Fuzzy c-Means Clusterization and ANN- MLP Prediction of Malign Breast Cancer in a Cohort of Patients

Alessandro Massaro
LUM University Giuseppe Degennaro https://orcid.org/0000-0003-1744-783X
Alberto Costantiello
LUM University Giuseppe Degennaro https://orcid.org/0000-0002-0004-9635
Nicola Magaletti
LUM University Giuseppe Degennaro https://orcid.org/0000-0002-4001-5747
Gabriele Cosoli
LUM University Giuseppe Degennaro
Vito Giardinelli
LUM University Giuseppe Degennaro
Angelo Leogrande (leogrande.cultore@lum.it)
LUM University Giuseppe Degennaro https://orcid.org/0000-0003-1381-4006

Research Article

Keywords: Breast Cancer Risk, Machine Learning, Algorithms, Prediction, Clusterization

Posted Date: August 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1953135/v1

Alessandro Massaro°, Alberto Costantiello*, Nicola Magaletti°, Gabriele Cosoli°, Vito O.M. Giardinelli°, Angelo Leogrande*°*

Fuzzy c-Means Clusterization and ANN- MLP Prediction of Malign Breast Cancer in a Cohort of Patients

** LUM - Libera Università Mediterranea "Giuseppe Degennaro", S.S. 100 - Km.18, Parco il Baricentro, 70010, Bari, Italy, °LUM Enterprise S.r.l., S.S. 100 - Km.18, Parco il Baricentro, 70010, Bari, Italy. Corresponding author: Angelo Leogrande, email: leogrande.cultore@lum.it.*

Abstract

A model based on to cluster and predict the presence of malign breast cancer. The algorithms are applied to an open dataset of 569 patients that have either benign or malign breast cancer. First the article presents a detailed data description that is followed by a correlation matrix analysis and a regression analysis to verify the select the best variables for the clustering analysis and the predictive model. The clusterization is realized with the fuzzy c-Means algorithms. The prediction is performed by the Artificial Neural Network (ANN) Multilayer Perceptron (MLP) algorithm showing the best performance in comparison with the results of other machine learning algorithms. The use of all the tools in a definite sequence is the model proposed in this work providing an auto consistent approach to automatize the risk calculus with a good efficiency. The proposed approach can be applied for other typologies of cancers.

Keywords: Breast Cancer Risk, Machine Learning, Algorithms, Prediction, Clusterization.

1. Introduction-Research Question

The research question is to find an auto consistent approach based on the simultaneous adoption of correlation matrix, regression, and fuzzy c-Means and ANN-MLP algorithms to optimize an approach clustering and predicting patients with breast cancer. The innovative research methodology is suitable to optimize prediction with a good accuracy and can be applied for other typology of cancers. The proposed method is also a solution for a pre-screening of patient's cohorts having a high probability to have a cancer. Breast cancer is one of the leading causes of death in woman in the world. The diagnosis of breast cancer generates consequences in terms of psychosocial morbidity [1].

2. Research Methodology

The research methodology is oriented to investigate a set of tools to finds clusters and predict the presence of malign breast cancer among the analysed cohorts of patients. The first phase of the analysis consists in the description of the dataset with the distinction between benign and malign breast cancer. The description of the dataset shows the inner characteristics of the variables with the suggestion of the possible relationships among variables. After having described the dataset a regression analysis is proposed to find the best variables to use to predict the value of malign breast cancer. Specifically, the choice of the number of variables is realized with the application of the p—value. With the built model it is realized a clusterization with the algorithm fuzzy c-means. The choice of the algorithm fuzzy c-Means algorithm is a necessity because the investigated variable is dichotomous i.e. it is 0 for benign breast cancer and 1 for malign breast cancer. Data obtained as an

output of the clusterization with fuzzy c-Means are used to compute a percentage ratio between the number of patients that have malign breast cancer and the total number of patients for each cluster. This ratio is relevant not only per sé but also as a quantitative tool to compare the pre- machine learning clusterization with the post machine learning clusterization. The following phase consists in the realization of a prediction based on the application of ANN-MLP machine learning algorithm. The results are divided in three categories based on the probability to develop breast cancer i.e.: high probability, medium probability, and low probability. The data that are in the context of high probability are used for the second clusterization with the fuzzy c-Means algorithm optimized with the Silhouette coefficient. Furthermore, the obtained data are analysed with the computation of the ratio between the number of patients with malign breast cancer and the total number of patients for each cluster. At this stage of analysis there is a confrontation between the malign breast cancer ratio of the pre-machine learning clusterization and the malign breast cancer of the post machine learning clusterization. The confrontation is useful to verify if the information of the clusterization with the pre-machine learning algorithm are under, equal, or over-estimated in respect with the post-machine learning probability to develop malign breast cancer. Specifically in the evaluation of the efficacy of the clusterization, it is used the following ratio:

$$\boldsymbol{MalignBreastCancerRatio_i = \frac{NumberOfPatientsWithMalignBreastCancer_i}{NumberOfPatients_i}}$$

Where i= number of the clusters. This ratio is used either to order clusters in the same clusterization, either to confront the performance of clusters among different clusterization i.e. the pre-machine learning clusterization vs. the post-machine learning clusterization.
The application of this kind of methodology is relevant for a series of motivations:

- it consents to describe correctly the dataset distinguishing between patients with malign and benign breast cancer,
- it is useful to find the best variables to create a model to investigate the level of malign breast cancer,
- the choice of the best variables creates the conditions to improve the efficiency of either the cluster analysis and the machine learning prediction;
- using the malign breast cancer ratio, it is possible to create a comparison between pre-machine learning and post machine learning clusterization.

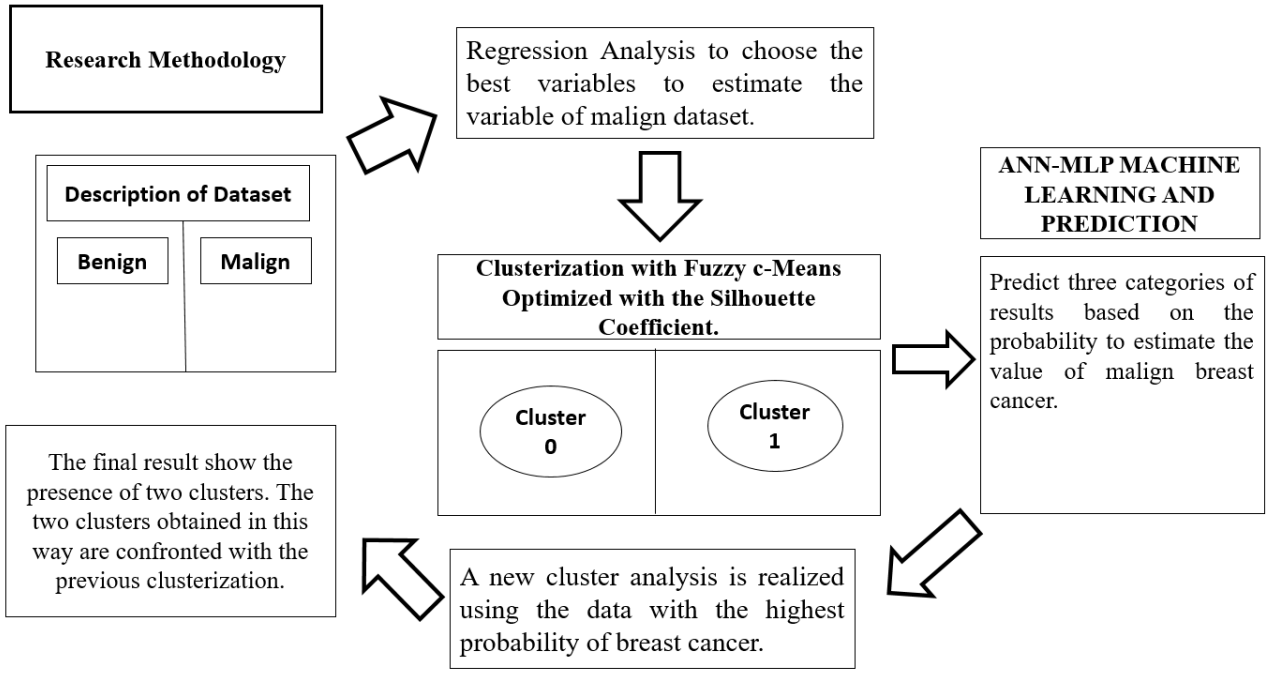


Figure 1. Research Methodology. The first phase of the analysis is based on the description of dataset, followed by a regression analysis used to choose the optimal set of variables to perform clusterization and prediction. The choice of the best variables is realized by the optimization of the p-value. After having found the best model through the usage of the regression, analysis it is realized a clusterization with fuzzy c –Means optimized with the Silhouette Coefficient. Furthermore, with the same variable a prediction is realized with the ANN-MLP algorithm. The results of the prediction that show the highest probability of developing malign breast cancer are used to make a new clusterization with fuzzy c-Means.

As we can see the applied research methodology is based on four metric tools that are: regression analysis, clusterization with fuzzy c-Means, machine learning and predictions. The use of this set of analytical tools is necessary to verify the presence of relationships among variables that can be used to predict the level of malign breast cancer among the analysed dataset and to verify the efficiency of the pre-machine learning clusterization with the post-machine learning clusterization.

3. Literature Review

Breast Cancer and Machine Learning. [2] use functional Magnetic Resonance Imaging-fMRI to predict the cognitive decline associated to breast cancer. [3] apply data mining and machine learning technique to predict breast cancer using a set of tools i.e.: Decision Tree, Naïve Bayes, and Artificial Neural Network. [4] use convolutional neural network-CNN to classify breast ultrasound images in four categories i.e.: mass, fatty tissue, fibro glandular tissue, skin. Results show that Accuracy, Precision, Recall and F1 Measure reached over 80%. [5] propose a method based on machine learning technique to predict malign breast cancer based on a set of algorithms i.e.: k-Nearest Neighbourhood, Logistic regression, Decision Tree, Random Forest, Support Vector Machine and deep learning using Adam Gradient Descent Learning. The authors found that the best predictive accuracy is achieved with Adam gradient Descent Learning with a value of 98.24%. [6] use deep learning convolutional neural network to predict breast cancer based on Mammograph MIAS database. [7] consider a set of machine learning algorithms to predict breast cancer i.e.: Decision Tree, Random Forest, K-nearest Neighbours, Support Vector Machine, Logistic Regression, and Artificial Neural Network. Results

show that KNN has the higher accuracy. [8] use Artificial Intelligence to predict breast cancer through the application of Convolutional Neural Network-CNN, Logic Based, as Random Forest-RF, Support Vector Machines-SVM and Bayesian methods. [9] apply a set of machine learning algorithm to image processing to detect breast cancer. Specifically, the authors apply Fuzzy SVM, Bayesian Classifier and Random Forest. [10] perform Bayesian optimization to distinguish between malign breast cancer and benign breast cancer. [11] consider a set of algorithms to predict the level of malign breast cancer using Machine Learning-ML, Support Vector Machine-SVM, Artificial Neural Network-ANN, K-Nearest Neighbour-KNN, Decision Tree-DT. [12] use a set of machine learning algorithms to predict the level of breast cancer. [13] implement a series of machine learning algorithms to predict the level of malign breast cancer using kNN, decision tree, Binary SVM and AdaBoost. The model shows the presence of a predictive accuracy equal to 99.12% in the case of kNN, equal to 98.86% for Binary SVM model. [14] use machine learning technique to predict the survival 5-year rate of patients with breast cancer. Analyse the performance of a series of algorithms to predict the breast cancer, the result shows that Support Vector Machine is the best predictor with an accuracy of 97.07%. [15] critically consider a series of 1,879 articles and find that the most relevant algorithms are: Support Vector machine-SVM, Artificial Neural Network-ANN, Decision Tree-DT, Naïve Bayes-NB, and k-Nearest Neighbour KNN. Use Support Vector Machine SVM and Random Forest-RF for the detection of breast cancer through the application of miRNA biomarkers [16]. Use machine-learning technique to estimate the financial distress of patients with breast cancer [17]. Apply a set of machine learning technique to predict malign breast cancer using Wisconsin Breast Cancer Database-WBCD through the application of Artificial Neural Network-ANNs, Support Vector Machine-SVMs, Decision Tree-DTs, and k-Nearest Neighbours k-NNS [18]. [19] propose a confrontation between Deep Learning-DL and Conventional Machine Learning-CML for the prediction of breast cancer. The authors show that DL tends to perform better than CML. [20] apply a set of machine learning algorithms to analyze ultrasound images to correctly identify the presence of breast cancer. The authors specifically uses k-Nearest Neighbour, Support Vector Machine, Decision Tree, Naïve Bayes to classify data and CNN to classify breast cancer. [21] estimate the probability of developing breast cancer through the application of different machine learning algorithms based on blood pressure. The authors employ Artificial Neural Network-ANN, standard extreme Learning Machine-ELM, Support Vector Machine-SVM, and k-Nearest Neighbour-kNN. [22] use Principal Component Analysis-PCA and Artificial Neural Network-ANN to predict the value of breast cancer. [23] use a set of algorithms to predict the level of ML algorithms namely support vector machines-SVM, Logistic Regression-LR, and k-Nearest Neighbours-KNN. [24] apply Support Vector Machine-SVM, Naïve Bayesian, Linear Discriminant, Quadratic Discriminant, Logistic Regression, k-Nearest Neighbour k-NN, and Random Forest. The authors find that k-NN algorithm has the best classification accuracy with a value of 92,105%. Use Random Forest, Gradient Boosted Decision Trees, and Logistic Regression to precision oncology [25]. [26] use a set of machine learning algorithms to predict breast cancer such as Random Forest, kNN k-Nearest Neighbour and Naïve Bayes using Wisconsin Diagnosis Breast Cancer. [27] apply Employ Decision Tree, Naïve Bayes, and Sequential Minimal Optimization to predict breast cancer. The authors use two different datasets i.e.: Wisconsin Breast Cancer and Breast Cancer dataset. Results show that the Sequential Minimal Optimization has better results in predicting breast cancer using Wisconsin Breast Cancer while Decision Tree has the best performance in predicting breast cancer using Breast Cancer dataset. [28] use a set of machine learning algorithms to predict breast cancer. Specifically, the authors use SVM, kNN, MLP, Decision Tree, Random Forest, Logistic Regression, AdaBoost, Gradient Boosting Machines. The authors show that SVM offers the best results. [29] apply Support Vector Machines-SVMs to predict breast cancer with an accuracy of 0,825. [30] use Naïve Bayes, Random Forest, AdaBoost, Support Vector Machine (SVM), Least

square SVM, AdaBag, Logistic Regression-LR and Linear Discriminant Analysis to predict breast cancer survival. Results show that SVM and LDA have a greater accuracy equal to 93%. Predict breast cancer using machine-learning techniques i.e.: Linea Regression, Random Forest, Multi-layer Perceptron and Decision Tree-DT [31]. Use random forest to predict breast cancer with a high level of accuracy [32]. [33] use a set of algorithms such as random forest, support vector machine, logistic regression and Bayesian classification algorithms to predict breast cancer. Results shows that the best predictive algorithm is Random Forest with an area under receiver operating curve equal to 0.75. [34] use logistic regression, k-nearest neighbours, support vector machine, naïve Bayes, decision tree, random forest, and rotation forest. The author finds that the logistic regression is the best predictor with a level of accuracy of 98.1%. [35] use a set nine different machine learning algorithms to predict breast cancer. Specifically, the authors apply Logistic Regression, Gaussian Naïve Bayes, Linear Support Vector Machine, RBF Support vector Machine, Decision Tree, Random Forest, XgBoost, Gradient Boosting, KNN. The results shows that KNN and logistic regression have the highest level of accuracy equal to 98%. [36] apply Support Vector Regression, Lasso Regression, Kernel Ridge regression, K-Neighbourhood Regression, and Decision Tree regression to predict breast cancer. [37] apply simple logistic regression, support vector machine and multilayer perceptron network in association with a voting scheme to predict Breast Cancer. The authors find that the level of accuracy in the case of majority-based voting is equal to 99.42%. [38] use Support Vector Machine-SVM, Decision Tree, Naïve Bayes, and k-Nearest Neighbours kNN to predict breast cancer. The results show that SVM has the best level accuracy in prediction with a level of 97.13%. [39] apply Decision Tree C4.5, Decision Tree C5.0, Gradient Boosting Model-GBM, Artificial Neural Network-ANN, Support Vector Machine-SVM to predict the probability of lymphedema in breast cancer survivor. The authors find that ANN has an accuracy of 93.75% in predicting lymphedema.

In a broader sense, it is possible to apply new technologies to detect disease. Specifically, it is necessary to apply new technologies to detect tumour masses [40]. Augmented data can be applied to predict main diseases [41]. Machine learning algorithms can also be used to predict hypertension risk [42] and diabetes [43]. Finally, telemedicine is an essential tool in the passage from industry 4.0 to industry 5.0 [44].

4. The Metric Models

We test the following model with an application of set of regression technique. Specifically, to test the hypothesis are applied both dichotomous and non-dichotomous models. Dichotomous models are: Probit, Logit, Tobit. Non-dichotomous models are: OLS, WLS heteroskedasticity. The application of either dichotomous or non-dichotomous is proposed to realize a comparative analysis and to verify if the model is persistent in differentiated models.

$$\begin{aligned} \boldsymbol{BreastCancer_i} &= \boldsymbol{a_1} + \boldsymbol{b_1}(\boldsymbol{PerimeterMean})_i + \boldsymbol{b_2}(\boldsymbol{ConcavitySe})_i \\ &+ \boldsymbol{b_3}(\boldsymbol{ConcavepointsSe})_i + \boldsymbol{b_4}(\boldsymbol{TextureWorst})_t + \boldsymbol{b_5}(\boldsymbol{ConcavityWorst})_i \\ &+ \boldsymbol{b_6}(\boldsymbol{SymmetryWorst})_i \end{aligned}$$

Applying the following model, we find that breast cancer is positively associated to:

- *Perimeter Mean;*
- *Concave Point Standard Error;*
- *Texture Worst;*
- *Concavity Worst;*
- *Symmetry Worst;*

We also find that there is a negative relationship between breast cancer and the following variable:

- *Concavity Standard Error:*

Synthesis of the Regression Analysis						
	Logit	**Probit**	**Tobit**	**OLS**	**WLS Corrected for Heteroskedasticity**	
Variables	**Coefficient and P-Value**	**Coefficient and P-Value**	**Coefficient and P-Value**	**Coefficient and P-Value**	**Coefficient and P-Value**	**Mean**
Const	-35,4349 ***	-18,0211 ***	−3,35122 ***	−1,37832 ***	−1,13263 ***	-26,728
Perimeter Mean	0,197032 ***	0,0993254 ***	0,0153865 ***	0,00880793 ***	0,0080476 ***	0,06572
Concavity Standard Error	-107,4 ***	-63,2087 ***	−14,7680 ***	−4,86494 ***	−3,68499 ***	-85,304
Concave Point Standard Error	368,87 ***	214,896 ***	50,8355 ***	18,71 ***	11,6268 ***	132,988
Texture Worst	0,285998 ***	0,148432 ***	0,0302107 ***	0,0137337 ***	0,009635 ***	0,0976
Concavity Worst	15,9264 ***	8,54891 ***	1,73397 ***	0,798545 ***	0,615872 ***	5,52474
Symmetry Worst	11,0114 *	5,00465 *	1,6312 ***	1,05249 ***	1,05809 ***	3,95157

Table 3. The Metric Results obtained either with binary either with non-binary metric model. The analysis shows that all the variables are positively associated with breast cancer with the sole exception of Concavity Standard Error that has a negative association.

The application of those regression models is useful to verify the cross-model persistency of the relationships among variables. Specifically in the proposed model there are two different types of models i.e.; dichotomous and non-dichotomous. Dichotomous model are those models that are that are able to verify the presence of statistical relationships among bijective variables i.e. variables that can alternatively assume a value of 0 and 1 such as Logit, Probit and Tobit. Non-dichotomous models are those models that can be used for continuous variable. In this case, the non-dichotomous models are OLS and WLS corrected heteroskedasticity. The confrontation among dichotomous and non-dichotomous models shows the confirmation of the relationships among variables. The choice of the variables is realized through the minimization of the p-value.

5. Clusterization with c-Means Algorithm optimized with the Silhouette Coefficient

In the following part it is analysed the methodology that is applied to maximize the number of clusters. The algorithm of clustering is fuzzy c-Means. In this case we apply fuzzy c-Means instead of k-Means since the investigate variable is not continuous but dichotomous. In effect, fuzzy c-Means is more efficient in the case of dichotomous variables. The Silhouette Coefficient is used to optimize the number of clusters. The following results are obtained after an empirical investigation. The result

shows the presence of an inverse relationship between the number of clusters and the value of the overall Silhouette Coefficient computed using KNIME. Based on this empirical analysis we choose a number of clusters equal to 2 to optimize the level of the Silhouette Coefficient that is equal to 0,25686.

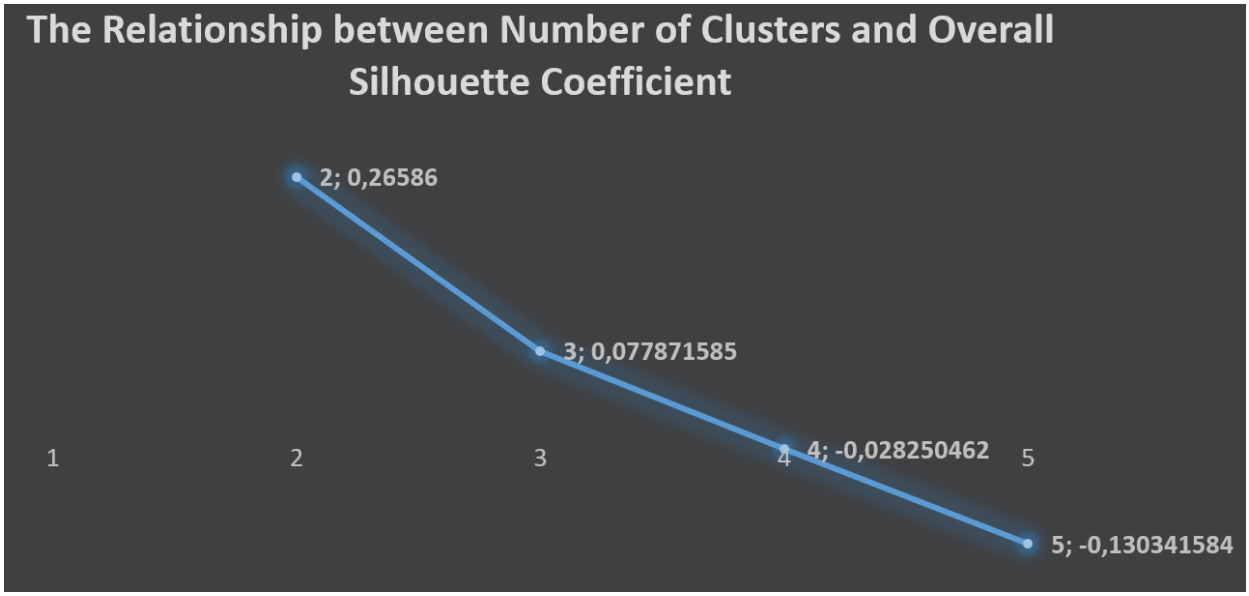


Figure 2. The relationship between the number of clusters and the overall Silhouette Coefficient. The relationship show the presence of a negative relationship between the number of clusters and the level of the Silhouette Coefficient. To optimize the level of clusters we choose a number of clusters equal to 2.

Clusterization with the Optimization of the Silhouette Coefficient		
Variables	Values	
Clusters	0	1
Number of Observations	96	40
Diagnosis -Number of Patients with Malign Breast Cancer	66	5
Perimeter Mean Median Value	101,4	80,24
Concave Points Mean Median Value	0,07	0,03
Concave Points Standard Error Median Value	0,01	0,01
Texture Worst Median Value	29,09	20,71
Concavity Worst Median Value	0,35	0,14
Symmetry Worst Median Value	0,3	0,27
*Diagnosis/Number of Observations*100=MalignBreastCancer*	68,75	12,5

Table 4. Results of clusterization with fuzzy c-Means optimized with the Silhouette Coefficient.

The two clusters have the following characteristics:

- Cluster 0: the number of observations is equal to 96, the median value of the diagnosis is equal to 66. This means that there are 96 observations in the cluster 0 and the number of patients with malign breast cancer is equal to 66. It derives that the percentage of patients with breast cancer is equal to 68,75%. The Median Value of Perimeter Mean is equal to 101,4, the Concave Points Mean Median value is equal to 0,01, the Texture Worst Median Value is equal

to 29,09, the Concavity Worst Median Value is equal to 0,35, the Symmetry Worst Median value is equal to 0,3.

- Cluster 1: with a number of observations, equal to 40, with a number of patients with Malign Breast Cancer equals to 5. The percentage of patients with malign breast cancer computed in respect to the number of observations that is equal to 12,5%. The median value of the Perimeter Mean is equal to 80,4, the median value of Concave Points is equal to 0,03, the median value of Concave Points Standard Error is equal to 0,01, the Median value of Texture Worst is equal to 20,71, the median value of the Symmetry Worst is equal to 0,27.

As we can observe from the analysis of the clusterization with the fuzzy c-Means algorithm optimized with the Silhouette coefficient the cluster 0 has a presence of patients with malign breast cancer equal to 67% while the cluster 1 has a correspondent value of 12,5%. The clustering algorithm has divided the data in two clusters one of which has a greater number of patients with breast cancer while the other has a lower level of the observed variable. To verify the efficacy of the proposed model we can analyse the data of the cluster 0 to verify if in this case the model preserves its statistically significance. In this sense we run a series of regression based on the data of cluster 0 to confirm the validity of the model. We found no statistical significance of the tested model for the observations of the cluster 0. This means that the model has a general validity to find relationship and predict the value of the presence of malign breast cancer in the observed population. This means that to develop a model that can predict malign breast cancer with greater certainty it is necessary to consider a dataset that is constituted only of data on malign breast cancer.

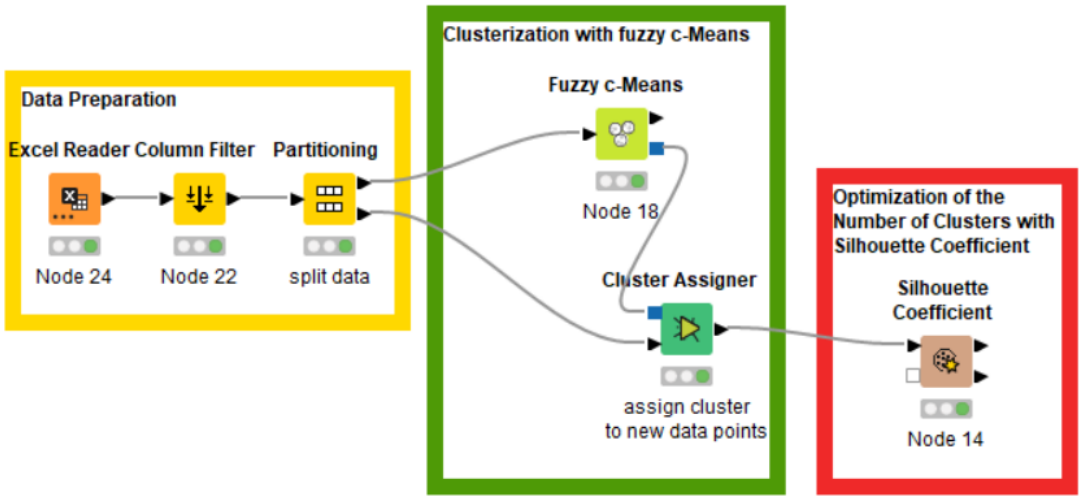


Figure 3. The workflow of the KNIME algorithm is based on three different phases that are Data Preparation, Clusterization with fuzzy c-Means and optimization of the number of clusters with Silhouette Coefficient. The workflow synthesizes the complex activity of choosing variables, computing the clusterization with fuzzy c-Means and the valuation of the Silhouette Coefficient.

6. Machine Learning and Prediction with Original Data

In this section, we show the predictive performance of a series of machine learning algorithms that have been confronted. Each algorithm has been trained with the 80% of the feasible data while the remaining 20% has been used for the prediction. Specifically, each algorithm has been evaluated for its ability to maximize the level of R-squared and minimize the level of statistical errors that are Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. A ranking is built for each of

this statistical measure and for each algorithm, it is assigned a number in the ranking. By this means the algorithms that have totalized the lower levels of ranking can also be considered has the level.

Ranking of Machine Learning Algorithms based on Predictive Performance						
Rank	**Algorithm**	***R^2***	***mean absolute error***	***mean squared error***	***root mean squared error***	***Sum***
1	*Random Forest Regression*	1	5	1	1	8
1	*ANN*	2	2	2	2	8
2	*Tree Ensemble Regression*	3	6	3	3	15
2	*PNN*	4	3	4	4	15
3	*Gradient Boosted Trees Regression*	5	4	5	5	19
3	*Simple Regression Tree*	6	1	6	6	19
4	*Polynomial Regression*	7	7	7	7	28
5	*Linear Regression*	8	8	8	8	32

Figure 4. Ranking of machine learning algorithms based on predictive performance. Random Forest Regression and ANN-Artificial Neural Network are both at the same rank with a payoff equal to 8. But, since the value of ANN-Artificial Neural Network in each is rank is better than the correspondent value of Random Forest Regression for the fact that the first has a position of 5 in Mean Absolute Error then we choose to prefer ANN-Artificial Neural Network to Random Forest Regression as the best performing algorithm to predict the level of breast cancer.

Using the best predictive algorithm, i.e. ANN-Artificial Neural Network, it is possible to predict the probability to develop breast cancer. The main results of the prediction of breast cancer has synthetized below:

Results of the Prediction of Breast Cancer Based on the Best Predictive Algorithm i.e. ANN-Artificial Neural Network with MLP-Multilayer Perceptron			
Range	*Number of patients*	*Level of Alert*	*Description*
>0,70	28	High	28 persons have a high probability to develop breast cancer. Among them there are 25 persons for which the probability to develop breast cancer is very high with a predicted value between [0,90;1] in a scale between 0 and 1.
0,50<x<0,70	5	Medium	5 persons have a medium probability to develop breast cancer based on the prediction algorithm.
0,0<x<0,50	81	Low	81 persons have a low probability to develop breast cancer.

Table 1. Results of the Prediction of Breast Cancer Based on the Best Predictive Algorithm i.e. ANN-Artificial Neural Network with MLP-Multilayer Perceptron.

As we can see with the analysis of the usage of ANN-MLP to predict the probability of development of malign breast cancer there are 28 patients that have a high probability to develop malign breast cancer, 5 patients with a medium probability to develop malign breast cancer and 81 patients that have a low probability to develop breast cancer. It is necessary to underline that the prediction is realized on a population that is already affected by breast cancer and that are divided in two parts i.e.: malign breast cancer and benign breast cancer. This means that the healthy population is excluded from the analysis. This means that the proposed estimation cannot applied to the whole population, but that probability should explicitly be referred to patients that already have been diagnosed with the

presence of breast cancer both malign and benign. After having realized the prediction with the ANN-MLP machine learning algorithm a new clusterization is realized with the fuzzy c-Means algorithm using as output the set of data that show the highest probability to be affected by malign breast cancer i.e. that observations that have a value comprehended between 0,70 and 1.00. Specifically, the post-machine learning clusterization show the presence of two clusters in which the malign breast cancer ratio is lower than the pre-machine learning clusterization. The difference among the pre-machine learning clusterization and the post-machine learning clusterization is a signal of the difference in terms of the accuracy of the analysis. The post-machine learning clusterization show a level of accuracy elevated in respect to the pre-machine learning clusterization and this difference let us infer that the probability of patients to develop malign breast cancer in the post-machine learning prediction is lower that the correspond value of the pre-machine learning prediction.

7. Discussion of the Results

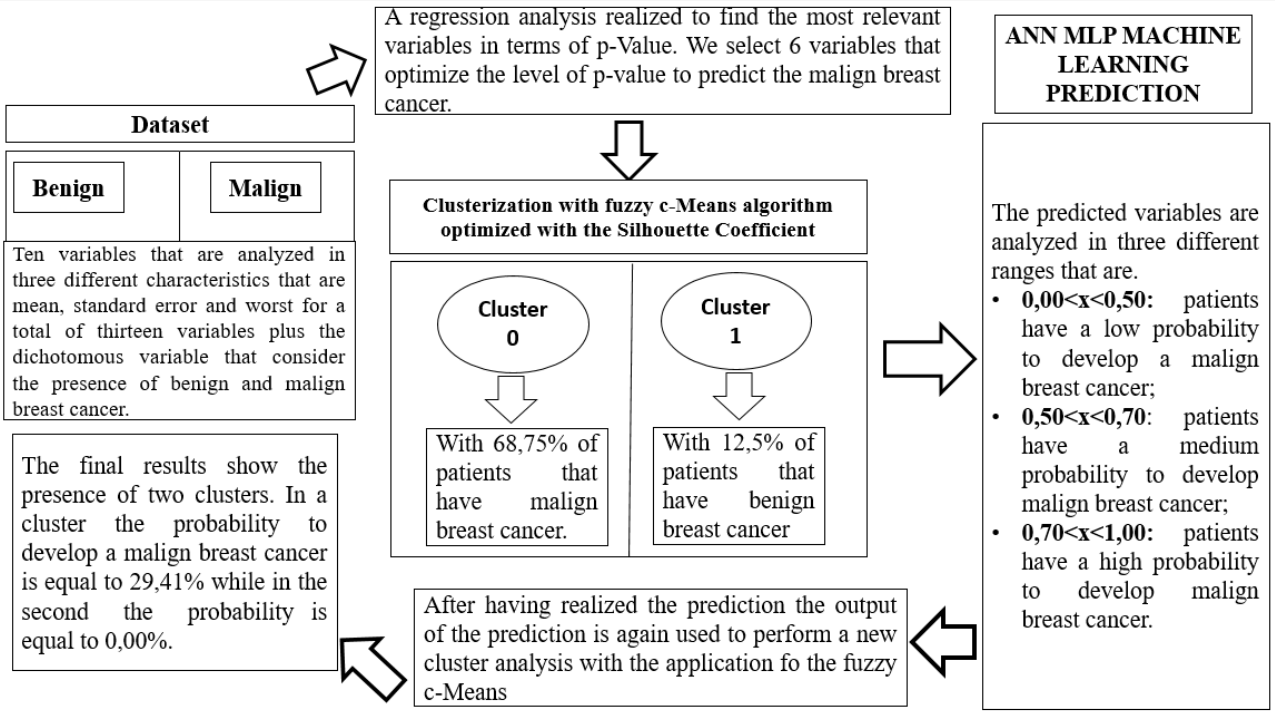


Figure 5. The methodology. The applied methodology starts with a description of data. In the second part, a series of regression analysis is realized to choose the best variable for the econometric model to predict the level of malign breast cancer. The choice of the variable is realized with the maximization of the p-values. A cluster analysis with fuzzy c-Means algorithm optimized with the Silhouette Coefficient is realized. We found two different clusters: one in which the percentage of malign breast cancer is equal to 68,75% and the other in which the percentage of malign breast cancer is equal to 12,5%. Furthermore, with the six variables that we have found we run a prediction with the best predictive algorithm i.e. ANN-MLP. With the predictive data we run a second cluster analysis with the fuzzy c-Means algorithm, and we fund two clusters in which with a significant lower percentage of malign breast cancer.

8. Test Scores, Accuracy and ROC Curve

Furthermore, we promote a comparison among ten different machine-learning algorithms to verify the level of accuracy, AUC, F1, Precision and Recall. We found the following order in terms of accuracy:

- Neural Network with a level of accuracy equal to 0,961;

- Support Vector Machine with a level of accuracy equal to 0,96;
- Stochastic Gradient Descent with a level of accuracy equal to 0,956;
- Gradient Boosting with a level of accuracy equal to 0,947;
- Random Forest with a level of accuracy equal to 0,946;
- Naïve Bayes with a level of accuracy equal to 0,926;
- Tree With a level of accuracy equal to 0,921;
- Logistic Regression with a level of accuracy equal to 0,919.

In the following table there is the numerical representation of the value of the performance of algorithms in terms of different statistical measures that are:

- Area Under ROC-AUC: that is the area under the receiver operating curve;
- Classification Accuracy-AC: that is defined as the proportion of correctly classified examples;
- F1: a weighted harmonic mean of precision and recall;
- Precision is defined as the proportion of true positives among instances classified as positive;
- Recall: the proportion of true positives among all positive instances in the data.

Ranking of Algorithms Based on Accuracy

Rank	Model	AUC	F1	Precision	Recall	CA
1	*Neural Network*	0,99	0,961	0,961	0,961	0,961
2	*SVM*	0,99	0,959	0,96	0,96	0,96
3	*SGD*	0,95	0,956	0,956	0,956	0,956
4	*Gradient Boosting*	0,98	0,947	0,947	0,947	0,947
5	*Random Forest*	0,98	0,945	0,945	0,946	0,946
6	*Naive Bayes*	0,98	0,926	0,927	0,926	0,926
7	*Tree*	0,92	0,921	0,921	0,921	0,921
8	*Logistic Regression*	0,97	0,919	0,919	0,919	0,919
9	*Adaboost*	0,9	0,912	0,912	0,912	0,912
10	*CN2 rule inducer*	0,87	0,847	0,888	0,879	0,879

Figure 6. Ranking of Algorithm Based on Accuracy

The analysis shows that the ANN is not only the best predictor in terms of minimization of statistical errors, but it is also the best algorithm in terms of accuracy. Furthermore, the ANN has also good performances in other statistical measures such as: AUC with a value equal to 0,987, F1 with a value equal to 0,961, Precision with a value equal to 0,961, Recall with a value equal to 0,961. Those statistical measures

9. Conclusions

The paper proposed an innovative model based on a pre-analysis of the variables to be considered for the risk evaluation of breast cancers. The goal of the paper is to explain the adopted auto consistent model, and the procedure useful for a pre-screening the variable having a major weight for the

prediction estimation. Furthermore, the model provides a method to analyse the variables of a dataset which will be different in cases of different breast cancers.

10. References

[1] H. Matthews, E. A. Grunfeld and A. Turner, "The efficacy of interventions to improve psychosocial outcomes following surgical treatment for breast cancer: a systematic review and meta-analysis," *Psycho-oncology,* vol. 26, no. 5, pp. 593-607, 2017.

[2] S. R. Kesler, A. Rao, D. W. Blayney, I. A. Oakley-Girvan, M. Karuturi and O. Palesh, "Predicting long-term cognitive outcome following breast cancer with pre-treatment resting state fMRI and random forest machine learning," *Frontiers in human neuroscience,* vol. 11, no. 555, 2017.

[3] S. Eltalhi and H. Kutrani, "Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review," *IOSR Journal of Dental and Medical Sciences,* vol. 18, no. 4, pp. 85-94, 2019.

[4] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics,* vol. 91, pp. 1-9, 2019.

[5] P. Gupta and S. Garg, "Breast cancer prediction using varying parameters of machine learning models," *Procedia Computer Science,* vol. 171, pp. 593-601, 2020.

[6] N. Khuriwal and N. Mishra, "Breast cancer detection from histopathological images using deep learning," *2018 3rd international conference and workshops on recent advances and innovations in engineering (ICRAIE) ,* pp. 1-4, 2018.

[7] M. M. Islam and T. N. Poly, "Machine learning models of breast cancer risk prediction," *BioRxiv,* no. 723304, 2019.

[8] A. A. Nahid and Y. Kong, "Involvement of machine learning for breast cancer image classification: a survey," *Computational and mathematical methods in medicine,* 2017.

[9] S. Chaudhury, A. N. Krishna, S. Gupta, K. S. Sankaran, K. S. K. Sau and F. Sammy, "Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer," *Computational and Mathematical Methods in Medicine,* 2022.

[10] E. Michael, H. Ma, H. Li and S. Qi, "An optimized framework for breast cancer classification using machine learning," *BioMed Research International,* 2022.

[11] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath and M. A. Bamiah, "Early detection of breast cancer using machine learning techniques," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* vol. 10, no. 3-2, pp. 21-27, 2018.

[12] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC),* pp. 226-229, 2017.

[13] S. Laghmati, B. Cherradi, A. Tmiri, O. Daanouni and S. Hamida, "Classification of patients with breast cancer using neighbourhood component analysis and supervised machine learning techniques," *2020*

3rd International Conference on Advanced Communication Technologies and Networking (CommNet) , pp. 1-6, 2020.

[14] M. Zhao, Y. Tang, H. Kim and K. Hasegawa, "Machine learning with k-means dimensional reduction for predicting survival outcomes in patients with breast cancer," *Cancer informatics,* vol. 17, no. 1176935118810215, 2018.

[15] R. D. Nindrea, T. Aryandono, L. Lazuardi and I. Dwiprahasto, "Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis," *Asian Pacific journal of cancer prevention: APJCP,* vol. 7, no. 1747, p. 19, 2018.

[16] O. Rehman, H. Zhuang, A. Muhamed Ali, A. Ibrahim and Z. Li, "Validation of miRNAs as breast cancer biomarkers with a machine learning approach," *Cancers,* vol. 3, no. 431, p. 11, 2019.

[17] C. Sidey-Gibbons, A. Pfob, M. Asaad, S. Boukovalas, Y. L. Lin, J. C. Selber and A. C. Offodile, "Development of machine learning algorithms for the prediction of financial toxicity in localized breast cancer following surgical treatment," *JCO clinical cancer informatics,* vol. 5, pp. 338-347, 2021.

[18] W. Yue, Z. Wang, H. Chen, A. Payne and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs,* vol. 2, no. 13, p. 2, 2018.

[19] S. Boumaraf, X. Liu, Y. Wan, Z. Zheng, C. Ferkous and X. B. D. Ma, "Conventional machine learning versus deep learning for magnification dependent histopathological breast cancer image classification: A comparative study with visual explanation," *Diagnostics,* 2021.

[20] Y. Pourasad, E. Zarouri, M. Salemizadeh Parizi and A. Salih Mohammed, "Presentation of novel architecture for diagnosis and identifying breast cancer location based on ultrasound images using machine learning," *Diagnostics,* vol. 11, no. 10, p. 1870, 2021.

[21] M. F. Aslan, Y. Celik, K. Sabancı and A. Durdu, "Breast cancer diagnosis by different machine learning methods using blood analysis data," *International Journal of Intelligent Systems and Applications in Engineering,* 2018.

[22] B. Sahu, S. Mohanty and S. Rout, "A hybrid approach for breast cancer classification and diagnosis," *EAI Endorsed Transactions on Scalable Information Systems,* vol. 6, no. 20, 2019.

[23] G. Battineni, N. Chintalapudi and F. Amenta, "Performance analysis of different machine learning algorithms in breast cancer predictions," *EAI Endorsed Transactions on Pervasive Health and Technology,* vol. 6, no. 23, pp. e4-e4, 2020.

[24] B. K. Singh, "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm.," *Biocybernetics and Biomedical Engineering,* vol. 39, no. 2, pp. 393-409, 2019.

[25] J. Reddy, W. D. Lindsay, C. G. Berlind, C. A. Ahern and B. D. Smith, "Applying a machine learning approach to predict acute toxicities during radiation for breast cancer patients," *International Journal of Radiation Oncology, Biology, Physics,* p. 10, 2018.

[26] S. Sharma, A. Aggarwal and T. Choudhury, "Breast cancer detection using machine learning algorithms," *2018 International conference on computational techniques, electronics and mechanical systems (CTEMS),* pp. 114-118, 2018.

[27] S. A. Mohammed, S. Darrab, S. A. Noaman and G. Saake, “Analysis of breast cancer detection using different machine learning techniques,” in *International Conference on Data Mining and Big Data*, Singapore, Springer, 2020, pp. 108-117.

[28] S. Turgut, M. Dağtekin and T. Ensari, “Microarray breast cancer data classification using machine learning methods,” *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT),* pp. 1-3, 2018.

[29] K. Kourou, G. Manikis, P. Poikonen-Saksela, K. Mazzocco, R. Pat-Horenczyk, B. Sousa and D. I. Fotiadis, “A machine learning-based pipeline for modeling medical, socio-demographic, lifestyle and self-reported psychological traits as predictors of mental health outcomes after breast cancer diagnosis: An initial effort to define resilience effects,” *Computers in Biology and Medicine,* vol. 131, no. 104266, 2021.

[30] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi and J. Poorolajal, “Prediction of survival and metastasis in breast cancer patients using machine learning classifiers,” *Clinical Epidemiology and Global Health,* vol. 7, no. 8, pp. 293-299, 2019.

[31] M. Gupta and B. Gupta, “A comparative study of breast cancer diagnosis using supervised machine learning techniques,” *2018 second international conference on computing methodologies and communication (ICCMC),* pp. 997-1002, 2018.

[32] B. Dai, R. C. Chen, S. Z. Zhu and W. W. Zhang, “Using random forest algorithm for breast cancer diagnosis,” *2018 International Symposium on Computer, Consumer and Control (IS3C),* pp. 449-452, 2018.

[33] Y. J. Tseng, C. E. Huang, C. N. Wen, P. Y. Lai, M. H. Wu, Y. C. Sun and J. J. Lu, “Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies,” *International journal of medi,* vol. 128, pp. 79-86, 2019.

[34] M. F. Ak, “A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications,” *Healthcare,* vol. 8, no. 2, p. 111, 2020.

[35] N. Al-Azzam and I. Shatnawi, “Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer,” *Annals of Medicine and Surgery,* vol. 62, pp. 53-64, 2021.

[36] I. Mihaylov, M. Nisheva and D. Vassilev, “Application of machine learning models for survival prognosis in breast cancer studies,” *Information,* vol. 3, no. 93, p. 10, 2019.

[37] A. S. Assiri, S. Nazir and S. A. Velastin, “Breast tumor classification using an ensemble machine learning method,” *Journal of Imaging,* vol. 6, no. 39, p. 6, 2020.

[38] H. Asri, H. Mousannif, H. Al Moatassime and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science,* vol. 83, pp. 1064-1069, 2016.

[39] J. Lötsch, R. Sipilä, T. Tasmuth, D. Kringel, A. M. Estlander, T. Meretoja and A. Ultsch, “Machine-learning-derived classifier predicts absence of persistent pain after breast cancer surgery with high accuracy,” *Breast cancer research and treatment,* vol. 171, no. 2, pp. 399-411, 2018.

[40] A. Massaro, F. Spano and A. Athanassiou, “Modeling and innovative technology of optical 3D antenna sensors as micro rectangular apertures,” *Opt Quant Electron,* vol. 44, p. 213–218, 2012.

[41] A. Massaro, N. Magaletti, V. Giardinelli, G. Cosoli, A. Leogrande and F. Cannone, “Original Data Vs High Performance Augmented Data for ANN Prediction of Glycemic Status in Diabetes Patients,” *SSRN 4082839,* 2022.

[42] A. Massaro, V. Giardinelli, G. Cosoli, N. Magaletti and A.-. Leogrande, “The Prediction of Hypertension Risk,” *SSRN,* 2022.

[43] A. Massaro, N. Magaletti, G. Cosoli, V. O. Giardinelli and A. Leogrande, “The Prediction of Diabetes,” *University Library of Munich, Germany,* vol. 113372, 2022.

[44] A. Massaro, Electronics in Advanced Research Industries: Industry 4.0 to Industry 5.0 Advances, Hoboken, New Jersey: John Wiley & Sons, 2021.

11. Figure Index

12. Table

Table 1. Descriptive Statistics. Those values show the numerical characteristics of the variables that are in the dataset. The analysis of the summary statistics is the first and most basic approach to data description able to shed lights on relevant connections and relationships among the variables.

Table 2. KPI Correlation Matrix. Those classifications are useful to individualize the variables that can be eligible for a metric modelling. Even if there is not a coincidence between the correlation index and the presence of statistical significance in metric models analysed with p-value.

Table 3. The Metric Results obtained either with binary either with non-binary metric model. The analysis shows that all the variables are positively associated with breast cancer with the sole exception of Concavity Standard Error that has a negative association.

13. Appendix

13.1 Data Description

In this paragraph, we describe data. Data are collected from a public database from Kaggle[1]. The dataset is based on 569 observations. Specifically, we have the following characteristics of the analysed dataset that are:

- Diagnosis: with an average value of 0,373, a median value equal to 0,000, a Standard Deviation value of 0,484, a minimum value of 0,000, a maximum value of 1,00;
- Radius Mean: with an average value of 707, a median value of 13,9 a standard deviation equal to 2,43e+003, a minimum value of 7,76, a maximum value of 9,90e+003;
- Texture Mean: with an average value equal to 19,3, a Median value of 18,8, a Standard Deviation equal to 4,30, a Minimum value of 9,71, a maximum value of 39,3;
- Perimeter Mean: with an average value of 92,00, a Median value of 86,2, a Standard Deviation of 24,3, a Minimum value of 43,8, a Maximum value of 39,3;
- Area mean: with an average value of 655, a median value of 551 a standard deviation equal to 352, a minimum value of 144,00, a maximum value of 2,50e+003;
- Smoothness mean: with an average value of 0,0964, a median value of 0,0959, a standard deviation equal to 0,0141, a minimum value of 0,0526, a medium value of 0,163.
- Compactness Mean: with an average value of 0,104, with a median value of 0,0926, a standard deviation equal to 0,0528, a minimum value of 0,0194, a maximum value of 0,345;
- Concavity Mean: with an average value of 0,0888; a median value of 0,615, a standard deviation equal to 0,0797, a minimum value of 0,0194, a maximum value of 0,427;
- Concave Points Mean: with an average value of 0,0489, a median value of 0,0335, a standard deviation of 0,0388, a minimum value of 0,000, a maximum value of 0,201;

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

- Symmetry Mean: with an average value of 0,181, a median value of 0,179, a standard deviation of 0,0274, a minimum value of 0,106, a maximum value of 0,304;
- Fractal Dimension Mean: with an average value of 0,0628, a median value of 0,0615, a standard deviation equal to 0,00706, a minimum 0,0500, a maximum value of 0,0974;
- Radius Standard Deviation: with an average value of 48,6, a median value of 0,324, a standard deviation of 263, a minimum value of 0,112, a maximum value of 2,87e+003;
- Texture Standard Deviation: with an average value of 803, a median value of 1,03e+003; a standard deviation 844, a minimum value of 0,360, a maximum value of 4,89e+003;
- Perimeter Standard Deviation: with an average value of 2,55e+00, a median value of 2,16e+003, a standard deviation equal to 1,76e+003, a minimum value of 0,757, a maximum value of 9.81e+003;
- Smoothness Standard Deviation: with an average value of 0,00704, a median value of 0,00638, a standard deviation of 0,00300, a minimum value of 0,00171, a maximum value of 0,0311;
- Compactness Standard Deviation: with an average value of 0,00704, a median value of 0,00638, a standard deviation of 0,0179, a minimum value of 0,00225, a maximum value of 0,135;
- Concavity Standard Deviation: with an average value of 0,0319, a median value of 0,0259, a standard deviation of 0,0302, a minimum value of 0,000, a maximum value of 0,396;
- Concavity Standard Deviation: with an average value of 0,0319, a median of 0,0259, a standard deviation of 0,0302, a minimum value of 0,000 a maximum value of 0,396.
- Concave Points Standard Deviation: with an average value of 0,0118, a median value of 0,0109, a standard deviation of 0,00617, a minimum value of 0,000 a maximum value of 0,0528;
- Symmetry Standard Deviation: with an average value of 0,0205, a median of 0,0187, a standard deviation of 0,0827, a minimum value of 0,00788, a maximum value of 0,0790;
- Fractal Dimension Standard Deviation: with an average value of 0,00279, a median value of 0,00219, a standard deviation of 0,00265, a minimum value of 0,000895 and a maximum value of 0,0298;
- Radius Worst: with an average value of 315, a median value of 15,2, a standard deviation of 1,66e+003, a minimum value of 7,93, a maximum value of 9,98e+003;
- Texture worst: with an average value of 25,7, a median value of 25,4, a standard deviation of 6,15, a minimum value of 12,00, a maximum value of 49,5;
- Perimeter worst: with an average value of 107, a median value of 97,7, a standard deviation of 33,6, a minimum value of 50,4, a maximum value of 251.
- Area Worst: with an average value of 881, a median value of 687, a standard deviation of 569, a minimum value of 185, a maximum value of 4,25e+003;
- Smoothness Worst: with an average value of 0,132, a median value of 0,131, a standard deviation of 0,0228, a minimum value of 0,0712, a maximum value of 0,223;
- Compactness Worst: with an average value of 2,11, a median value of 0,212, a standard deviation of 69,9 a minimum value of 0,0000 a maximum value of 1,25e+0003;
- Concave Points Worst; with an average value of 0,115, a median value of 0,0999, a standard deviation of 0,0657, a minimum value of 0,0000 a maximum value of 1,25e+003;
- Symmetry Worst: with an average value of 0,290, a median value of 0,282, a standard deviation of 0,0657, a minimum value of 0,000 a maximum value of 0,291;

- Fractal Dimension Worst: with an average value of 0,0839, a median value of 0,0800, a standard deviation of 0,0181, a minimum value of 0,0500, a maximum value of 0,207.

Descriptive statistics, using observations 1 - 569						
Rank	**Variabile**	**Average**	**Median**	**Standard Deviation**	**Min**	**Max**
1	*Diagnosis*	0,373	0	0,484	0	1
2	*Radius Mean*	707	13,9	2,43E+03	7,76	9,90E+03
3	*Texture Mean*	19,3	18,8	4,3	9,71	39,3
4	*Perimeter Mean*	92	86,2	24,3	43,8	189
5	*Area Mean*	655	551	352	144	2,50E+03
6	*Smoothness Mean*	0,0964	0,0959	0,0141	0,0526	0,163
7	*Compactness Mean*	0,104	0,0926	0,0528	0,0194	0,345
8	*Concavity Mean*	0,0888	0,0615	0,0797	0	0,427
9	*Concave Points Mean*	0,0489	0,0335	0,0388	0	0,201
10	*Symmetry Mean*	0,181	0,179	0,0274	0,106	0,304
11	*Fractal Dimension Mean*	0,0628	0,0615	0,00706	0,05	0,0974
12	*Radius Standard Deviation*	48,6	0,324	263	0,112	2,87E+03
13	*Texture Standard Deviation*	803	1,03E+03	844	0,36	4,89E+03
14	*Perimeter Standard Deviation*	2,55E+03	2,16E+03	1,76E+03	0,757	9,81E+03
15	*Area Standard Deviation*	316	25,8	1,53E+03	10,1	9,83E+03
16	*Smoothness Standard Deviation*	0,00704	0,00638	0,003	0,00171	0,0311
17	*Compactness Standard Deviation*	0,0255	0,0204	0,0179	0,00225	0,135
18	*Concavity Standard Deviation*	0,0319	0,0259	0,0302	0	0,396
19	*Concavepoints Standard Deviation*	0,0118	0,0109	0,00617	0	0,0528
20	*Symmetry Standard Deviation*	0,0205	0,0187	0,00827	0,00788	0,079
21	*Fractal Dimension Standard Deviation*	0,00379	0,00319	0,00265	0,0009	0,0298
22	*Radius Worst*	315	15,2	1,66E+03	7,93	9,98E+03
23	*Texture Worst*	25,7	25,4	6,15	12	49,5
24	*Perimeter Worst*	107	97,7	33,6	50,4	251
25	*Area Worst*	881	687	569	185	4,25E+03
26	*Smoothness Worst*	0,132	0,131	0,0228	0,0712	0,223
27	*Compactness Worst*	2,11	0,212	44,3	0,0273	1,06E+03
28	*Concavity Worst*	4,41	0,227	69,9	0	1,25E+03
29	*Concavepoints Worst*	0,115	0,0999	0,0657	0	0,291
30	*Symmetry Worst*	0,29	0,282	0,0619	0,157	0,664
31	*Fractal Dimension Worst*	0,0839	0,08	0,0181	0,055	0,207

Table 1. Descriptive Statistics. Those values show the numerical characteristics of the variables that are in the dataset. The analysis of the summary statistics is the first and most basic approach to data description able to shed lights on relevant connections and relationships among the variables.

As it is clear from the analysis of the dataset there are essentially 10 variables that are considered in three different characteristics that are: mean, standard deviation and worst. That choice of the dataset builder can be considered as a redundant proposition of data. However, should be considered

effectively as the necessity to verify which numerical manifestation of the variable is essential for the determination of the cause that can predict the manifestation of malign breast cancer.

13.2 Correlation Matrix

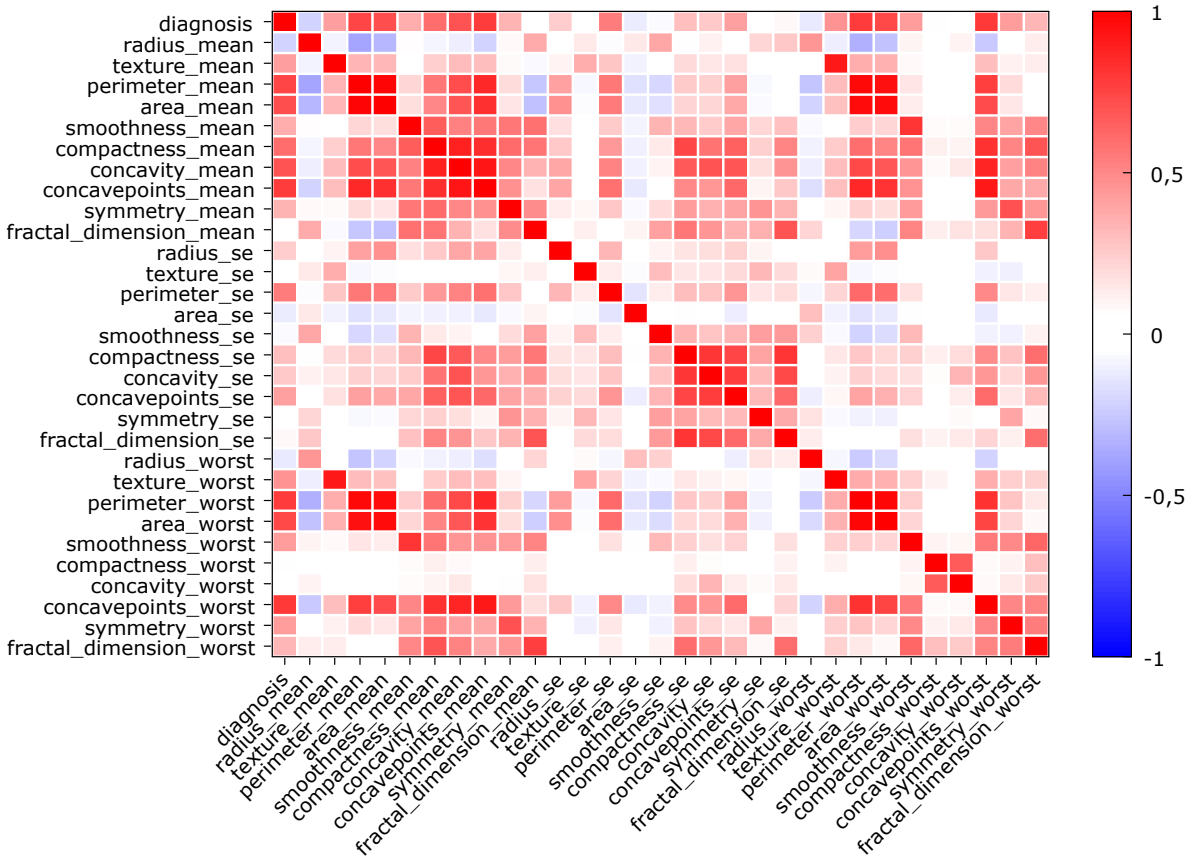


Figure 7. Correlation matrix. The correlation matrix has the ability to show the connections among the variables expressed in terms of correlation index. Specifically the graphical representation of the correlation matrix is useful to visualize the positive and negative correlations, and the intensity of the colour is a sign of the relevance of the correlation.

We analyse the correlation matrix considering the presence of three different targets i.e.:

- Strong correlation: that are values of the correlations in the following interval [0,70; 1,00]; [-0,70; -1,00]. The variables that are in the following intervals can be considered eligible as candidate to either a descriptive and e predictive model;
- Average correlation: that are values of correlation in the interval [0,50; 0,70] and [-0,50;-0,70]. The variables that have such a kind of correlation can also be considered eligible for a descriptive and predictive model even if those represent a second best in terms of correlation in respect to the variables that are present in the previous category;
- Weak Correlation: is the condition of variables that are in the following interval i.e.: [0,00; 0,50] and [-0,00;-0,50]. Those variables are considered as essentially as the last relevant variables in terms of correlation.

Choice of Correlation

Strong Correlation	]0,70;1,00];]-0,70;-1,00]
Average Correlation	]0,50;0,70];]-0,50;-0,70]
Weak Correlation	[0,50;-0,50]

Table 2. KPI Correlation Matrix. Those classifications are useful to individualize the variables that can be eligible for a metric modelling. Even if there is not a coincidence between the correlation index and the presence of statistical significance in metric models analysed with p-value.

Based on this kind of classification it is possible to find some relevant correlation among the different variables to optimize the process of modelling. Specifically we found that the main relevant variables in the model are:

- Concave Points Worst that has a correlation with breast cancer of a value of 0,7936;
- Perimeters Worst that has a positive correlation with breast cancer with a value of 0,7829;
- Concave Points Mean that has a positive correlation with breast cancer equal to 0,7766;
- Perimeter Mean that has a positive correlation with breast cancer equal to 0,7426;
- Area worst that as a positive correlation with breast cancer equal to 0,7338;
- Area Mean that has a positive correlation with breast cancer equal to 0,709

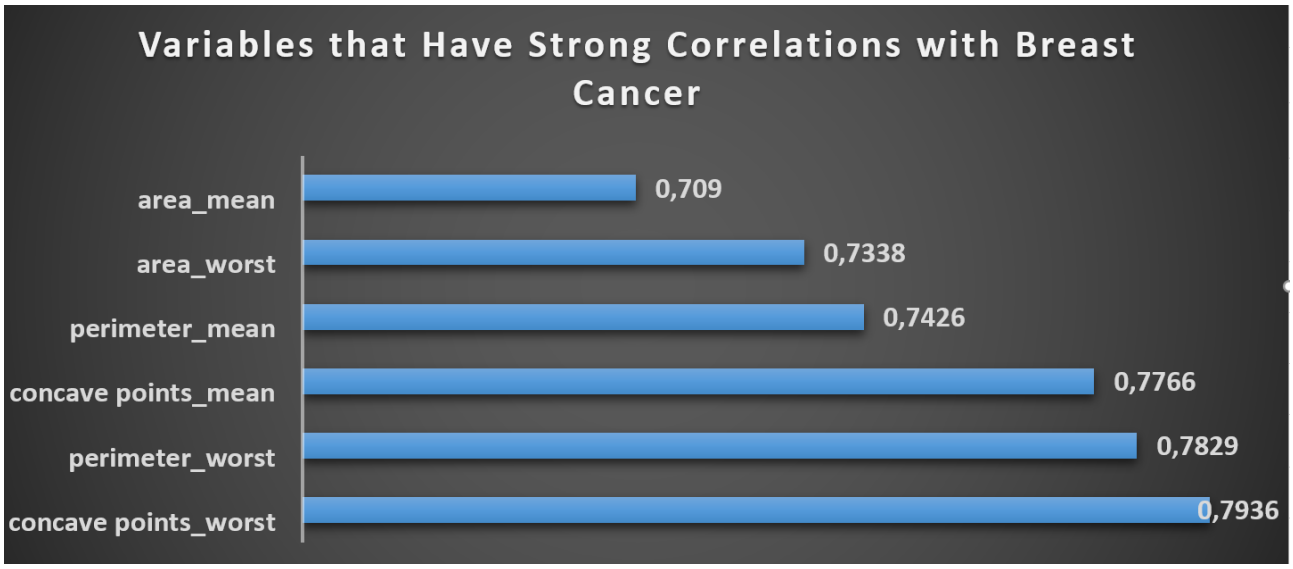


Figure 8. Variables that have a strong correlation with Breast Cancer. Strong Correlation is defined as a value of the correlation index that is in the intervals]0,70;1] and]-0,70;-1].

Furthermore, we found three different variables that have a medium correlation with respect to brain cancer that are:

- Concavity mean with a positive correlation equal to 0,6964;
- Compactness Mean with a positive correlation equal to 0,5965;
- Perimeter Standard Error with a positive correlation equal to 0,5345;

Finally we have 21 variables that are weakly correlated in respect to breast cancer that are:

- Texture Worst: with a positive correlation with breast cancer equal to 0,4569;
- Smoothness Worst: with a positive correlation with breast cancer equal to 0,4125;
- Symmetry Worst: with a positive correlation with breast cancer equal to 0,4163;
- Texture Mean: with a positive correlation with breast cancer equal to 0,4152;
- Concave Points Standard Error: with a positive correlation with breast cancer equal to 0,408;
- Smoothness Mean: with a positive correlation with breast cancer equal to 0,3586;

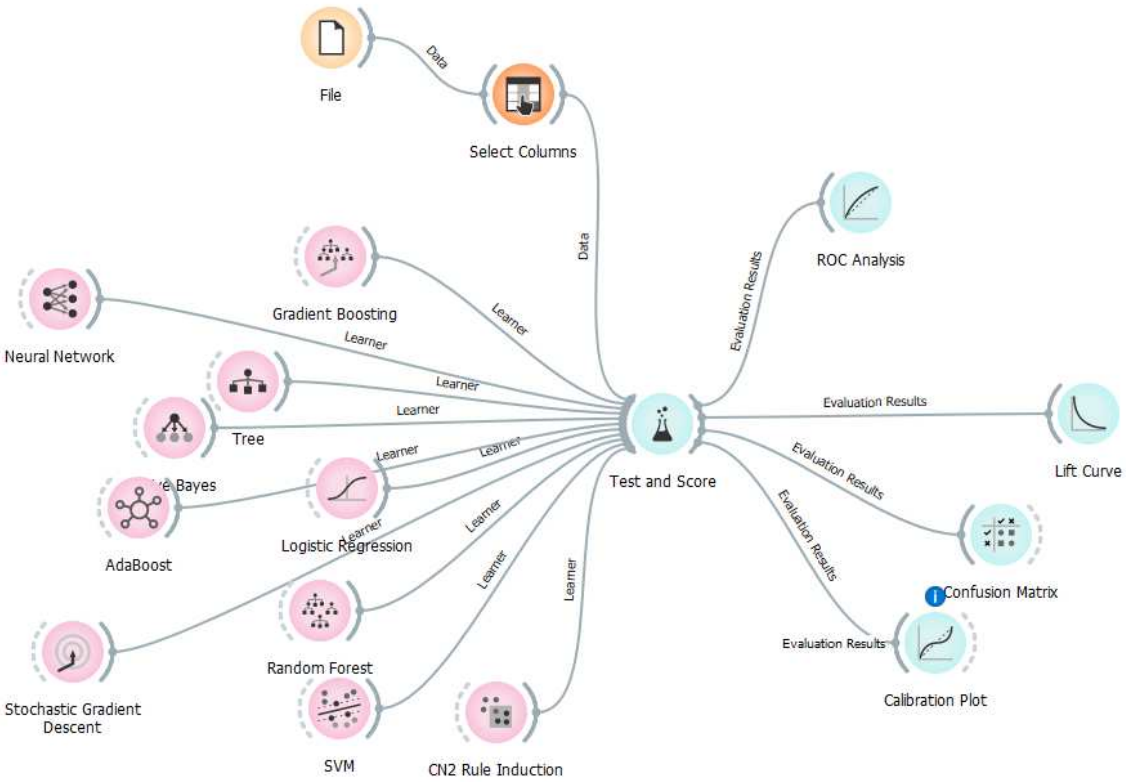
- Symmetry Mean: with a positive correlation with breast cancer equal to 0,3305;
- Fractal Dimension Worst: with a positive correlation with breast cancer equal to 0,3239;
- Compactness Standard Error: with a positive correlation with breast cancer equal to 0,293;
- Concavity Standard Error: with a positive correlation with breast cancer 0,2537;
- Radius Standard Error: with a positive correlation with breast cancer equal to 0,2388;
- Fractal Dimension Standard Error: with a positive correlation with breast cancer equal to 0,078;
- Compactness Worst: with a positive correlation with breast cancer equal to 0,0565;
- Concavity Worst: with a positive correlation with breast cancer equal to 0,0138;
- Symmetry Standard Error: with a negative correlation with breast cancer equal to -0,0065;
- Texture Standard Error: with a negative correlation with breast cancer equal to -0,0108;
- Fractal Dimension Mean: with a negative correlation with breast cancer equal to -0,0128;
- Smoothness Standard Error: with a negative correlation with breast cancer equal to -0,067;
- Area Standard Error: with a negative correlation with breast cancer equal to -0,1226;
- Radius Worst: with a negative correlation with breast cancer equal to -0,137;
- Radius Mean: with a negative correlation with breast cancer equal to -0,2188.

As we can see each variable has at least three different definitions that are the mean value, the standard error, and the worst value. Those differences are introduced to verify effectively which metric can effectively have a rule in definition of the main causes of the breast cancer. As we will see in the next paragraph some variable is relevant in the sense of standard error, some other in these sense of the mean value and some other in the sense of the worst value. However, as we can see in the next section to estimate the determinants that have an impact in terms of breast cancer it is not sufficient to consider exclusively the value of the variables that show strong correlations but also the model also present variables that have medium and lower correlation. The fact that there are different values between high correlations and the variables predicted in the regression analysis should not be considered has a counterfactual. In effect while correlation tends to find the linear relationships among variables, the statistical meaning of the regression analysis consists in the determination of a set of variables that can be used to predict the value of the investigated variable. In this sense the fact that some variables that have high correlation are not present in the main regression model, and vice versa, should be considered because of the different goals that correlation and regression models tend to achieve.

Analysis of Correlation Matrix			
Ranking	Variables	Correlation Index	Typology of Correlation
	Diagnosis	1	Variable of Interest
1	concave points_worst	0,7936	Strong Correlation
2	perimeter_worst	0,7829	
3	concave points_mean	0,7766	
4	perimeter_mean	0,7426	
5	area_worst	0,7338	
6	area_mean	0,709	
7	concavity_mean	0,6964	Medium Correlation
8	compactness_mean	0,5965	
9	perimeter_se	0,5345	
10	texture_worst	0,4569	Weak Correlation
11	smoothness_worst	0,4215	
12	symmetry_worst	0,4163	
13	texture_mean	0,4152	
14	concave points_se	0,408	
15	smoothness_mean	0,3586	
16	symmetry_mean	0,3305	
17	fractal_dimension_worst	0,3239	
18	compactness_se	0,293	
19	concavity_se	0,2537	
20	radius_se	0,2388	
21	fractal_dimension_se	0,078	
22	compactness_worst	0,0565	
23	concavity_worst	0,0138	
24	symmetry_se	-0,0065	
25	texture_se	-0,0108	
26	fractal_dimension_mean	-0,0128	
27	smoothness_se	-0,067	
28	area_se	-0,1226	
29	radius_worst	-0,137	
30	radius_mean	-0,2188	

Figure 9. A Synthesis of the main correlations that variables have in respect to breast cancer. The correlations are considered based on three different criteria that are: strong correlation, medium correlation and weak correlation. The variables that show the presence of strong correlation should be eligible for modelling.

13.3 Accuracy, Test and Score and ROC Curve

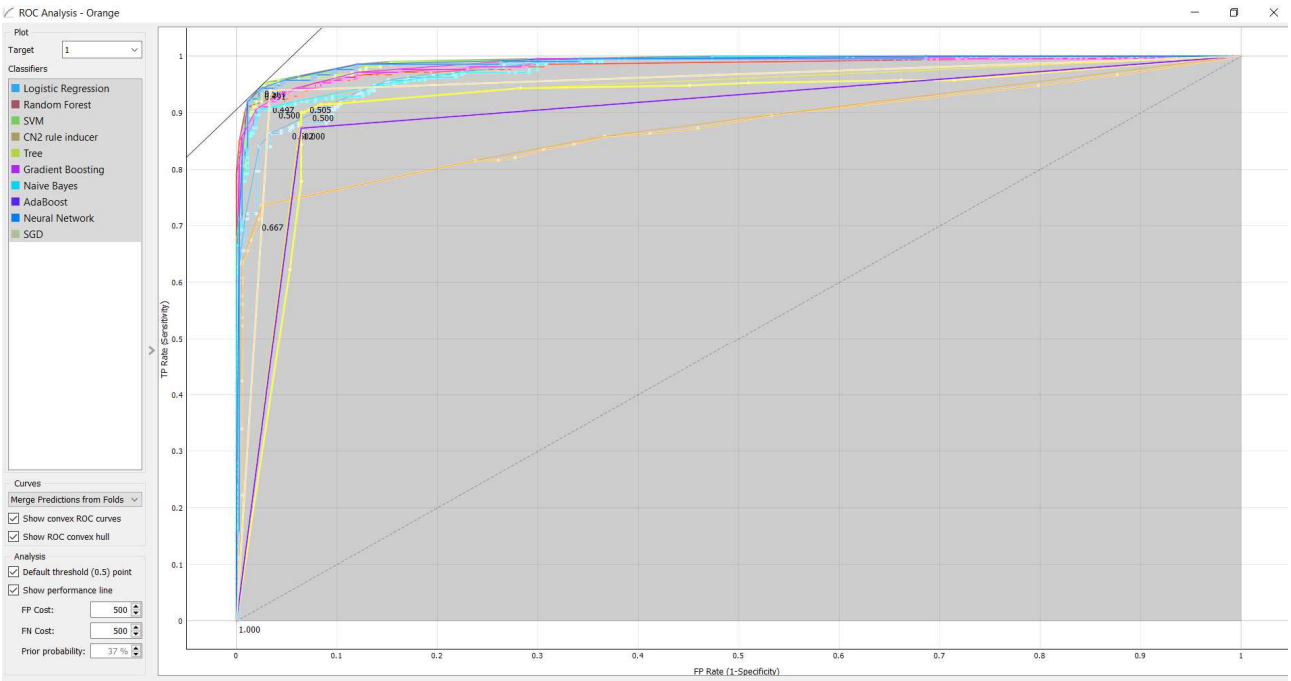


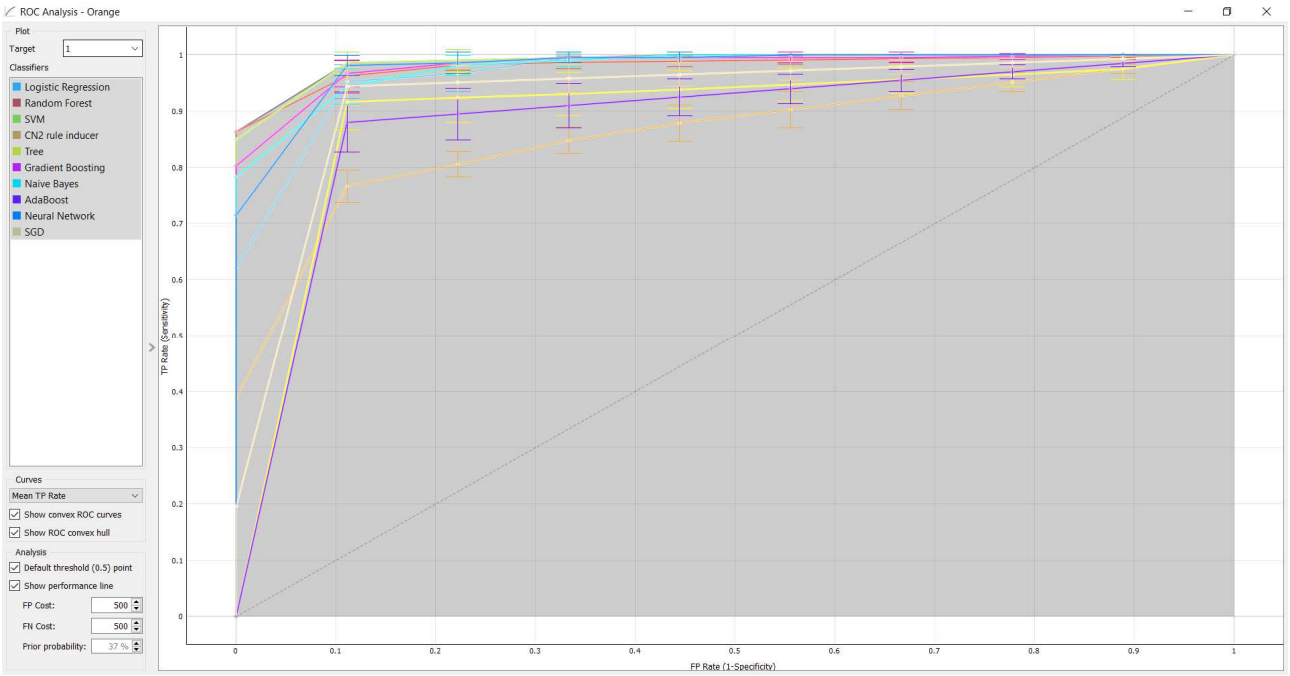
Evaluation results for target (None, show average over classes)

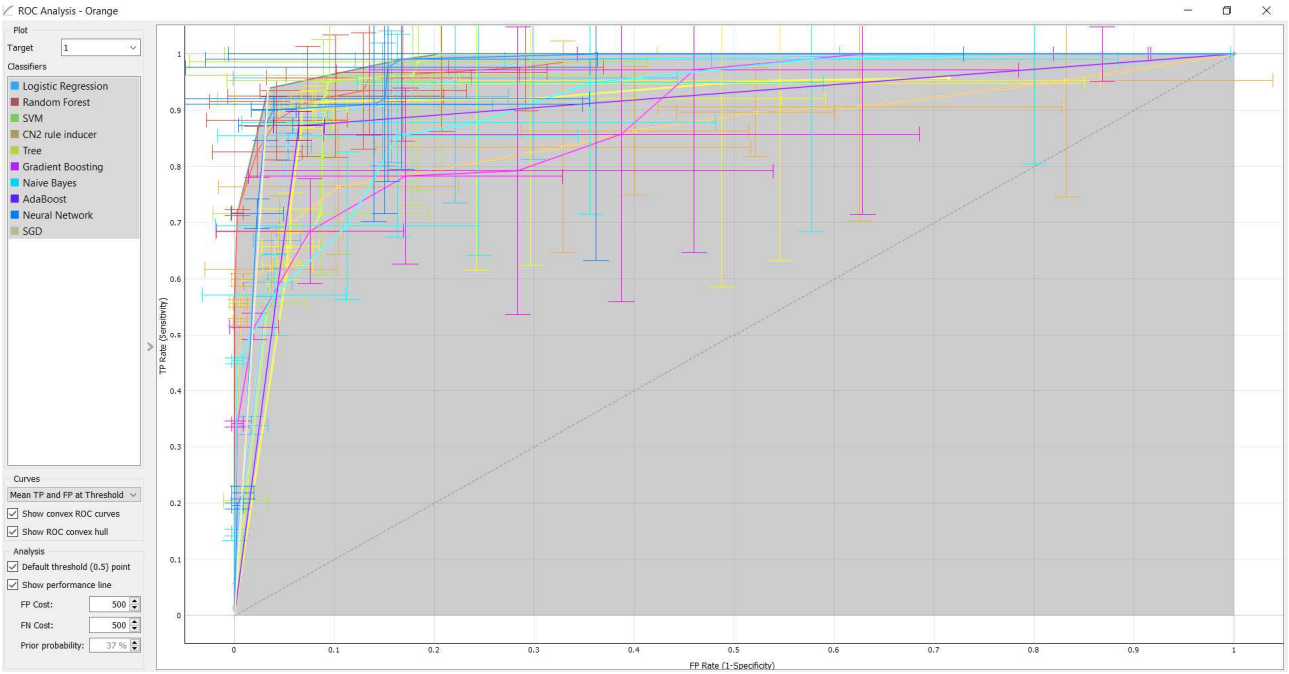
Model	AUC	CA	F1	Precision	Recall
Tree	0.917	0.921	0.921	0.921	0.921
SVM	0.988	0.960	0.959	0.960	0.960
SGD	0.953	0.956	0.956	0.956	0.956
Random Forest	0.981	0.946	0.945	0.945	0.946
Neural Network	0.987	0.961	0.961	0.961	0.961
Naive Bayes	0.980	0.926	0.926	0.927	0.926
Logistic Regression	0.973	0.919	0.919	0.919	0.919
Gradient Boosting	0.984	0.947	0.947	0.947	0.947
CN2 rule inducer	0.872	0.879	0.874	0.888	0.879
AdaBoost	0.904	0.912	0.912	0.912	0.912

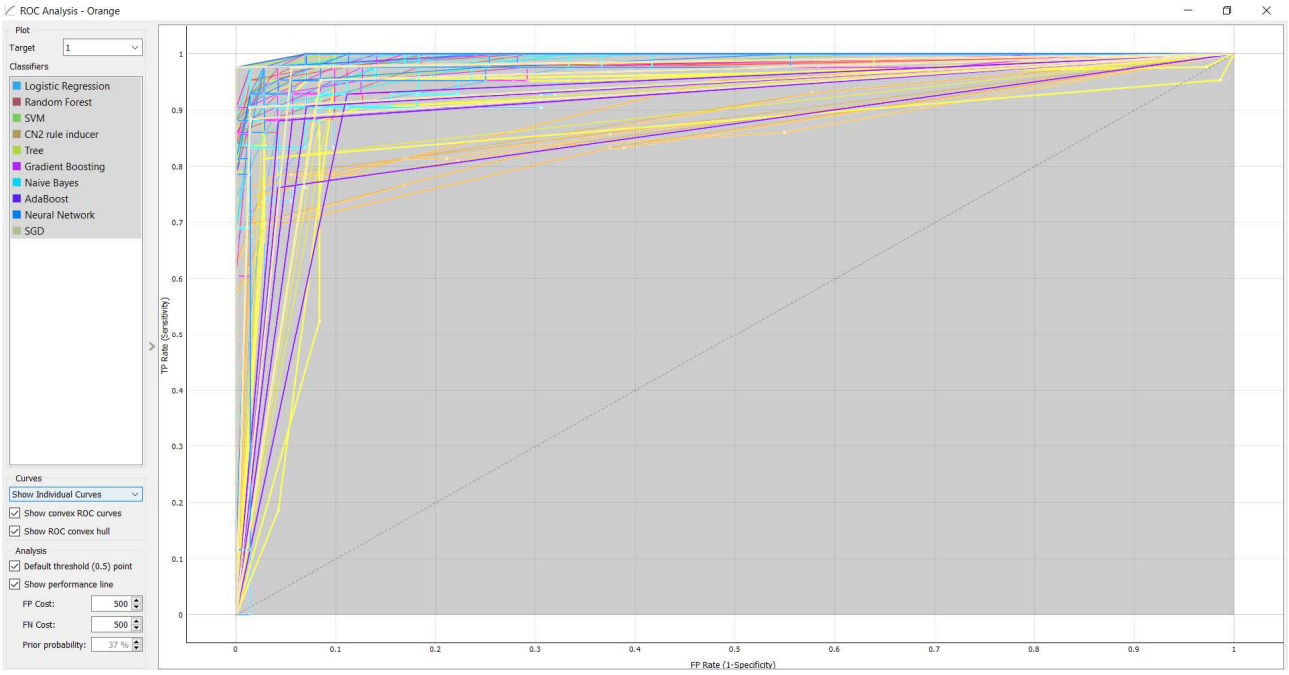
Compare models by: Area under ROC curve — Negligible diff.:

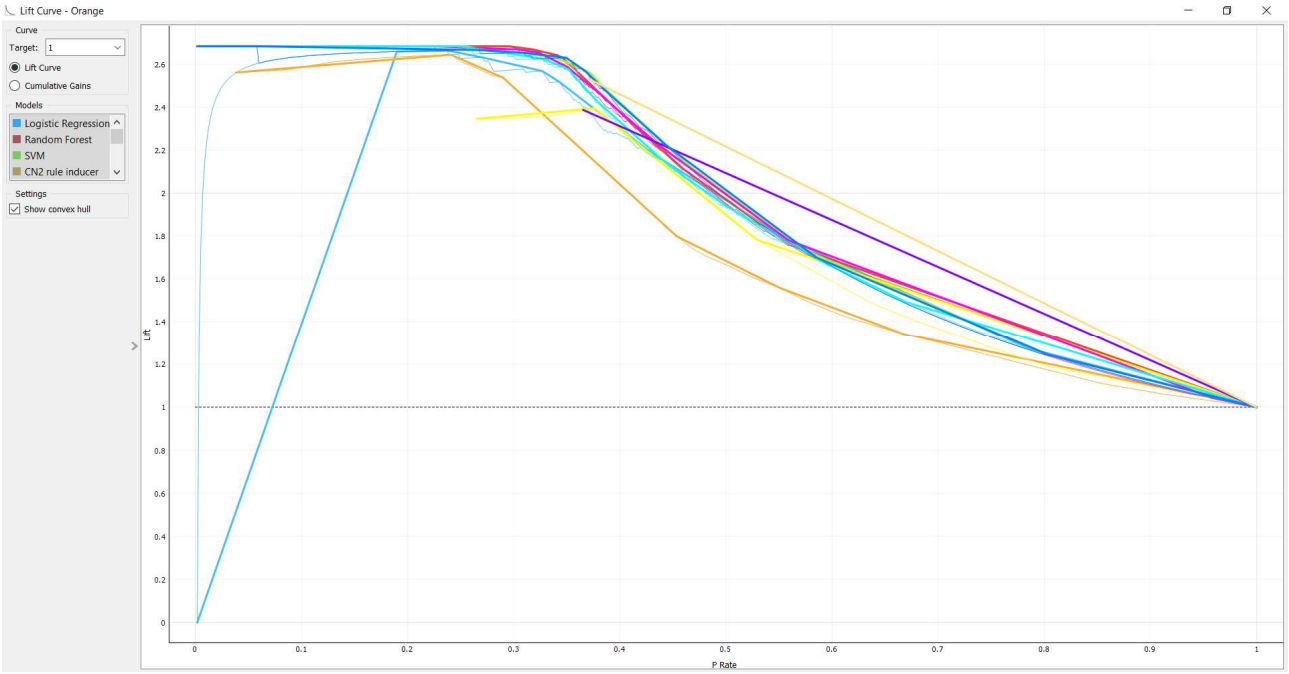
	Tree	SVM	SGD	Random Forest	Neural Network	Naive Bayes	Logistic Regression	Gradient Boosting	CN2 rule inducer	AdaBoost
Tree		0.005	0.007	0.008	0.004	0.011	0.006	0.007	0.981	0.579
SVM	0.995		0.975	0.812	0.607	0.883	0.955	0.908	0.999	0.999
SGD	0.993	0.025		0.055	0.019	0.064	0.081	0.037	0.998	0.964
Random Forest	0.992	0.188	0.945		0.149	0.537	0.745	0.380	0.999	0.995
Neural Network	0.996	0.393	0.981	0.851		0.818	0.935	0.743	1.000	0.998
Naive Bayes	0.989	0.117	0.936	0.463	0.182		0.724	0.348	0.999	0.998
Logistic Regression	0.994	0.045	0.919	0.255	0.065	0.276		0.061	0.999	1.000
Gradient Boosting	0.993	0.092	0.963	0.620	0.257	0.652	0.939		0.999	0.999
CN2 rule inducer	0.019	0.001	0.002	0.001	0.000	0.001	0.001	0.001		0.076
AdaBoost	0.421	0.001	0.036	0.005	0.002	0.002	0.000	0.001	0.924	

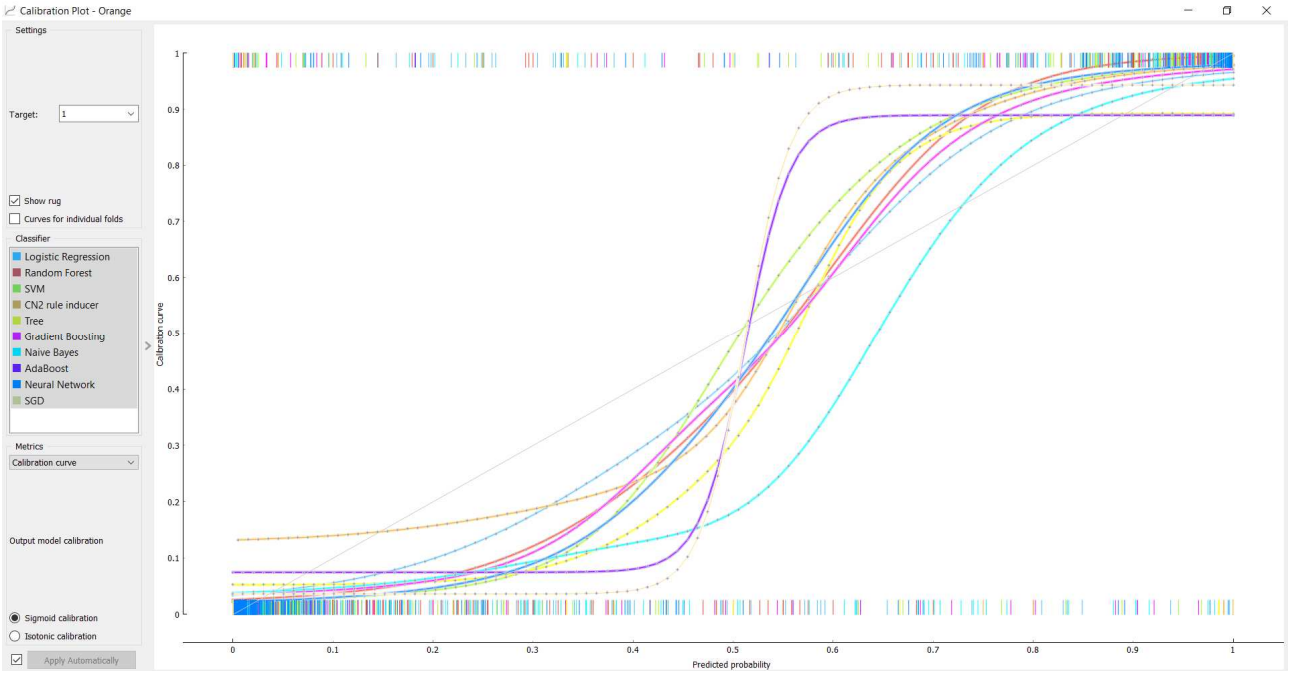

ROC Analysis - Orange
Plot
Target
1
Classifiers
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Neural Network
SGD
Curves
Merge Predictions from Folds
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost:
500
FN Cost:
500
Prior probability:
37 %
0.667
1.000
TP Rate (Sensitivity)
FP Rate (1-Specificity)

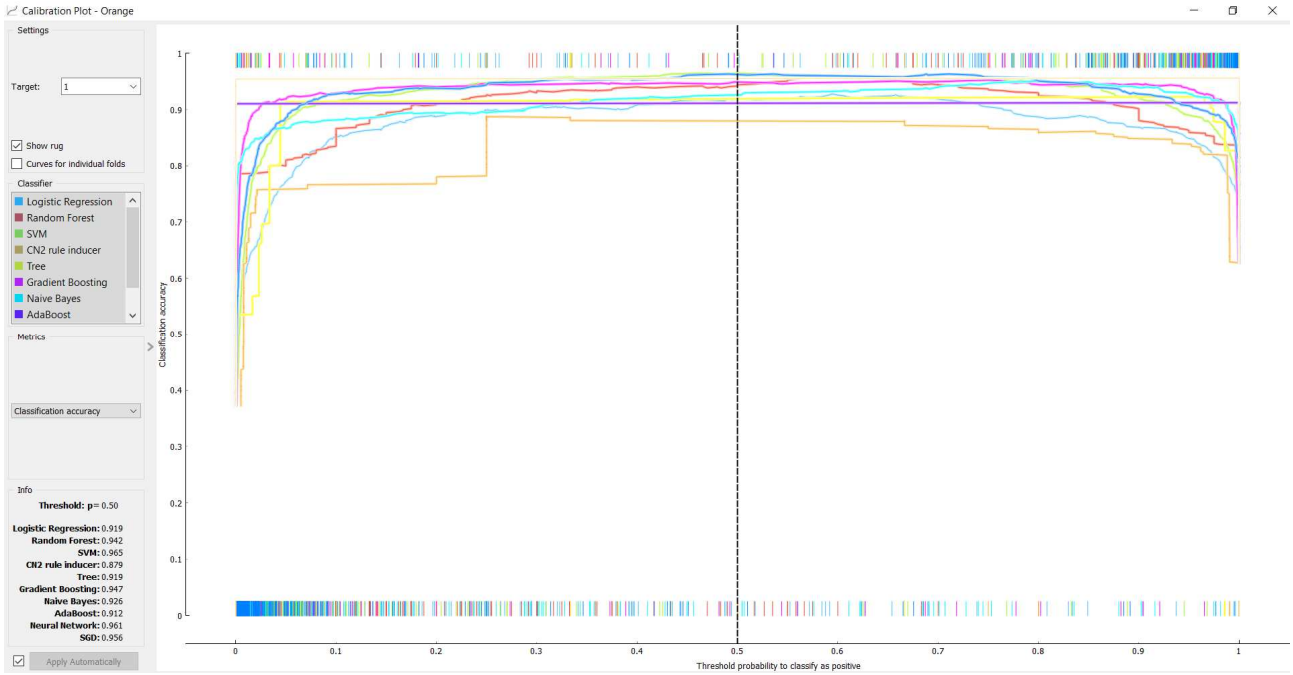

ROC Analysis - Orange
Plot
Target
1
Classifiers
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Neural Network
SGD
Curves
Mean TP Rate
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost:
500
FN Cost:
500
Prior probability:
37 %
TP Rate (Sensitivity)
FP Rate (1-Specificity)

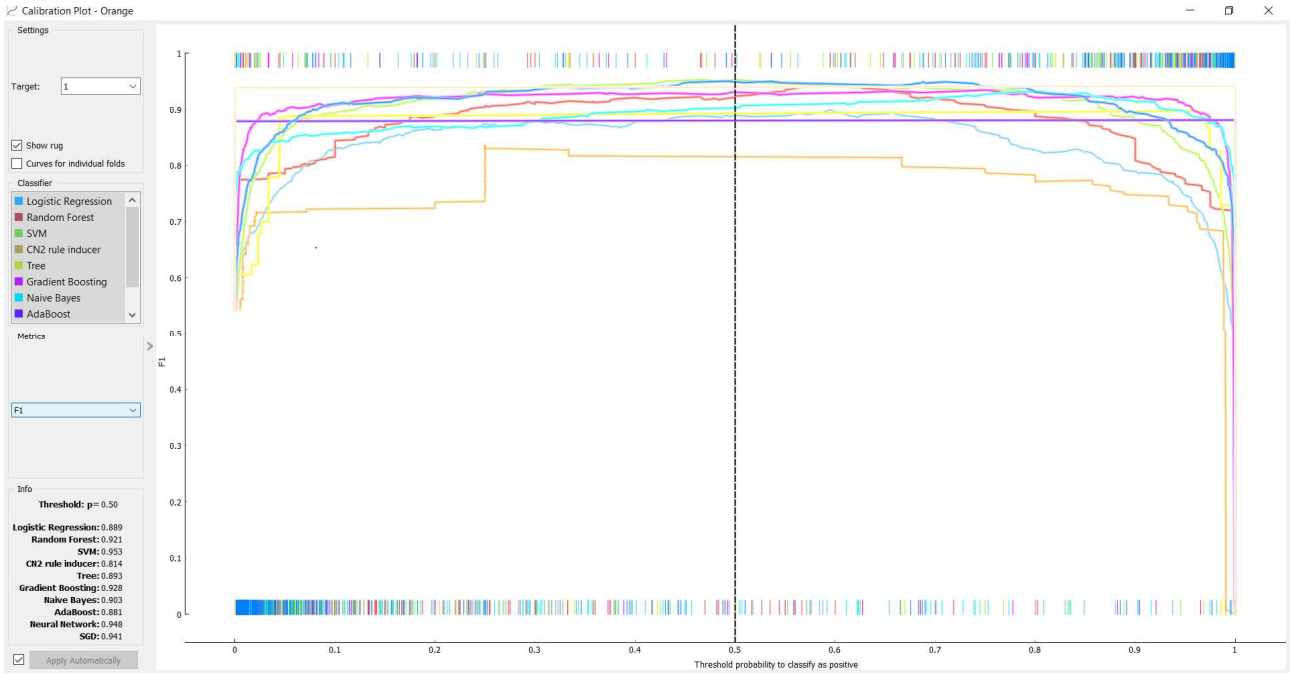

ROC Analysis - Orange
Plot
Target
1
Classifiers
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Neural Network
SGD
Curves
Mean TP and FP at Threshold
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost:
500
FN Cost:
500
Prior probability:
37 %
TP Rate (Sensitivity)
FP Rate (1-Specificity)

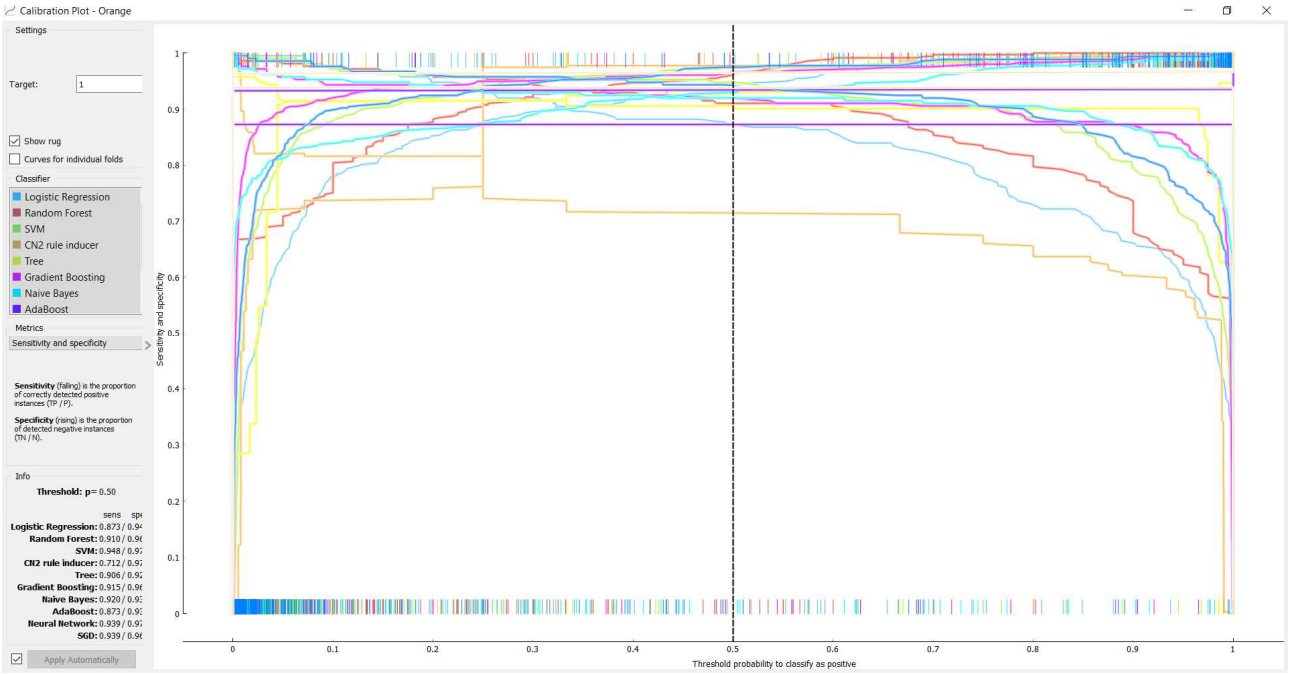

ROC Analysis - Orange
Plot
Target
1
Classifiers
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Neural Network
SGD
Curves
Show Individual Curves
Show convex ROC curves
Show ROC convex hull
Analysis
Default threshold (0.5) point
Show performance line
FP Cost:
500
FN Cost:
500
Prior probability:
37 %
TP Rate (Sensitivity)
FP Rate (1-Specificity)

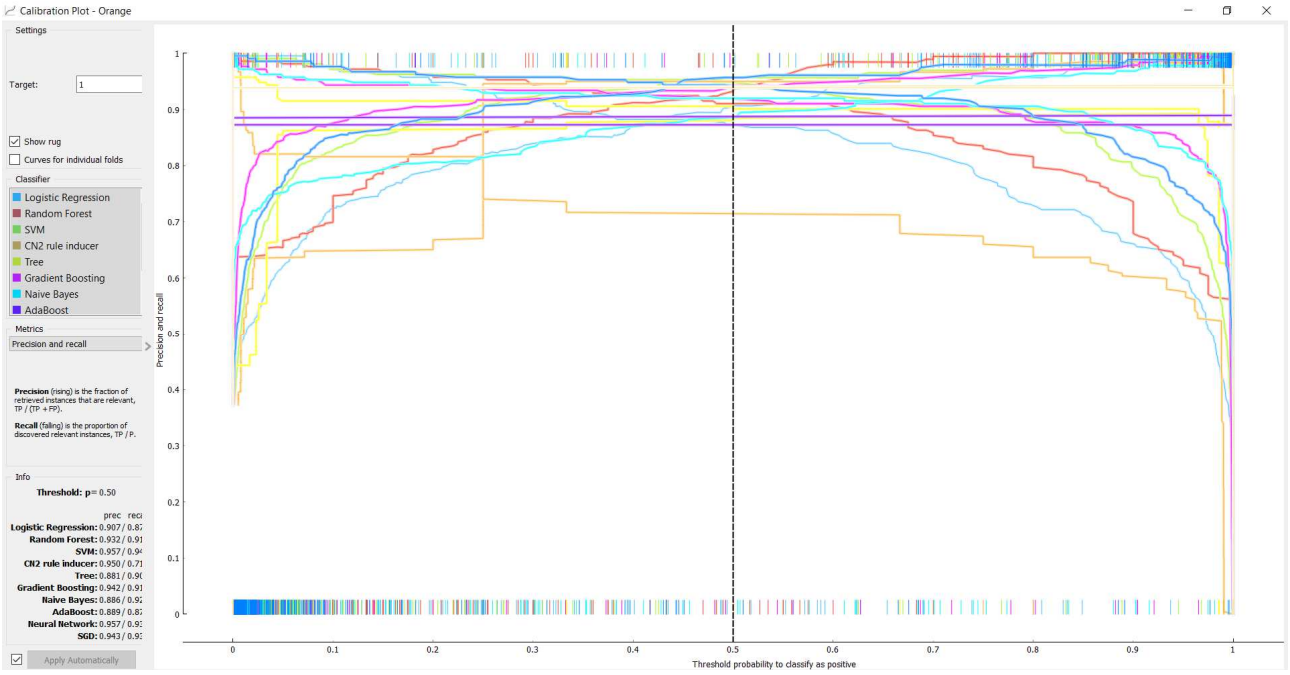

Lift Curve - Orange
Curve
Target: 1
Lift Curve
Cumulative Gains
Models
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Settings
Show convex hull
Lift
P Rate

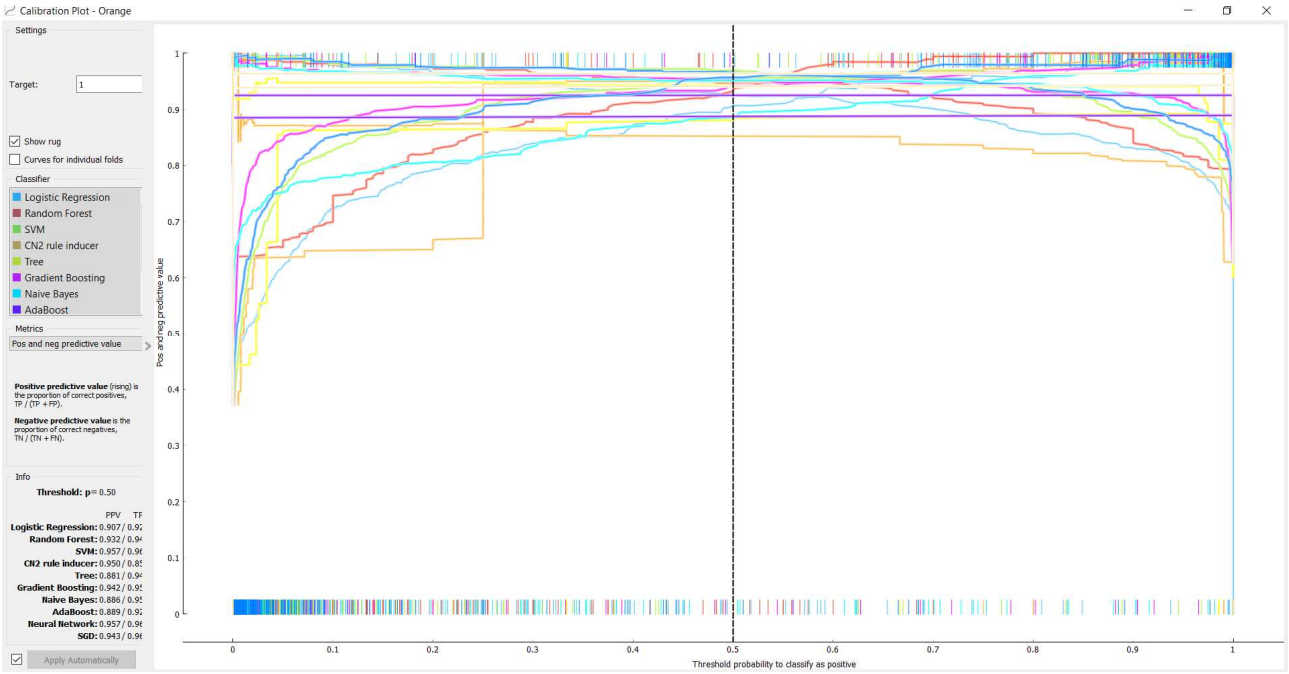

Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Neural Network
SGD
Metrics
Calibration curve
Output model calibration
Sigmoid calibration
Isotonic calibration
Apply Automatically
Calibration curve
Predicted probability

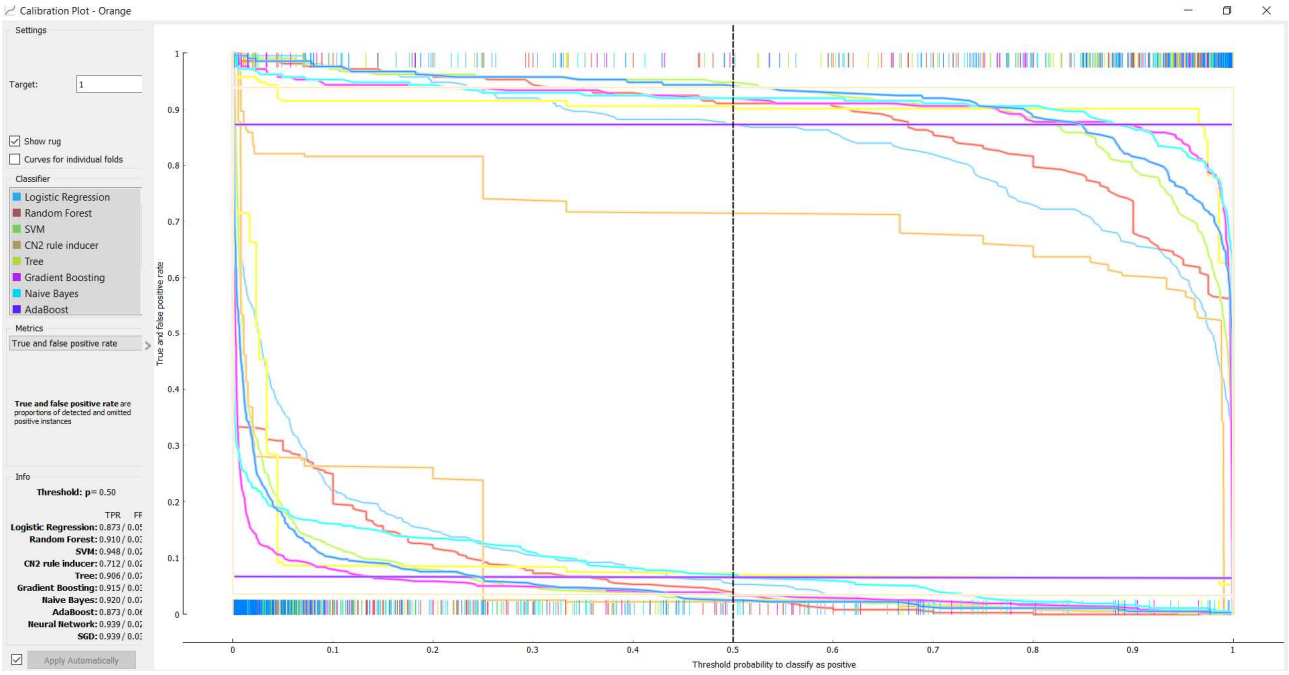

Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
Classification accuracy
Info
Threshold: p= 0.50
Logistic Regression: 0.919
Random Forest: 0.942
SVM: 0.965
CN2 rule inducer: 0.879
Tree: 0.919
Gradient Boosting: 0.947
Naive Bayes: 0.926
AdaBoost: 0.912
Neural Network: 0.961
SGD: 0.956
Apply Automatically
Classification accuracy
Threshold probability to classify as positive


Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
F1
Info
Threshold: p= 0.50
Logistic Regression: 0.889
Random Forest: 0.921
SVM: 0.953
CN2 rule inducer: 0.814
Tree: 0.893
Gradient Boosting: 0.928
Naive Bayes: 0.903
AdaBoost: 0.881
Neural Network: 0.948
SGD: 0.941
Apply Automatically
F1
Threshold probability to classify as positive


Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
Sensitivity and specificity
Sensitivity (falling) is the proportion of correctly detected positive instances (TP / P).
Specificity (rising) is the proportion of detected negative instances (TN / N).
Info
Threshold: p= 0.50
sens spe
Logistic Regression: 0.873 / 0.94
Random Forest: 0.910 / 0.9
SVM: 0.948 / 0.9
CN2 rule inducer: 0.712 / 0.9
Tree: 0.906 / 0.9
Gradient Boosting: 0.915 / 0.9
Naive Bayes: 0.920 / 0.9
AdaBoost: 0.873 / 0.9
Neural Network: 0.939 / 0.9
SGD: 0.939 / 0.9
Apply Automatically
Sensitivity and specificity
Threshold probability to classify as positive


Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
Precision and recall
Precision (rising) is the fraction of retrieved instances that are relevant, TP / (TP + FP).
Recall (falling) is the proportion of discovered relevant instances, TP / P.
Info
Threshold: p= 0.50
prec rec
Logistic Regression: 0.907 / 0.87
Random Forest: 0.932 / 0.91
SVM: 0.957 / 0.94
CN2 rule inducer: 0.950 / 0.71
Tree: 0.881 / 0.90
Gradient Boosting: 0.942 / 0.91
Naive Bayes: 0.886 / 0.92
AdaBoost: 0.889 / 0.87
Neural Network: 0.957 / 0.93
SGD: 0.943 / 0.93
Apply Automatically
Precision and recall
Threshold probability to classify as positive


Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
Pos and neg predictive value
Positive predictive value (rising) is the proportion of correct positives, TP / (TP + FP).
Negative predictive value is the proportion of correct negatives, TN / (TN + FN).
Info
Threshold: p= 0.50
PPV TF
Logistic Regression: 0.907 / 0.9
Random Forest: 0.932 / 0.9
SVM: 0.957 / 0.9
CN2 rule inducer: 0.950 / 0.8
Tree: 0.881 / 0.9
Gradient Boosting: 0.942 / 0.9
Naive Bayes: 0.886 / 0.9
AdaBoost: 0.889 / 0.9
Neural Network: 0.957 / 0.9
SGD: 0.943 / 0.9
Apply Automatically
Pos and neg predictive value
Threshold probability to classify as positive


Calibration Plot - Orange
Settings
Target: 1
Show rug
Curves for individual folds
Classifier
Logistic Regression
Random Forest
SVM
CN2 rule inducer
Tree
Gradient Boosting
Naive Bayes
AdaBoost
Metrics
True and false positive rate
True and false positive rate are proportions of detected and omitted positive instances
Info
Threshold: p= 0.50
TPR FF
Logistic Regression: 0.873 / 0.0
Random Forest: 0.910 / 0.0
SVM: 0.948 / 0.0
CN2 rule inducer: 0.712 / 0.0
Tree: 0.906 / 0.0
Gradient Boosting: 0.915 / 0.0
Naive Bayes: 0.920 / 0.0
AdaBoost: 0.873 / 0.0
Neural Network: 0.939 / 0.0
SGD: 0.939 / 0.0
Apply Automatically
True and false positive rate
Threshold probability to classify as positive