, this report aims to highlight the benefits of incorporating machine learning techniques into breast cancer diagnostics, address existing challenges, and explore future research directions for optimizing detection methods and treatment approaches.

Problem Statement –

One of the challenges to breast cancer detection is the variability of the diagnostic imaging and the complexity involved in the interpretation of clinical data. Traditional methods, lead to false positives or negatives that affect timely diagnosis and treatment. The following report addresses the problem of how WEKA, should be effectively used to improve breast cancer detection. Specifically, this paper is concerned with the application of WEKA's algorithms in decision trees, support vector machines, and neural networks for the analysis of complex datasets with a view toward enhancing the diagnostic accuracy of early detection for the better clinical outcomes.

Theoretical Background –

Correlation coefficient:- It is a statistical measure of the strength of a linear relationship between two variables.

Mean Absolute Error:- It informs about the size of mistakes in a set of predictions, without considering their direction. It's measured as the average absolute difference between the predicted and actual values.

Root Mean square error:- It is one of the most commonly used measures for evaluating the quality of predictions.

Relative Error:- It can be defined as the ratio of absolute error to the size of measurement. The absolute error can be obtained simply by dividing the measured value by the absolute error.

Root relative squared error:- RRSE is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error.

Kappa statistic:- It is frequently used to test interrater reliability.

TP Rate:- TP rate as a measure of how good our model is at finding the positive cases it's supposed to detect.

FP Rate:- It measures the proportion of actual negative instances that are incorrectly classified as positive by the model.

MCC:- Matthews Correlation Coefficient is a performance metric used to evaluate the quality of binary classifications.

ROC rate:- The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across different threshold settings, providing insight into the trade-offs between sensitivity and specificity.

PRC area:- PRC is another important tool for evaluating the performance of classification models, particularly when dealing with imbalanced datasets. The PRC plots Precision against Recall for different threshold values, providing insights into the trade-offs between these two metrics.

1. Decision Stump Tree

Decision Stump is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature.

Table 1.1.1 Summary for Decision Stump Tree 10 Fold

Correctly Classified Instances	196	68.53%
Incorrectly Classified Instances	90	31.47%
Kappa statistic	0.2257	
Mean absolute error	0.3801	
Root mean squared error	0.4404	
Relative absolute error	90.85%	
Root relative squared error	96.34%	
Total Number of Instances	286	

Table 1.1.2 Accuracy measures for Decision Stump Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.796	0.576	0.766	0.796	0.78	0.226	0.588	0.742	no-recurrence-events
	0.424	0.204	0.468	0.424	0.444	0.226	0.588	0.361	recurrence-events
Weighted Avg.	0.685	0.466	0.677	0.685	0.681	0.226	0.588	0.629	

Table 1.2.1 Summary for Decision Stump Tree 5 Fold

Correctly Classified Instances	200	69.93%
Incorrectly Classified Instances	86	30.07%
Kappa statistic	0.2898	
Mean absolute error	0.3828	
Root mean squared error	0.4417	
Relative absolute error	91.48%	
Root relative squared error	96.64%	
Total Number of Instances	286	

Table 1.2.2 Accuracy measures for Decision Stump Tree 5 Fold

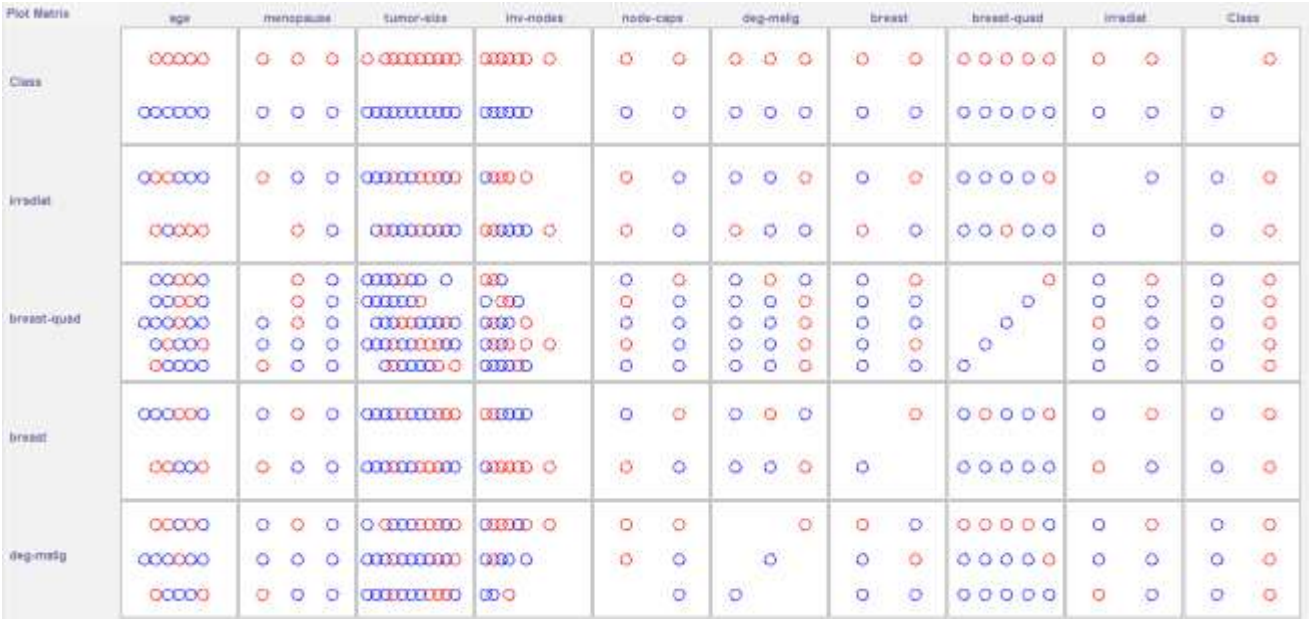
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.776	0.482	0.792	0.776	0.784	0.29	0.608	0.754	no-recurrence-events
	0.518	0.224	0.494	0.518	0.506	0.29	0.608	0.379	recurrence-events
Weighted Avg.	0.699	0.406	0.703	0.699	0.701	0.29	0.608	0.642	

Table 1.3.1 Summary for Decision Stump Tree 80% Split

Correctly Classified Instances	40	70.18%
Incorrectly Classified Instances	17	29.82%
Kappa statistic	0.2963	
Mean absolute error	0.3905	
Root mean squared error	0.4743	
Relative absolute error	88.26%	
Root relative squared error	96.76%	
Total Number of Instances	57	

Table 1.3.2 Accuracy measures for Decision Stump Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.889	0.619	0.711	0.889	0.79	0.319	0.635	0.702	no-recurrence-events
	0.381	0.111	0.667	0.381	0.485	0.319	0.635	0.482	recurrence-events
Weighted Avg.	0.702	0.432	0.695	0.702	0.678	0.319	0.635	0.621	



2. Hoeffding Tree

It is an incremental decision tree that is capable of learning from the data streams. The basic assumption about the data is that data is not changing over time helps in building a Hoeffding tree.

Table 2.1.1 Summary for Hoeffding Tree 10 Fold

Correctly Classified Instances	200	69.93%
Incorrectly Classified Instances	86	30.07%
Kappa statistic	0.2105	
Mean absolute error	0.36	
Root mean squared error	0.4693	
Relative absolute error	86.05%	
Root relative squared error	102.68%	
Total Number of Instances	286	

Table 2.1.2 Accuracy measures for Hoeffding Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.851	0.659	0.753	0.851	0.799	0.217	0.65	0.801	no-recurrence-events
	0.341	0.149	0.492	0.341	0.403	0.217	0.65	0.461	recurrence-events
Weighted Avg.	0.699	0.507	0.676	0.699	0.681	0.217	0.65	0.7	

Table 2.2.1 Summary for Hoeffding Tree 5 Fold

Correctly Classified Instances	208	72.73%
Incorrectly Classified Instances	78	27.27%
Kappa statistic	0.3096	
Mean absolute error	0.3282	
Root mean squared error	0.454	
Relative absolute error	78.42%	
Root relative squared error	99.35%	
Total Number of Instances	286	

Table 2.2.2 Accuracy measures for Hoeffding Tree 5 Fold

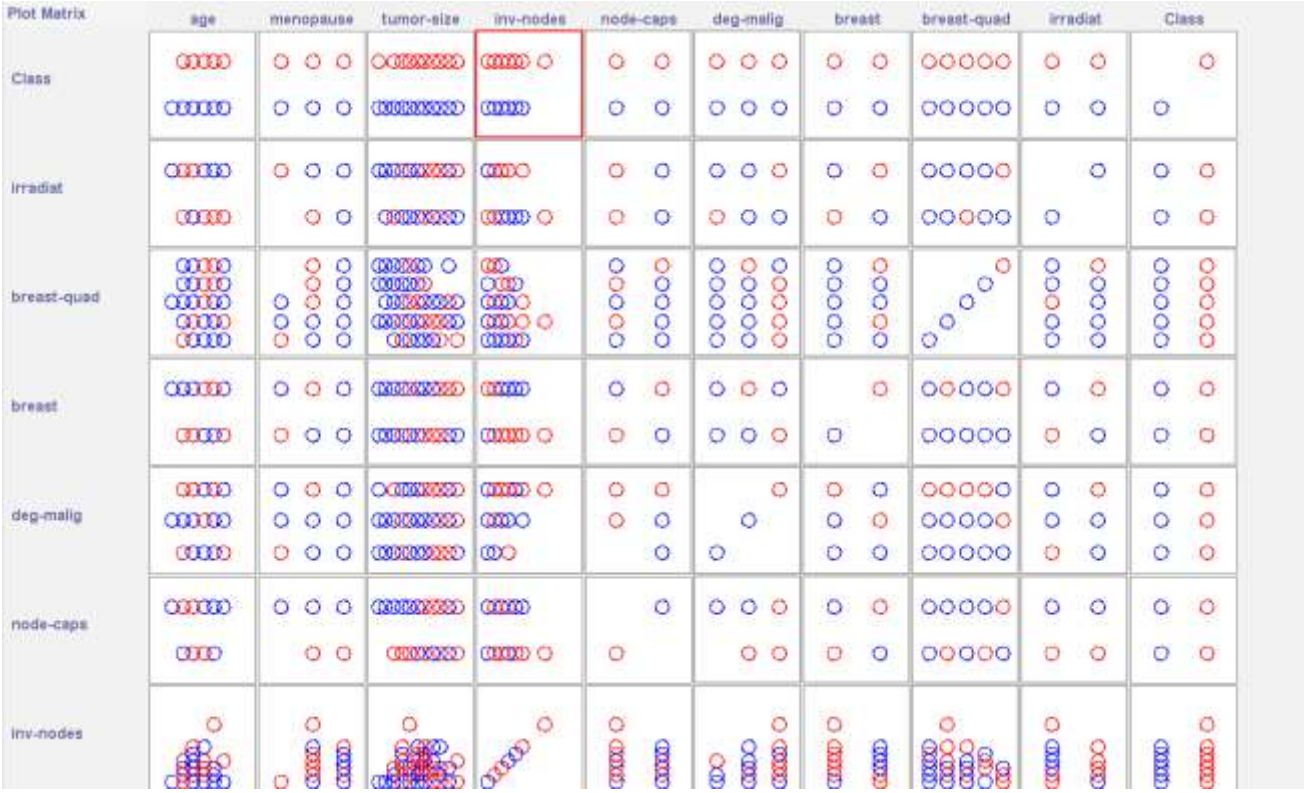
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.846	0.553	0.783	0.846	0.813	0.313	0.698	0.838	no-recurrence-events
	0.447	0.154	0.551	0.447	0.494	0.313	0.698	0.486	recurrence-events
Weighted Avg.	0.727	0.434	0.714	0.727	0.718	0.313	0.698	0.733	

Table 2.3.1 Summary for Hoeffding Tree 80% Split

Correctly Classified Instances	36	63.16%
Incorrectly Classified Instances	21	36.84%
Kappa statistic	0	
Mean absolute error	0.4425	
Root mean squared error	0.4902	
Relative absolute error	100.00%	
Root relative squared error	100.00%	
Total Number of Instances	57	

Table 2.3.2 Accuracy measures for Hoeffding Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1	1	0.632	1	0.774	?	0.5	0.632	no-recurrence-events
	0	0	?	0	?	?	0.5	0.368	recurrence-events
Weighted Avg.	0.632	0.632	?	0.632	?	?	0.5	0.535	



3. J48 Tree

It creates decision trees by recursively partitioning data based on attribute values. J48 employs information gain or gain ratio to select the best attribute for splitting.

Table 3.1.1 Summary for J48 Tree 10 Fold

Correctly Classified Instances	216	75.52%
Incorrectly Classified Instances	70	24.48%
Kappa statistic	0.2826	
Mean absolute error	0.3676	
Root mean squared error	0.4324	
Relative absolute error	87.86%	
Root relative squared error	94.61%	
Total Number of Instances	286	

Table 3.1.2 Accuracy measures for J48 Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.96	0.729	0.757	0.96	0.846	0.339	0.584	0.736	no-recurrence-events
	0.271	0.04	0.742	0.271	0.397	0.339	0.584	0.436	recurrence-events
Weighted Avg.	0.775	0.524	0.752	0.755	0.713	0.339	0.584	0.647	

Table 3.2.1 Summary for J48 Tree 5 Fold

Correctly Classified Instances	212	74.13%
Incorrectly Classified Instances	74	25.87%
Kappa statistic	0.2288	
Mean absolute error	0.3726	
Root mean squared error	0.4435	
Relative absolute error	89.04%	
Root relative squared error	97.04%	
Total Number of Instances	286	

Table 3.2.2 Accuracy measures for J48 Tree 5 Fold

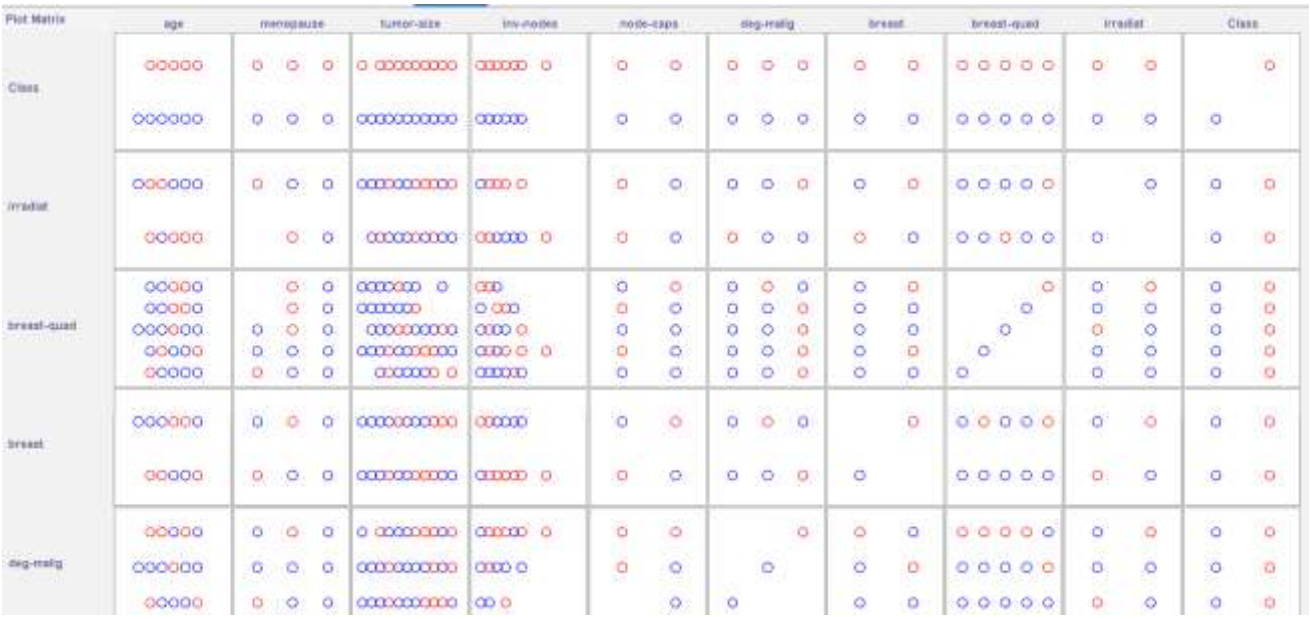
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.96	0.776	0.745	0.96	0.839	0.287	0.582	0.728	no-recurrence-events
	0.224	0.04	0.704	0.224	0.339	0.287	0.582	0.444	recurrence-events
Weighted Avg.	0.741	0.558	0.733	0.741	0.691	0.287	0.582	0.643	

Table 3.3.1 Summary for J48 Tree 80% Split

Correctly Classified Instances	41	71.93%
Incorrectly Classified Instances	16	28.07%
Kappa statistic	0.2995	
Mean absolute error	0.3707	
Root mean squared error	0.4619	
Relative absolute error	83.77%	
Root relative squared error	94.24%	
Total Number of Instances	57	

Table 3.3.2 Accuracy measures for J48 Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.972	0.714	0.7	0.972	0.814	0.379	0.628	0.711	no-recurrence-events
	0.286	0.028	0.857	0.286	0.429	0.379	0.628	0.522	recurrence-events
Weighted Avg.	0.719	0.461	0.758	0.719	0.672	0.379	0.628	0.641	



4. LMT Tree

It is a classification model with an associated supervised training algorithm that combines logistic regression (LR) and decision tree learning.

Table 4.1.1 Summary for LMT Tree 10 Fold

Correctly Classified Instances	215	75.17%
Incorrectly Classified Instances	71	24.83%
Kappa statistic	0.3042	
Mean absolute error	0.3589	
Root mean squared error	0.4291	
Relative absolute error	85.77%	
Root relative squared error	93.88%	
Total Number of Instances	286	

Table 4.1.2 Accuracy measures for LMT Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.93	0.671	0.766	0.93	0.84	0.335	0.675	0.797	no-recurrence-events
	0.329	0.07	0.667	0.329	0.441	0.335	0.675	0.49	recurrence-events
Weighted Avg.	0.752	0.492	0.737	0.752	0.722	0.335	0.675	0.706	

Table 4.2.1 Summary for LMT Tree 5 Fold

Correctly Classified Instances	217	75.87%
Incorrectly Classified Instances	69	24.12%
Kappa statistic	0.3072	
Mean absolute error	0.3583	
Root mean squared error	0.4291	
Relative absolute error	85.63%	
Root relative squared error	93.23%	
Total Number of Instances	286	

Table 4.2.2 Accuracy measures for LMT Tree 5 Fold

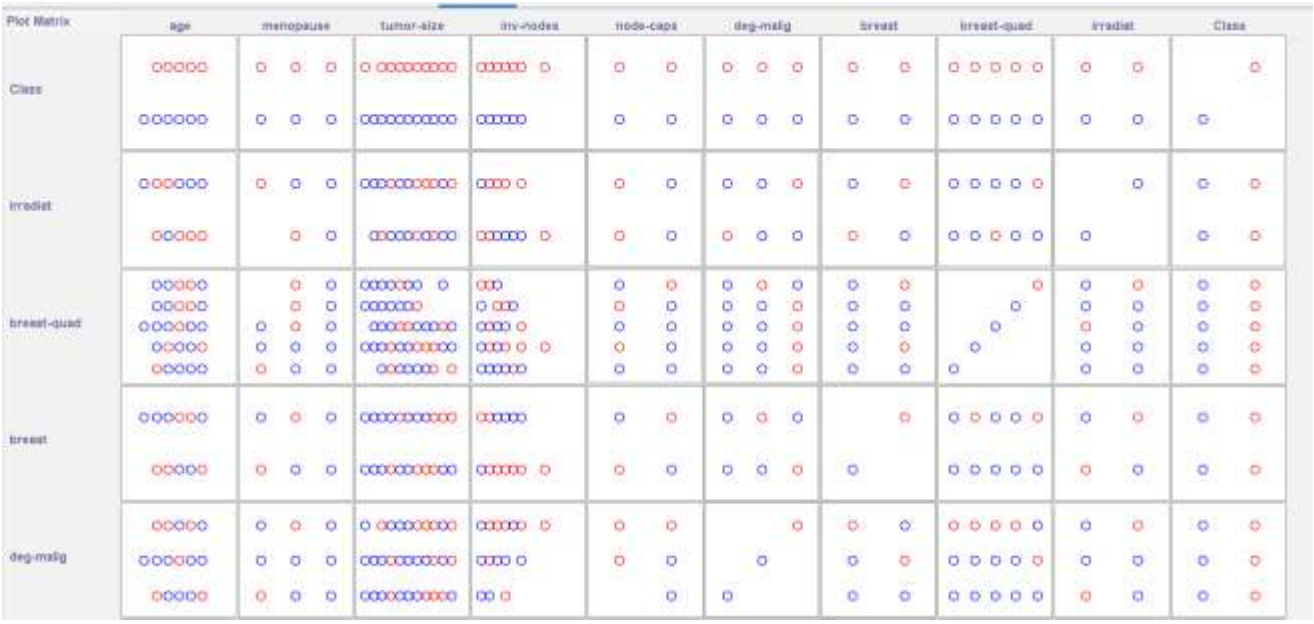
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.95	0.694	0.764	0.95	0.847	0.353	0.676	0.798	no-recurrence-events
	0.306	0.05	0.722	0.306	0.43	0.353	0.676	0.484	recurrence-events
Weighted Avg.	0.759	0.503	0.752	0.759	0.723	0.353	0.676	0.705	

Table 4.3.1 Summary for LMT Tree 80% Split

Correctly Classified Instances	42	73.68%
Incorrectly Classified Instances	15	26.32%
Kappa statistic	0.3508	
Mean absolute error	0.3694	
Root mean squared error	0.4548	
Relative absolute error	83.49%	
Root relative squared error	92.78%	
Total Number of Instances	57	

Table 4.3.2 Accuracy measures for LMT Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.972	0.667	0.714	0.972	0.824	0.424	0.669	0.723	no-recurrence-events
	0.333	0.028	0.875	0.333	0.483	0.424	0.669	0.577	recurrence-events
Weighted Avg.	0.737	0.431	0.773	0.737	0.698	0.424	0.669	0.669	



5. Random Forest Tree

It is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.”

Table 5.1.1 Summary for Random Forest Tree 10 Fold

Correctly Classified Instances	199	69.58%
Incorrectly Classified Instances	87	30.42%
Kappa statistic	0.1736	
Mean absolute error	0.3727	
Root mean squared error	0.4613	
Relative absolute error	89.09%	
Root relative squared error	100.92%	
Total Number of Instances	286	

Table 5.1.2 Accuracy measures for Random Forest Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.871	0.718	0.742	0.871	0.801	0.184	0.634	0.798	no-recurrence-events
	0.282	0.129	0.48	0.282	0.356	0.184	0.634	0.409	recurrence-events
Weighted Avg.	0.696	0.543	0.664	0.696	0.669	0.184	0.634	0.682	

Table 5.2.1 Summary for Random Forest Tree 5 Fold

Correctly Classified Instances	193	67.48%
Incorrectly Classified Instances	93	32.52%
Kappa statistic	0.1494	
Mean absolute error	0.3717	
Root mean squared error	0.4652	
Relative absolute error	88.82%	
Root relative squared error	101.79%	
Total Number of Instances	286	

Table 5.2.2 Accuracy measures for Random Forest Tree 5 Fold

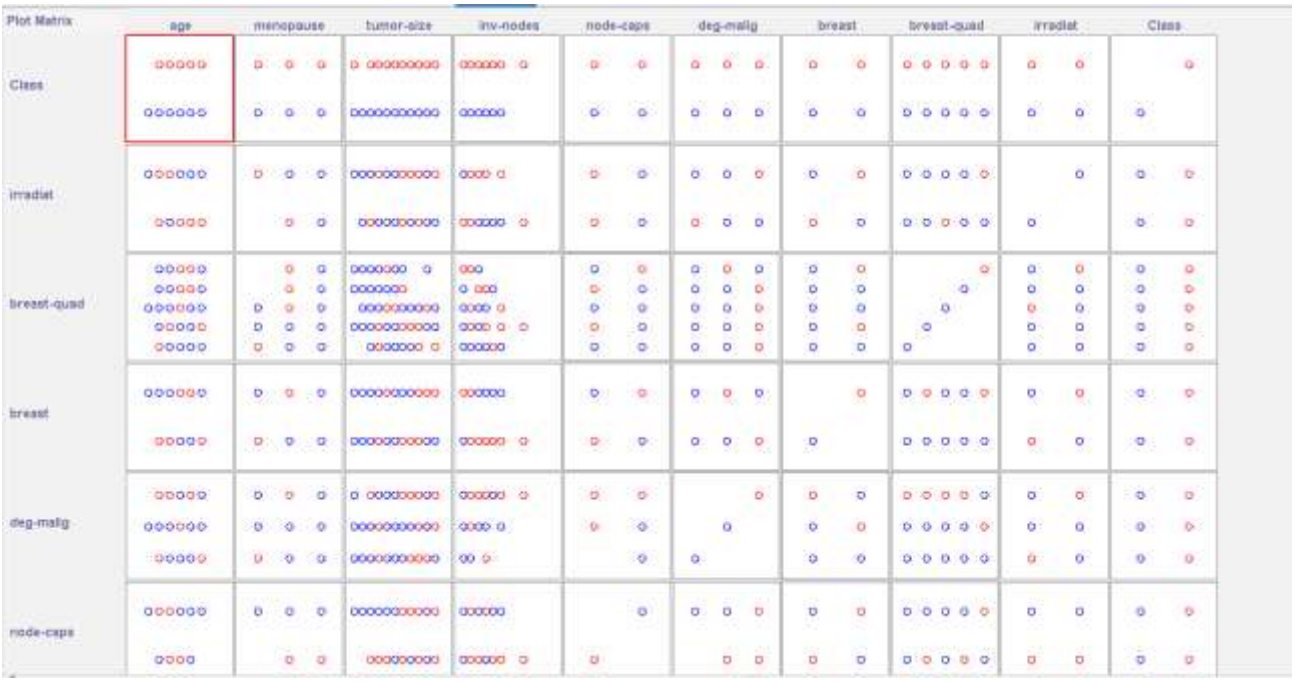
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.831	0.694	0.739	0.831	0.782	0.153	0.636	0.808	no-recurrence-events
	0.306	0.169	0.433	0.306	0.359	0.153	0.636	0.408	recurrence-events
Weighted Avg.	0.675	0.538	0.648	0.675	0.656	0.153	0.636	0.689	

Table 5.3.1 Summary for Random Forest Tree 80% Split

Correctly Classified Instances	41	71.93%
Incorrectly Classified Instances	16	28.07%
Kappa statistic	0.3304	
Mean absolute error	0.3632	
Root mean squared error	0.4661	
Relative absolute error	82.08%	
Root relative squared error	95.08%	
Total Number of Instances	57	

Table 5.3.2 Accuracy measures for Random Forest Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.917	0.619	0.717	0.917	0.805	0.364	0.68	0.778	no-recurrence-events
	0.381	0.083	0.727	0.381	0.5	0.364	0.68	0.621	recurrence-events
Weighted Avg.	0.719	0.422	0.721	0.719	0.693	0.364	0.68	0.721	



6. Random Tree

It is a supervised Classifier; it is an ensemble learning algorithm that generates many individual learners. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables.

Table 6.1.1 Summary for Random Tree 10 Fold

Correctly Classified Instances	191	66.78%
Incorrectly Classified Instances	95	33.22%
Kappa statistic	0.1855	
Mean absolute error	0.3533	
Root mean squared error	0.5699	
Relative absolute error	84.44%	
Root relative squared error	124.68%	
Total Number of Instances	286	

Table 6.1.2 Accuracy measures for Random Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.781	0.6	0.755	0.781	0.768	0.186	0.588	0.746	no-recurrence-events
	0.4	0.219	0.436	0.4	0.417	0.186	0.588	0.351	recurrence-events
Weighted Avg.	0.668	0.487	0.66	0.668	0.664	0.186	0.588	0.629	

Table 6.2.1 Summary for Random Tree 5 Fold

Correctly Classified Instances	187	65.38%
Incorrectly Classified Instances	99	34.62%
Kappa statistic	0.1012	
Mean absolute error	0.3548	
Root mean squared error	0.574	
Relative absolute error	84.79%	
Root relative squared error	125.60%	
Total Number of Instances	286	

Table 6.2.2 Accuracy measures for Random Tree 5 Fold

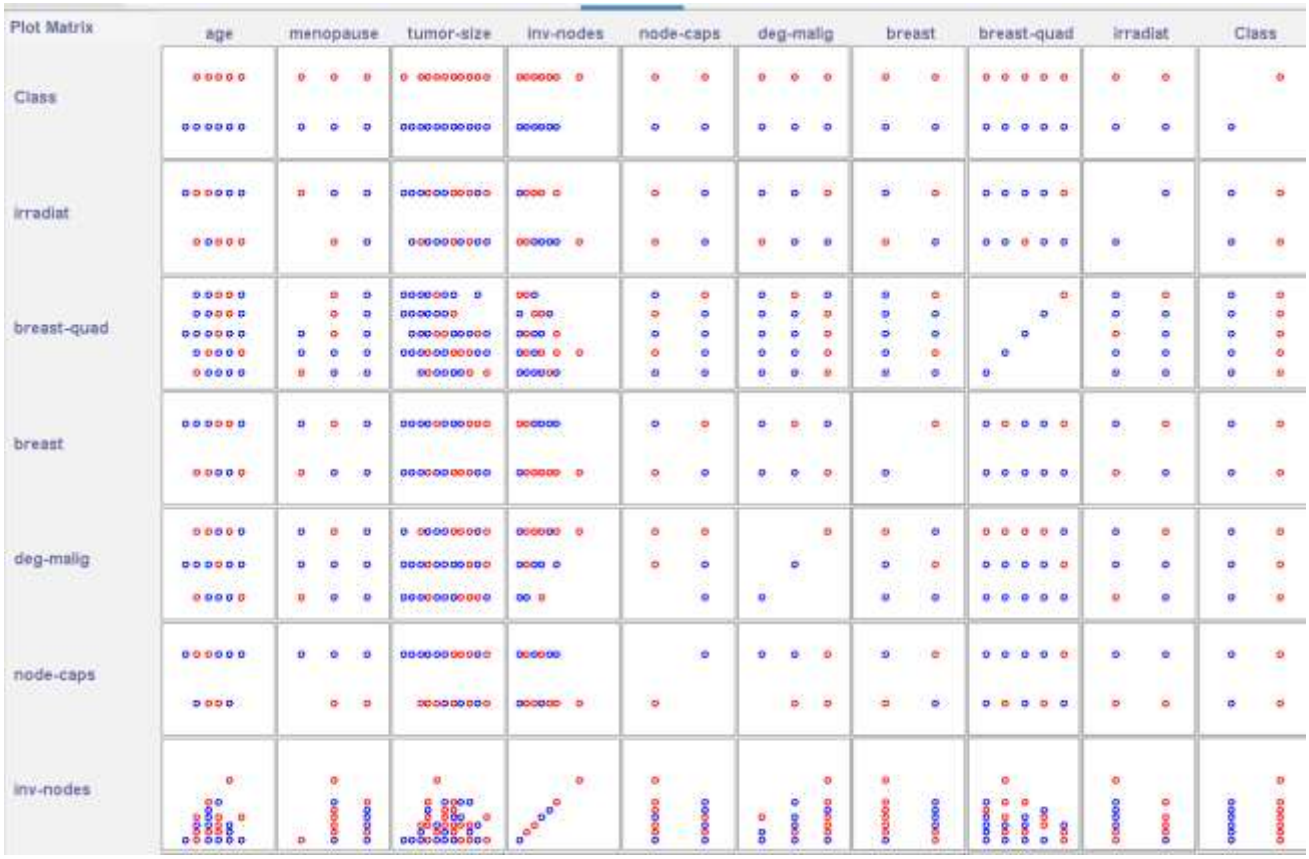
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.811	0.718	0.728	0.811	0.767	0.103	0.552	0.725	no-recurrence-events
	0.282	0.189	0.387	0.282	0.327	0.103	0.552	0.336	recurrence-events
Weighted Avg.	0.654	0.561	0.626	0.654	0.636	0.103	0.552	0.61	

Table 6.3.1 Summary for Random Tree 80% Split

Correctly Classified Instances	39	68.42%
Incorrectly Classified Instances	18	31.58%
Kappa statistic	0.212	
Mean absolute error	0.3005	
Root mean squared error	0.5123	
Relative absolute error	67.92%	
Root relative squared error	104.52%	
Total Number of Instances	57	

Table 6.3.2 Accuracy measures for Random Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.944	0.762	0.68	0.944	0.791	0.268	0.706	0.756	no-recurrence-events
	0.238	0.056	0.714	0.238	0.357	0.268	0.706	0.554	recurrence-events
Weighted Avg.	0.684	0.502	0.693	0.684	0.631	0.268	0.706	0.681	



7. REP Tree

It is a fast decision tree learner that builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning algorithm.

Table 7.1.1 Summary for REP Tree 10 Fold

Correctly Classified Instances	202	70.63%
Incorrectly Classified Instances	84	29.37%
Kappa statistic	0.1601	
Mean absolute error	0.3797	
Root mean squared error	0.4652	
Relative absolute error	90.74%	
Root relative squared error	101.78%	
Total Number of Instances	286	

Table 7.1.2 Accuracy measures for REP Tree 10 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.91	0.776	0.735	0.91	0.813	0.182	0.621	0.773	no-recurrence-events
	0.224	0.09	0.514	0.224	0.311	0.182	0.621	0.398	recurrence-events
Weighted Avg.	0.706	0.572	0.669	0.706	0.664	0.182	0.621	0.661	

Table 7.2.1 Summary for REP Tree 5 Fold

Correctly Classified Instances	193	67.48%
Incorrectly Classified Instances	93	32.52%
Kappa statistic	0.0738	
Mean absolute error	0.4016	
Root mean squared error	0.4724	
Relative absolute error	95.97%	
Root relative squared error	103.37%	
Total Number of Instances	286	

Table 7.2.2 Accuracy measures for REP Tree 5 Fold

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.886	0.824	0.718	0.886	0.793	0.084	0.545	0.722	no-recurrence-events
	0.176	0.114	0.395	0.176	0.244	0.084	0.545	0.358	recurrence-events
Weighted Avg.	0.675	0.613	0.622	0.675	0.63	0.084	0.545	0.614	

Table 7.3.1 Summary for REP Tree 80% Split

Correctly Classified Instances	35	61.40%
Incorrectly Classified Instances	22	38.60%
Kappa statistic	0.0793	
Mean absolute error	0.4205	
Root mean squared error	0.4941	
Relative absolute error	95.04%	
Root relative squared error	100.80%	
Total Number of Instances	57	

Table 7.3.2 Accuracy measures for REP Tree 80% Split

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.833	0.762	0.652	0.833	0.732	0.087	0.628	0.714	no-recurrence-events
	0.238	0.167	0.455	0.238	0.312	0.087	0.628	0.434	recurrence-events
Weighted Avg.	0.614	0.543	0.579	0.614	0.577	0.087	0.628	0.611