Machine Learning Techniques Applied to the Classification of Breast Cancer Recurrence

Diego P. Ardila[1]

[1]Departamento de ELE – Pontifícia Universidade Católica do Río de Janeiro (PUC-Rio)
Caixa Postal 38097 – Río de Janeiro – RJ – Brazil

diego.paez@aluno.puc-rio.br

Abstract. *Breast cancer affected 2.3 million women in 2020, resulting in 685,000 deaths worldwide. Death from breast cancer is mainly associated with metastasis and relapse. This work aims to analyze data corresponding to patients diagnosed with breast cancer, apply data mining to predict disease recurrence, and compare the performance of machine learning techniques in breast cancer relapse classification.*

Resumo. *O câncer de mama afetou 2,3 milhões de mulheres em 2020, resultando em 685.000 mortes em todo o mundo. A morte por câncer de mama está principalmente associada a metástases e recaídas. Este trabalho visa analisar dados correspondentes a pacientes diagnosticadas com câncer de mama, aplicar mineração de dados para prever a recorrência de doenças e comparar o desempenho de técnicas de aprendizagem de máquinas na classificação de recidivas de câncer de mama.*

1. Introduction

Breast cancer affected 2.3 million women in 2020 and caused 685,000 deaths worldwide. Consequently, according to information provided by the World Health Organization, it is the most frequent malignant pathology among women. Several researchers consider early detection and prediction the best alternative to fight against this highly invasive malignant pathology [1]. Death from breast cancer is mainly associated with metastasis and relapse. Metastatic relapse can occur months to years after the initial diagnosis and treatment of breast cancer.

Therefore, for researchers using data mining approaches, predicting breast cancer recurrence is a significant challenge. An essential aspect of evaluating breast cancer behavior is its recurrence, which is adequately related to mortality. Despite its relevance, it is rarely recorded in large part of breast cancer datasets, which hampers research in its prediction.

Several data mining techniques have been used in the literature investigated for breast cancer classification. Kumar et al. [2] performed technical comparisons to predict malignant and benign breast cancer. Temesgen Abera Asfaw [3] demonstrated that logical regression (LR) has the best classification accuracy, 96.93%, for detecting breast cancer using the UCI Wisconsin breast cancer dataset. Another research [4] studied LR in breast cancer detection. They concluded that using a β-weighting factor to the existing logistic function significantly improves accuracy, sensitivity, and specificity.

2. Methodology

This project proposes four configurations using different data preprocessing techniques to analyze and compare the performance of Machine Learning (ML) models applied to recurrence classification in breast cancer.

Figure 1 describes the methodological process employed for the development of the project. First, there is a General Pre-Processing block (GPP) of the data applied to the dataset. Then, configurations MC-1, MC-2, MC-3, and MC-4 defined for the evaluation of the models follow a line in which different operations are performed on the data before the training models.

The MC-1 configuration uses the raw data to perform the training, testing, and validation of the Logistic Regression (LR) [5], Naive Bayes (NB) [6], Support Vector Machine (SVM) [7], [8] and K-Nearest Neighbors (KNN) [9] models. The MC-2 configuration applies the GPP on the dataset before executing the training, testing, and validation block. On the other hand, the MC-3 and MC-4 configurations use the data resulting from the GPP, to which the Principal Component Analysis (PCA) [10] feature extraction technique is applied to select the most relevant attributes within the dataset. The result of this processing is involved in two ways; MC-3 implements the same training, testing, and validation block used by the previous configurations. MC-4 applies the SMOTE oversample technique [11]-[13] on the training data to balance the target class and execute the training, testing, and validation blocks.

Finally, the training, testing, and validation process is performed in two iterations. In the first iteration, the default parameters (Penalty, C, Solver, Var Smoothing, Gamma, Kernel, Leaf Size, n_neighbors, distances) of the LR, NB, SVM, and KNN models of the SkLearn library [14] are applied. In the second iteration, the parameters of the different models are optimized, thus evaluating the performance of the models in two instantiations.

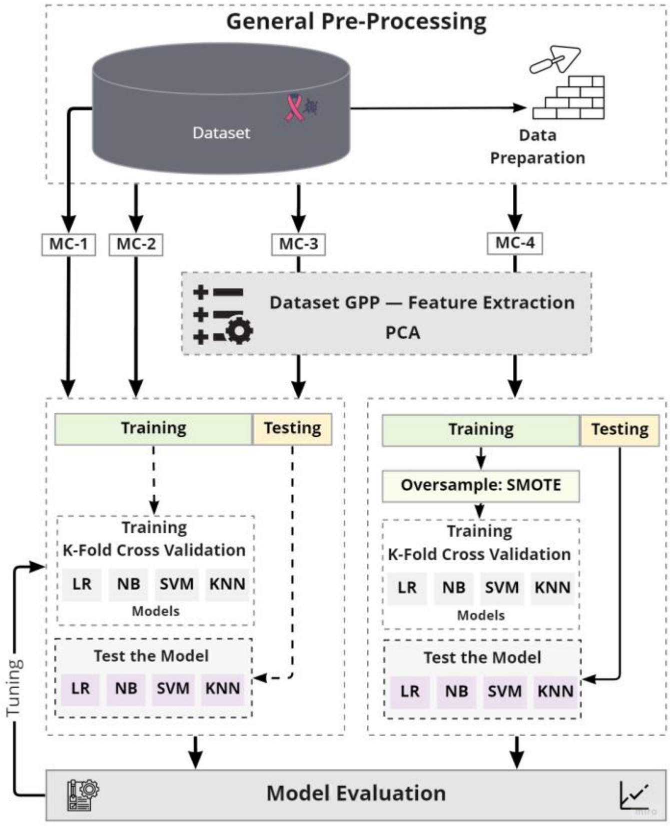

Figure 1. Description of the Methodological Process.

2.1. Dataset Description

In this study, two versions of the same dataset were used; the difference is the number of target classes (2 and 4) each had. This dataset contains 344 instances with 19 attributes, distributed in four categorical with integer coding, eight categorical with binary coding, and seven continuous. The target variable is a recurrence, and for versions 1 and 2 of the dataset, it is divided into (no recurrence, with recurrence) and (no recurrence, early recurrence, medium recurrence, late recurrence). Figure 2 shows the name of each attribute and its distribution.

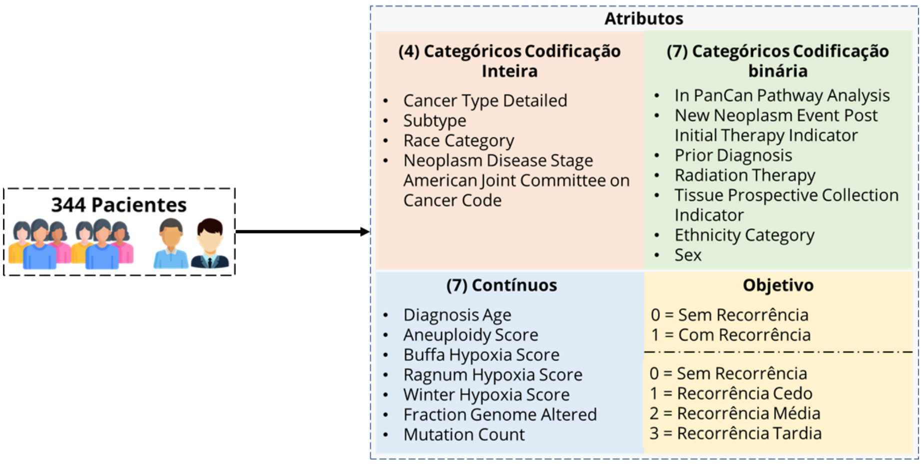

Figure 2. Attribute Distribution Datasets

2.2. Preprocessing and Exploratory Data Analysis

General data preprocessing was applied to prepare the dataset for the four defined configurations (MC-1 to MC-4). No missing data (NaN) was found in the data exploration process, and all attributes were already numerically or binary categorized. The scatter diagram presented in Figure 3 shows that the sex attribute has a reduced number (4) of instances associated with the male sex compared to the female sex (340); for this reason, these instances are eliminated. Additionally, it is possible to observe that in both datasets, there is an imbalance in the target classes. Dataset 2 has four classes which are reduced to three by applying the K-Means method [15], [16], thus achieving a better data distribution.

As part of the performance evaluation process of ML models with different configurations, a new dataset was created in which binary-coded attributes were not altered. One-Hot coding [17] was applied to integer-coded categorical features, and continuous characteristics were standardized.

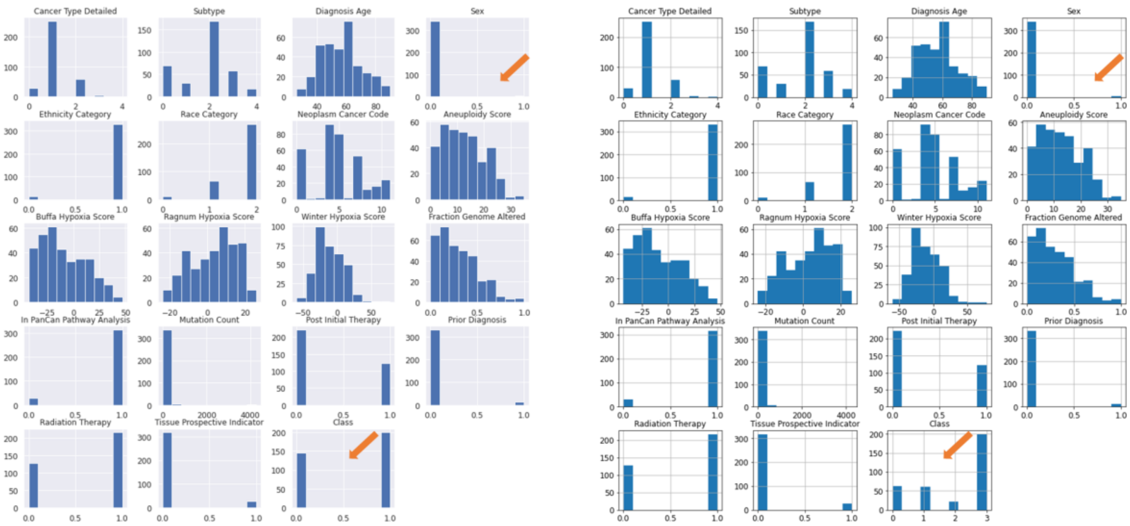

Figure 3. Attribute Scatter Diagrams

2.3. Model Evaluation

The evaluation of the models was performed in two phases. In the first phase, the default parameters defined in SkLearn were used to train the models: LR (penalty='l2', C=1.0, solver='lbfgs'), NB (var_smoothing=1e-09), SVM (C=1.0, kernel='rbf', gamma='scale'), KNN (n_neighbors=5, leaf_size=30, p=2). Data were separated into training and testing using a ratio of 80%-20%. The K-Fold Cross-Validation technique [18] was used to compare the models' performance with each other, and the performance metrics described in Table 1 were used to evaluate the classification ability of each model.

Table 1. Performance Metrics

Performance Metrics	Description
Confusion matrix	True Positives, true negatives, false positives, false negatives.
Accuracy	Overall model performance
Precision	What proportion of predicted Positives is actually Positive?
Recall	What proportion of real positives are correctly classified?
F1-Score	Combine precision and recall into a single number. Uses the harmonic mean.
AUC Roc Score	Area under the ROC curve, provides a single score and can be used to compare models.

Once the first phase of evaluation was completed, in the second phase, the model parameters were optimized using the Grid Search Cross-Validation technique [19], [20] with the search ranges defined in Table 2. This way, the model training process was repeated, and the performance metrics were calculated with the optimized parameters.

Table 2. Parameters Optimization Grid

Parameters Grid - Logistic Regresion	
Penalty	['l1','l2', "elasticnet"]
C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
Solver	['newton-cg', 'lbfgs', 'liblinear', "saga"]
Parameters Grid - Naive Bayes	
var smoothing	np.logspace(0,-9, num=1000) [1,…,1e-9]
Parameters Grid - SVM	
C	[0.1, 1, 10, 100, 1000]
Gamma	[1, 0.1, 0.01, 0.001, 0.0001]
Kernel	['rbf']
Parameters Grid - KNN	
Leaf Size	range(1,50)
n_neighbors	range(1,30)
distance	manhattan, euclidean

3. Results

Tables 3-4 present the results of the first phase of model evaluation; the tables are divided into the results of the four configurations defined in the methodology (MC-1 to MC-4) and show the performance metrics of each of the models evaluated. The best result of each of the experiments performed is highlighted in red. Consequently, Tables 5-6 show the results of the second phase of model evaluation.

Table 3. Performances Metrics Dataset 1 – Two Classes

Resultados MC-1 Dataset 1 – Duas Classes.

MC-1			Predicted	
			nonRec	Rec
LR	Actual	nonRec	23	1
		Rec	0	45

Precision: 0.986 **Accuracy: 0.986**
Recall: 0.985 **AUC ROC: 0.979**
F1-Score: 0.986

MC-1			Predicted	
			nonRec	Rec
NB	Actual	nonRec	23	1
		Rec	1	44

Precision: 0.971 Accuracy: 0.971
Recall: 0.971 AUC ROC: 0.968
F1-Score: 0.971

MC-1			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	0	24
		Rec	0	45

Precision: 0.425 Accuracy: 0.652
Recall: 0.515 AUC ROC: 0.500
F1-Score: 0.652

MC-1			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	9	15
		Rec	14	31

Precision: 0.576 Accuracy: 0.580
Recall: 0.578 AUC ROC: 0.532
F1-Score: 0.580

Resultados MC-2 Dataset 1 – Duas Classes.

MC-2			Predicted	
			nonRec	Rec
LR	Actual	NonRecur	24	0
		Recur	1	44

Precision: 0.986 **Accuracy: 0.986**
Recall: 0.985 **AUC ROC: 0.989**
F1-Score: 0.986

MC-2			Predicted	
			nonRec	Rec
NB	Actual	nonRec	23	1
		Rec	2	43

Precision: 0.957 Accuracy: 0.957
Recall: 0.957 AUC ROC: 0.957
F1-Score: 0.957

MC-2			Predicted	
			nonRec	Rec
SVM	Actual	NonRecur	22	2
		Recur	0	45

Precision: 0.972 Accuracy: 0.971
Recall: 0.971 AUC ROC: 0.958
F1-Score: 0.971

MC-2			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	13	11
		Rec	8	45

Precision: 0.718 Accuracy: 0.725
Recall: 0.720 AUC ROC: 0.682
F1-Score: 0.725

Resultados MC-3 Dataset 1 – Duas Classes.

MC-3			Predicted	
			nonRec	Rec
LR	Actual	nonRec	24	0
		Rec	1	44

Precision: 0.986 **Accuracy: 0.986**
Recall: 0.986 **AUC ROC: 0.989**
F1-Score: 0.986

MC-3			Predicted	
			nonRec	Rec
NB	Actual	nonRec	18	6
		Rec	2	43

Precision: 0.885 Accuracy: 0.884
Recall: 0.881 AUC ROC: 0.853
F1-Score: 0.884

MC-3			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	22	2
		Rec	0	45

Precision: 0.972 Accuracy: 0.971
Recall: 0.971 AUC ROC: 0.958
F1-Score: 0.971

MC-3			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	16	8
		Rec	7	38

Precision: 0.783 Accuracy: 0.783
Recall: 0.781 AUC ROC: 0.756
F1-Score: 0.783

Resultados MC-4 Dataset 1 – Duas Classes.

MC-4			Predicted	
			nonRec	Rec
LR	Actual	nonRec	24	0
		Rec	1	44

Precision: 0.986 Accuracy: 0.986
Recall: 0.986 AUC ROC: 0.989
F1-Score: 0.986

MC-4			Predicted	
			nonRec	Rec
NB	Actual	nonRec	19	5
		Rec	2	43

Precision: 0.899 Accuracy: 0.899
Recall: 0.887 AUC ROC: 0.874
F1-Score: 0.899

MC-4			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	24	0
		Rec	0	45

Precision: 1 **Accuracy: 1**
Recall: 1 **AUC ROC: 1**
F1-Score: 1

MC-4			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	18	6
		Rec	10	35

Precision: 0.780 Accuracy: 0.768
Recall: 0.772 AUC ROC: 0.764
F1-Score: 0.768

Table 4. Performances Metrics Dataset 2 – Three Classes

Resultados MC-1 Dataset 2 – Tres Classes.

MC-1			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	5	1
		earlyRec	1	11	1
		lateRec	0	0	45

Precision: 0.887 **Accuracy: 0.884**
Recall: 0.875 **AUC ROC: 0.883**
F1-Score: 0.884

MC-1			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	4	7	0
		earlyRec	0	12	1
		lateRec	1	0	44

Precision: 0.884 Accuracy: 0.870
Recall: 0.859 AUC ROC: 0.960
F1-Score: 0.870

MC-1			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	0	0	11
		earlyRec	0	0	13
		lateRec	0	0	45

Precision: 0.425 Accuracy: 0.652
Recall: 0.515 AUC ROC: 0.553
F1-Score: 0.652

MC-1			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	5	2	4
		earlyRec	1	1	11
		lateRec	4	10	31

Precision: 0.534 Accuracy: 0.536
Recall: 0.535 AUC ROC: 0.543
F1-Score: 0.536

Resultados MC-2 Dataset 2 – Tres Classes.

MC-2			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	6	0
		earlyRec	1	10	2
		lateRec	0	2	43

Precision: 0.861 Accuracy: 0.841
Recall: 0.839 AUC ROC: 0.943
F1-Score: 0.841

MC-2			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	5	6	0
		earlyRec	2	10	1
		lateRec	2	1	42

Precision: 0.836 Accuracy: 0.826
Recall: 0.828 AUC ROC: 0.838
F1-Score: 0.826

MC-2			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	4	6	1
		earlyRec	1	11	1
		lateRec	0	0	45

Precision: 0.874 **Accuracy: 0.870**
Recall: 0.856 **AUC ROC: 0.956**
F1-Score: 0.870

MC-2			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	3	3	5
		earlyRec	1	6	6
		lateRec	2	6	37

Precision: 0.658 Accuracy: 0.667
Recall: 0.656 AUC ROC: 0.727
F1-Score: 0.667

Resultados MC-3 Dataset 2 – Tres Classes.

MC-3			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	4	7	0
		earlyRec	2	9	2
		lateRec	0	1	44

Precision: 0.830 **Accuracy: 0.826**
Recall: 0.819 **AUC ROC: 0.955**
F1-Score: 0.826

MC-3			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	1	7	3
		earlyRec	0	10	3
		lateRec	2	0	43

Precision: 0.736 Accuracy: 0.783
Recall: 0.745 AUC ROC: 0.911
F1-Score: 0.783

MC-3			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	4	6	1
		earlyRec	1	11	1
		lateRec	0	0	45

Precision: 0.874 Accuracy: 0.696
Recall: 0.856 AUC ROC: 0.949
F1-Score: 0.870

MC-3			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	2	4	5
		earlyRec	2	8	3
		lateRec	2	5	38

Precision: 0.681 Accuracy: 0.696
Recall: 0.683 AUC ROC: 0.740
F1-Score: 0.696

Resultados MC-4 Dataset 2 – Tres Classes.

MC-4			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	6	0
		earlyRec	3	10	0
		lateRec	0	1	44

Precision: 0.863 **Accuracy: 0.855**
Recall: 0.854 **AUC ROC: 0.946**
F1-Score: 0.855

MC-4			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	3	8	0
		earlyRec	3	9	1
		lateRec	2	1	42

Precision: 0.791 Accuracy: 0.783
Recall: 0.782 AUC ROC: 0.904
F1-Score: 0.783

MC-4			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	6	4	1
		earlyRec	4	9	0
		lateRec	1	0	44

Precision: 0.855 Accuracy: 0.855
Recall: 0.855 AUC ROC: 0.919
F1-Score: 0.855

MC-4			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	5	5	1
		earlyRec	3	9	1
		lateRec	5	9	31

Precision: 0.748 Accuracy: 0.652
Recall: 0.679 AUC ROC: 0.728
F1-Score: 0.652

Table 5. Performances Metrics Dataset 1 – Two Classes – Tuned

Resultados MC-1 Dataset 1 – Duas Classes.

MC-1			Predicted	
			nonRec	Rec
LR	Actual	nonRec	23	1
		Rec	0	45

Precision: 0.986 **Accuracy: 0.986**
Recall: 0.985 **AUC ROC: 0.979**
F1-Score: 0.986

MC-1			Predicted	
			nonRec	Rec
NB	Actual	nonRec	23	1
		Rec	1	44

Precision: 0.971 Accuracy: 0.971
Recall: 0.971 AUC ROC: 0.968
F1-Score: 0.971

MC-1			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	21	3
		Rec	3	42

Precision: 0.913 Accuracy: 0.913
Recall: 0.913 AUC ROC: 0.904
F1-Score: 0.913

MC-1			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	6	18
		Rec	7	38

Precision: 0.603 Accuracy: 0.638
Recall: 0.604 AUC ROC: 0.547
F1-Score: 0.638

Resultados MC-2 Dataset 1 – Duas Classes.

MC-2			Predicted	
			nonRec	Rec
LR	Actual	NonRecur	24	0
		Recur	2	43

Precision: 0.973 **Accuracy: 0.971**
Recall: 0.971 **AUC ROC: 0.978**
F1-Score: 0.971

MC-2			Predicted	
			nonRec	Rec
NB	Actual	nonRec	23	1
		Rec	2	43

Precision: 0.957 Accuracy: 0.957
Recall: 0.957 AUC ROC: 0.957
F1-Score: 0.957

MC-2			Predicted	
			nonRec	Rec
SVM	Actual	NonRecur	24	0
		Recur	3	42

Precision: 0.961 Accuracy: 0.957
Recall: 0.957 AUC ROC: 0.967
F1-Score: 0.957

MC-2			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	16	8
		Rec	7	38

Precision: 0.781 Accuracy: 0.783
Recall: 0.781 AUC ROC: 0.756
F1-Score: 0.783

Resultados MC-3 Dataset 1 – Duas Classes.

MC-3			Predicted	
			nonRec	Rec
LR	Actual	nonRec	24	0
		Rec	1	44

Precision: 0.986 **Accuracy: 0.986**
Recall: 0.986 **AUC ROC: 0.989**
F1-Score: 0.986

MC-3			Predicted	
			nonRec	Rec
NB	Actual	nonRec	17	7
		Rec	2	43

Precision: 0.872 Accuracy: 0.870
Recall: 0.865 AUC ROC: 0.832
F1-Score: 0.870

MC-3			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	22	2
		Rec	0	45

Precision: 0.972 Accuracy: 0.971
Recall: 0.971 AUC ROC: 0.958
F1-Score: 0.971

MC-3			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	18	6
		Rec	6	39

Precision: 0.826 Accuracy: 0.826
Recall: 0.826 AUC ROC: 0.808
F1-Score: 0.826

Resultados MC-4 Dataset 1 – Duas Classes.

MC-4			Predicted	
			nonRec	Rec
LR	Actual	nonRec	24	0
		Rec	3	42

Precision: 0.961 **Accuracy: 0.957**
Recall: 0.957 **AUC ROC: 0.967**
F1-Score: 0.957

MC-4			Predicted	
			nonRec	Rec
NB	Actual	nonRec	20	4
		Rec	1	44

Precision: 0.929 Accuracy: 0.928
Recall: 0.926 AUC ROC: 0.906
F1-Score: 0.928

MC-4			Predicted	
			nonRec	Rec
SVM	Actual	nonRec	23	1
		Rec	3	42

Precision: 0.945 Accuracy: 0.942
Recall: 0.943 AUC ROC: 0.946
F1-Score: 0.826

MC-4			Predicted	
			nonRec	Rec
KNN	Actual	nonRec	16	8
		Rec	4	41

Precision: 0.824 Accuracy: 0.826
Recall: 0.822 AUC ROC: 0.789
F1-Score: 0.826

Table 6. Performances Metrics Dataset 2 – Three Classes – Tuned

Resultados MC-1 Dataset 2 – Tres Classes.

MC-1			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	5	1
		earlyRec	1	11	1
		lateRec	0	0	45

Precision: 0.923 **Accuracy: 0.899**
Recall: 0.891 **AUC ROC: 0.935**
F1-Score: 0.899

MC-1			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	3	8	0
		earlyRec	0	12	1
		lateRec	1	0	44

Precision: 0.870 Accuracy: 0.855
Recall: 0.838 AUC ROC: 0.938
F1-Score: 0.855

MC-1			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	6	3	2
		earlyRec	3	9	1
		lateRec	2	0	43

Precision: 0.838 Accuracy: 0.841
Recall: 0.839 AUC ROC: 0.884
F1-Score: 0.841

MC-1			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	1	1	9
		earlyRec	1	0	12
		lateRec	0	1	44

Precision: 0.521 Accuracy: 0.652
Recall: 0.546 AUC ROC: 0.627
F1-Score: 0.652

Resultados MC-2 Dataset 2 – Tres Classes.

MC-2			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	6	0
		earlyRec	0	12	1
		lateRec	0	0	45

Precision: 0.923 **Accuracy: 0.899**
Recall: 0.891 **AUC ROC: 0.959**
F1-Score: 0.899

MC-2			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	5	5	1
		earlyRec	1	11	1
		lateRec	1	1	43

Precision: 0.859 Accuracy: 0.855
Recall: 0.850 AUC ROC: 0.920
F1-Score: 0.855

MC-2			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	4	6	1
		earlyRec	1	11	1
		lateRec	0	0	45

Precision: 0.874 Accuracy: 0.870
Recall: 0.856 AUC ROC: 0.951
F1-Score: 0.870

MC-2			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	0	6	5
		earlyRec	1	4	8
		lateRec	0	3	42

Precision: 0.556 Accuracy: 0.667
Recall: 0.606 AUC ROC: 0.791
F1-Score: 0.667

Resultados MC-3 Dataset 2 – Tres Classes.

MC-3			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	3	8	0
		earlyRec	1	10	2
		lateRec	0	1	44

Precision: 0.843 Accuracy: 0.826
Recall: 0.812 AUC ROC: 0.944
F1-Score: 0.826

MC-3			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	1	7	3
		earlyRec	0	9	4
		lateRec	1	0	44

Precision: 0.748 Accuracy: 0.783
Recall: 0.739 AUC ROC: 0.925
F1-Score: 0.783

MC-3			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	6	5	0
		earlyRec	3	9	1
		lateRec	0	0	45

Precision: 0.865 **Accuracy: 0.870**
Recall: 0.866 **AUC ROC: 0.959**
F1-Score: 0.870

MC-3			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	4	2	5
		earlyRec	2	8	3
		lateRec	3	2	40

Precision: 0.740 Accuracy: 0.754
Recall: 0.745 AUC ROC: 0.802
F1-Score: 0.754

Resultados MC-4 Dataset 2 – Tres Classes.

MC-4			Predicted		
			nonRec	earlyRec	lateRec
LR	Actual	nonRec	5	6	0
		earlyRec	3	10	0
		lateRec	0	1	44

Precision: 0.863 **Accuracy: 0.855**
Recall: 0.854 **AUC ROC: 0.931**
F1-Score: 0.855

MC-4			Predicted		
			nonRec	earlyRec	lateRec
NB	Actual	nonRec	3	7	1
		earlyRec	2	10	1
		lateRec	2	1	42

Precision: 0.796 Accuracy: 0.797
Recall: 0.790 AUC ROC: 0.909
F1-Score: 0.797

MC-4			Predicted		
			nonRec	earlyRec	lateRec
SVM	Actual	nonRec	3	7	1
		earlyRec	3	9	1
		lateRec	1	2	42

Precision: 0.785 Accuracy: 0.783
Recall: 0.778 AUC ROC: 0.864
F1-Score: 0.783

MC-4			Predicted		
			nonRec	earlyRec	lateRec
KNN	Actual	nonRec	5	3	3
		earlyRec	4	5	4
		lateRec	4	7	35

Precision: 0.673 Accuracy: 0.652
Recall: 0.661 AUC ROC: 0.671
F1-Score: 0.652

Table 7 shows the performance of the models during training; in green color, the models that presented a better result are highlighted.

Table 7. Models Accuracy Cross-Validation

2 Classes				
Models	MC-1 CV	MC-2 CV	MC-3 CV	MC-4 CV
LR	**0,936**	**0,945**	**0,924**	**0,924**
NB	0,913	0,898	0,833	0,833
SVM	0,578	0,933	0,917	0,917
KNN	0,052	0,753	0,771	0,771

2 Classes Otimizado				
Models	MC-1 CV	MC-2 CV	MC-3 CV	MC-4 CV
LR	**0,942**	0,936	**0,924**	**0,928**
NB	0,924	0,930	0,844	0,084
SVM	0,881	**0,942**	0,913	0,917
KNN	0,619	0,832	0,826	0,778

3 Classes				
Models	MC-1 CV	MC-2 CV	MC-3 CV	MC-4 CV
LR	**0,808**	**0,822**	**0,811**	**0,811**
NB	0,788	0,621	0,749	0,749
SVM	0,578	0,808	0,775	0,775
KNN	0,467	0,680	0,670	0,670

3 Classes Otimizado				
Models	MC-1 CV	MC-2 CV	MC-3 CV	MC-4 CV
LR	**0,837**	**0,825**	**0,811**	**0,815**
NB	0,793	0,808	0,767	0,764
SVM	0,753	0,811	0,785	0,727
KNN	0,575	0,703	0,691	0,654

4. Conclusion

From the analysis of the results, it is possible to conclude that logistic regression presented the best number of results across all the experiments performed, given that on 13 out of 16 occasions, the performance metrics showed the best performance.

When comparing the results of dataset 1 vs. dataset 2, it is possible to observe that the models present a more incredible difficulty in performing the classification in dataset 2 (3 classes) since, in general, the best results of Precision and Recall vary between 0.854 and 0.923. Compared with dataset 1, Precision and Recall vary between 0.957 and 1.

No significant differences were observed in the performance metrics of dataset 1 as a function of the preprocessing techniques, i.e., the MC-1 configuration that used the raw data presents similar results to MC-2 to MC-4 in the LR.

The optimization applied to the parameters of the ML models in the second evaluation phase showed a significant improvement in the performance metrics of the SVM and KNN models. As a point to analyze, the best result obtained in dataset 1 was achieved with the SVM model of the MC-4 configuration. However, at the time of applying the parameter optimization, this result was not maintained. The metric used to compare the performance of the models with each other was observed, and no changes were observed in the experiment's performance when using the optimized parameters.

Finally, different methods of data analysis and processing oriented to breast cancer recurrence classification were evaluated in this work. It is considered a promising field that can contribute to the planning of preventive treatments that can improve patients' quality of life.

References

[1] V. P. C. Magboo and M. S. Magboo, "Machine Learning Classifiers on Breast Cancer Recurrences," *Procedia Computer Science*, vol. 192, pp. 2742–2752, Jan. 2021, doi: 10.1016/J.PROCS.2021.09.044.

[2] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, "Prediction of Malignant & Benign Breast Cancer: A Data Mining Approach in Healthcare Applications," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 37, pp. 435–442, Feb. 2019, doi: 10.48550/arxiv.1902.03825.

[3] T. A. Asfaw, "Comparative Analysis of Classification Approaches for Breast Cancer," *International Journal of Computer Engineering and Technology*, vol. 10, no. 4, pp. 10–16, 2019, Accessed: Jul. 04, 2022. [Online]. Available: www.jifactor.com

[4] L. Khairunnahar, M. A. Hasib, R. H. bin Rezanur, M. R. Islam, and M. K. Hosain, "Classification of malignant and benign tissue with logistic regression," *Informatics in Medicine Unlocked*, vol. 16, p. 100189, Jan. 2019, doi: 10.1016/J.IMU.2019.100189.

[5] SkLearn, "Logistic Regression." https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed Jul. 04, 2022).

[6] SkLearn, "Naive Bayes." https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Jul. 04, 2022).

[7] SkLearn, "Support Vector Machines." https://scikit-learn.org/stable/modules/svm.html (accessed Jul. 04, 2022).

[8] G. Rohith, "Support Vector Machine." https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed Jul. 04, 2022).

[9] SkLearn, "Nearest Neighbors." https://scikit-learn.org/stable/modules/neighbors.html (accessed Jul. 04, 2022).

[10] Z. Kai, "Feature Extraction using Principal Component Analysis." https://towardsdatascience.com/feature-extraction-using-principal-component-analysis-a-simplified-visual-demo-e5592ced100a (accessed Jul. 04, 2022).

[11] J. Korstanje, "SMOTE." https://towardsdatascience.com/smote-fdce2f605729 (accessed Jul. 04, 2022).

[12] J. Brownlee, "SMOTE for Imbalanced Classification with Python." https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (accessed Jul. 04, 2022).

[13] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *Lecture Notes in Computer Science*, vol. 3644, no. PART I, pp. 878–887, 2005, doi: 10.1007/11538059_91.

[14] SkLearn, "Supervised learning." https://scikit-learn.org/stable/supervised_learning.html#supervised-learning (accessed Jul. 04, 2022).

[15] SkLearn, "KMeans." https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (accessed Jul. 04, 2022).

[16] M. Garbade, "Understanding K-means Clustering in Machine Learning | by Dr. Michael J. Garbade | Towards Data Science." https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1 (accessed Jul. 04, 2022).

[17] SkLearn, "One Hot Encoder." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html (accessed Jul. 04, 2022).

[18] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation." https://machinelearningmastery.com/k-fold-cross-validation/ (accessed Jul. 04, 2022).

[19] SkLearn, "GridSearchCV." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed Jul. 04, 2022).

[20] N. Beheshti, "Cross-Validation and Grid Search." https://towardsdatascience.com/cross-validation-and-grid-search-efa64b127c1b (accessed Jul. 04, 2022).