Classification Model Evaluation Metrics

Article *in* International Journal of Advanced Computer Science and Applications · July 2021

DOI: 10.14569/IJACSA.2021.0120670

1 author:

Zeljko Vujovic
An independent researcher in Montenegro
11 PUBLICATIONS **17** CITATIONS

SEE PROFILE

UDC: 004.8 *(IJACSA)International Journal of Advanced Computer Science and Applications*,
DOI: 10.14569/IJACSA.2021.0120670 *Volume 12, Issue. 6, 2021*

Classification Model Evaluation Metrics

Željko Đ. Vujović
Boulevard Save Kovačevića 20/6, 81000 Podgorica, Montenegro
e-mail: etracon@t-com.me, cell. +382 67 852 298

Abstract - The *purpose* of this paper was to confirm the basic assumption that classification models are suitable for solving the problem of data set classifications. We selected four representative models: BaiesNet, NaiveBaies, MultilayerPerceptron, and J48, and applied them to a four-class classification of a specific set of hepatitis C virus data for Egyptian patients. We conducted the study using the WEKA software classification model, developed at Waikato University, New Zealand. Defeat results were obtained. None of the four classes envisaged has been determined reliably. We have described all 16 metrics, which are used to evaluate classification models, listed their characteristics, mutual differences, and the parameter that evaluates each of these metrics. We have presented comparative, tabular values that give each metric for each classification model in a concise form, detailed class accuracy with a table of best and worst metric values, confusion matrices for all four classification models, and a type I and II error table for all four classification models. In addition to the 16 metric classifications, which we described, we listed seven other metrics, which we did not use because we did not have the opportunity to show their application on the selected data set. Metrics were negatively rated selected, standard reliable, classification models. This led to the conclusion that the data in the selected data set should be pre-processed to be reliably classified by the classification model.

Keywords: classification model, classification models, evaluate classification models, worst metric values, four-class classification, metric classification, reliable classified classification models, detailed class accuracy

Subject areas: artificial intelligence and machine learning, software engineering

I INTRODUCTION

A specific set of data on the hepatitis C virus, consisting of 1385 instances described with 29 attributes, was considered. [12] The goal is to classify these instances into four classes, which represent hepatitis diseases: class a - Portal fibrosis, class, b - Little sepsis, class, c - A lot of sepsis, and class d - Cirrhosis.[6] This paper challenges this classification. Sources in the literature suggest that classification into five classes would be better: class a-liver inflammation, class b-fibrosis, class c-cirrhosis, class d – end-stage disease (ESLD), and class e-cancer. [15]

The initial assumption is that standard, generally accepted classification models, BayesNet, NaiveBayes, Multilayer-Perceptron, and J48, are suitable for such a classification. These models exist in the WEKA software and, as such, have been applied to the selected data set. Unsatisfactory results were obtained. Available instances are classified very poorly. That was the reason, motive, and incentive to consider why this is so? These four models were chosen at random. In this introduction, we give their generally accepted definitions.

A Bayesian network is defined as a system of event probabilities, nodes in a directed acyclic graph, in which, the probability of an event can be calculated from the probabilities of its predecessors in the graph. The nodes in the network are variable. They can be concrete values, randomly given, latent values , or hypotheses. They are characterized by the distribution of probabilities. Probability is a quantity that touches a presented state of knowledge or a state of belief. In Bayesian opinion, the probability is assigned to a hypothesis. In frequency thinking, the hypothesis is tested without assigning a probability. The result of Bayesian analysis is Bayesian inference. It updates the previous probability assigned to the hypothesis because more evidence and information have been obtained. [3], [16]

Naive Bayesian classifiers are based on naive assumptions of the mutual characteristics of independence. In this way, each distribution obtained can be independently estimated as a one-dimensional distribution. This alleviates the problems arising from the "curse of dimensionality". The “curse of dimensionality” is the problematic nature of the number of variables, which can be collected from a single sample. An example of this is the need for data sets that are scaled (arranged) exponentially with many characteristics.[3],[14] [16], [18].

A multilayer perceptron is defined as a system composed of a series of elements (nodes - "neurons") organized into layers. Layers process information so that they react dynamically to external inputs. The input layer has one neuron for each component, which exists in the input data. Communicates with hidden layers in the network. The entire processing of input data takes place in hidden layers. The input data are weighted (measured) by appropriate coefficients. The neuron accepts them, calculates their sum, and processes it with an activation function. It processes the processed data in a "forward" process. The last hidden layer is connected to the output layer. The output layer has one neuron for each possible output.[3], [14] ,[16], [18].

J48 is a machine learning model based on the decision tree. It was created using the ID3 algorithm (Iterative Dichtomizer 3), developed by the WEKA project development team. The decision tree presents and analyzes decision-making situations when one type of decision is derived from another type of decision. This facilitates understanding of selection problems, assessment of available versions of the decision, and coverage

of uncertain events, which affect outcomes and versions of the decision.[3],[14],[16],18].

The first idea was to consider the metrics used to evaluate the classification models used. 16 metrics used by WEKA software were reviewed, described, and explained. [4] In addition, it was stated that there are, in addition to the above, the following metrics: False discovery rate, [21] Log Loss, [22] Barier score, [23] Cumulative gain chart, [24] Lift curve, [25] Kolmogorov-Smirnov test, [26]. These metrics were not considered because they were not contained in the WEKA software, which was used. Therefore, they could not give their ratings of the classification model on the selected data set.

The research made a significant contribution to the interpretation of the 16 mentioned metrics, elements, and parameters that each of them uses to evaluate the classification models.

A significant contribution is also the question: why did the metrics negatively evaluate the classification models used on the selected data set?

As a result of this research, other questions arose. Is the number of attributes per instance of the observed data set too large? How many attributes are needed (optimal) and what are those attributes? Is it necessary to pre-process the data of the observed set? What are the techniques for pre-processing data in a set?

Unobtrusively, the question arose as to whether the four classes for the classification of instances of the observed set were correctly determined?

II METRICS

1. Accurately classified instances are the sum of true positive (TP) and true negative (TN).

2. Incorrectly classified instances are the sum of false positives (FPs) and false negatives (FNs).

3. Kappa statistic - Cohen's Kappa coefficient (k) is a measure of how many instances are classified model of machine learning, matched the data marked as the basic truth, controlling the accuracy of the random classifier as measured, expected accuracy. The accuracy of the Random Accuracy is 1 / k. Here k is the number of classes in the data set. In the case of binary classification k = 2, so the accuracy is 50%

$$K = \frac{(p0 - pe)}{(1 - pe)}$$

p0 - total accuracy of the module, pe - random accuracy (random accuracy of the model).

In the problem of binary classification pe = pe1 + pe2; pe1 - the probability that the predictions agree randomly with the actual values of class 1 - "good"; pe2 - the probability that the predictions agree randomly with the actual values of class 2 - "accidentally". The assumption is that the two classifiers (model prediction and actual class value) are independent. In this case, the probabilities pe1 and pe2 are calculated by multiplying the share of things in the class and the share of the predicted class.[2],[20].

4. Mean Absolute Error is the mean value of the absolute values of individual prediction errors of all instances in the test set. Each prediction error is the difference between the actual value and the predicted value for the instance.

The mean absolute error (MAE) *Ei* of an individual model and is calculated by the formula:

$$E_i = \frac{1}{n}\sum_{j=1}^{n}\left|P_{(ij)} - \sum_{j=1}^{n}\left|P_{(ij)} - T_j\right|\right|$$

where $P_{(ij)}$ is the value predicted by the individual model *i* for record *j* (of n records); and T_j is the target value for record *j*. For a perfect prediction, $P_{(ij)} = T_j$ and $E_i = 0$. Thus, the index E_i ranges from 0 to infinity, and 0 corresponds to the ideal. [14] [28]

5. Root mean squared error (RMSE) - The root mean square error is relative to what it would be if a simple predictor was used. Taking the square root of the relative square error, the error is reduced to the same dimensions as the predicted size.

The root mean square error (RMSE) E_i of an individual model and is calculated by the formula:

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(P_{(ij)} - T_j\right)^2}$$

Where $P_{(ij)}$ is the value predicted by the individual model *i* for record j (of n records), and T_j is the target value for the record *j*.For a perfect prediction, $P_{(ij)} = T_j$ and $E_i = 0$. Thus, the index E_i ranges from 0 to infinity, and 0 corresponds to the ideal.[27]

6. Relative absolute error (RAE) is the total absolute error and normalized by dividing by the total absolute error of the simple predictor (ZeroR classifier). The relative absolute error E_i of an individual model is evaluated by the equation:

$$E_i = \frac{\sum_{j=1}^{n}\left|P_{(ij)} - T_j\right|}{\sum_{j=1}^{n}\left|T_j - \bar{T}\right|}$$

Where $P_{(ij)}$ is the value predicted by the individual model *i* for record *j* (of *n* records); T_j is the target value for record *j*, and T is given by the formula:

$$\bar{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

For a perfect prediction, the counter is 0 and $E_i = 0$. Thus, the index E_i ranges from 0 to infinity, and 0 corresponds to the ideal.

A good prediction model produces a near-zero ratio. A bad model (one that is worse than a naive model) will produce a ratio greater than one x100%.[27]

7. Root relative squared error (RRSE) reduces the error to the same dimensions as the predicted size. Relative square error is the total square error divided by the total square error of a simple predictor. The root of the relative square error E_i of an individual model j is calculated by the formula:

$$E_i = \sqrt{\frac{\sum_{j=1}^{n} \left(P_{(ij)} - T_j\right)^2}{\sum_{j=1}^{n} \left(T_j - \bar{T}\right)^2}}$$

Where $P_{(ij)}$ is the value predicted by the individual model i for record j (of n records). For perfect prediction, the counter is equal to 0 and $Ei = 0$. The index Ei ranges from 0 to infinity, and 0 corresponds to the ideal. [28]

8. Confusion matrix for a binary classifier (Figure 1). Actual values are marked True (1) and False (0), and are predicted as Positive (1) and Negative (0). Estimates of the possibilities of classification models are derived from the expressions TP, TN, FP, FN, which exist in the confusion matrix. [10]

Class designation		Actual class	
		True (1)	False (0)
Predicted class	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 1 Confusion matrix for the binary classification problem [7]

TP (True Positive) - The data point in the confusion matrix is True Positive (TP) when a positive outcome is predicted and what happened is the same.

FP (False Positive) - The data point in the confusion matrix is false positive when a positive outcome is predicted, and what happened is a negative outcome. This scenario is known as a Type 1 Error. It is like a boon in bad foresight.

FN (False Negative) - The data point in the confusion matrix is false negative when a negative outcome is predicted, and what happened is a positive outcome. This scenario is well known as a Type 2 Error and is considered as dangerous as a Type 1 Error.

TN (True Negative) - The data point in the confusion matrix is True Negative (TN) when a negative outcome is predicted and what happens is the same. The results of the binary classification shown in Figure 2.

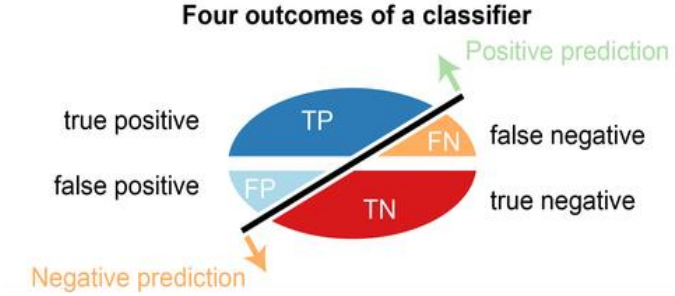


Figure 2. Elliptical representation of four binary results of the test set classification [7]

Confusion matrix for four-class classification (Figure 3). Four-class classification is a problem of classifying instances (examples) into four classes. Case of four classes: class A, class B, class C, and class D.[13],[17].

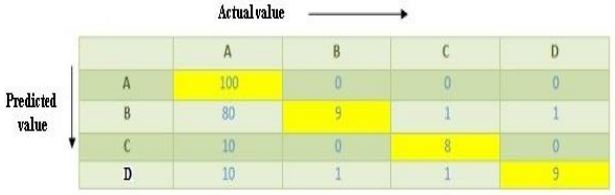
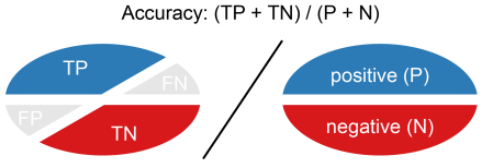


Figure 3 Confusion matrix for the four-class classification problem [8]

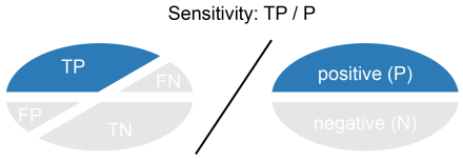
9. Accuracy is calculated as the sum of two accurate predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0, and the worst is 0.00. (Figure 4) [19]



$$ACC = \frac{TP + TN}{TP + TN + FN + FP} = \frac{TP + TN}{P + N}$$

Figure 4. Two ellipses show how accuracy is calculated [7],[11]

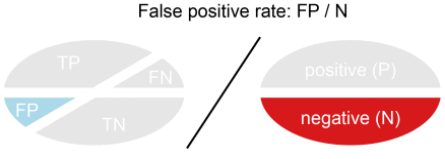
10. TP Rate - True Positive Rate (Sensitivity or Recall) is calculated as the number of accurate positive predictions (TP) divided by the total number of positive (P). Also called Sensitivity or Recall (REC). The best TP Rate is 1.0 and the worst 0.0. (Figure 5) [19]



$$SN = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Figure 5. Two ellipses show how the sensitivity is calculated [7]

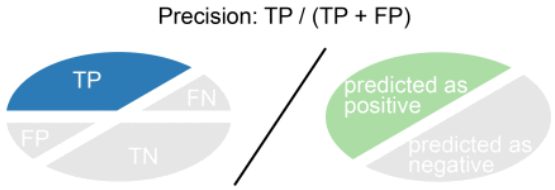
11. FP Rate - False Positive Rate is calculated as the number of false-positive predictions (FP) divided by the total number of negatives (N). The best false positive rate is 0.0 and the worst is 1.0. It can also be calculated as 1-specificity. (Figure 6) [19]



$$FPR = \frac{FP}{TN + FP} = 1 - SP$$

Figure 6. Two ellipses show how the False Positive Rate - FPR is calculated [7]

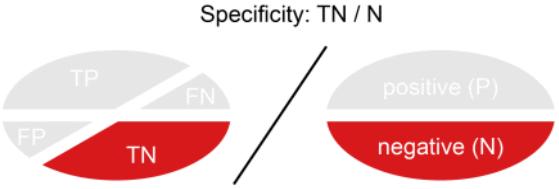
12. Precision is calculated as the number of correct positive predictions (TP), divided by the total number of positive predictions (TP + FP). The best accuracy is 1.0 and the worst 0.0. (Figure 7) [19]



$$PREC = \frac{TP}{TP + FP}$$

Figure 7. Two ellipses show how precision is calculated [7],[11]

13. True Negative Rate – TNR (Specificity) - is calculated as the number of correct negative predictions (TN) divided by the total number of negatives (N). The best specificity is 1.0 and the worst 0.0. . (Figure 8) [19]



$$SP = \frac{TN}{TN+FP} = \frac{TN}{N} \qquad SP = 1 - FPRate$$

Figure 8. Two ellipses show how specificity (SP) is calculated [7]

14. **F-Measure** or F-score is a measure of the accuracy of the test. It is calculated, based on precision and reminders, by the formula:

$$F\ Score = \frac{2 \times precision \times recall}{precision \times recall}$$ [7],[11],[19]

15. Matthews Correlation Coefficient (MCC) - is the correlation between the predicted classes and the basic truth. It is calculated based on the values from the confusion matrix.

$$MCC = \frac{TP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC is generally considered a balanced measure, which can be used even if the classes are of very different sizes. [7],[11],[19]

16. **ROC Area -** Receiver Operating Characteristic Area - The ROC curve is a graph that visualizes the trade-off between True Positive Rate and False Positive Rate. (Figure 9) For each threshold, we calculate True Positive Rate and False Positive Rate and plot them on one graph. The higher the True Positive Rate and the lower the False Positive Rate for each threshold, the better. Better classifiers have more curves on the left. The area below the ROC curve is called the ROC AUC score, a number that determines how good the ROC curve is. .[11]

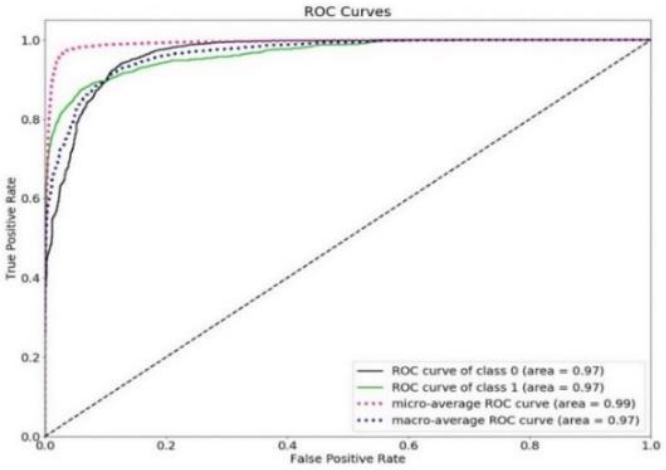


Figure 9 ROC curve [1],[5]

The ROC AUC Score shows how good the model is in ranking predictions. Indicates the probability that a randomly selected positive instance is ranked higher than a randomly negative instance. [7],[19]

17. **PRC Area** (Precision-Recall Curve Area) It is one number that describes the capabilities of the model. The PR AUC Score is the average of the precision scores calculated for each reminder threshold [0,0, 1,0]. The PRC curve is obtained by combining Positive Predictive Value and True Positive Rate. (Figure 10) For each threshold, Positive Predictive Value and True Positive Rate are calculated and the corresponding point of the graph is plotted. Preferably, the algorithm has high precision and high sensitivity. These two metrics are not independent. That is why a compromise is being made between them. A good PRC curve has a higher AUC. Research has shown that PRC is graphically more inforative than ROC graphs when estimating binary classifiers on unbalanced sets.[5],[9],[19].

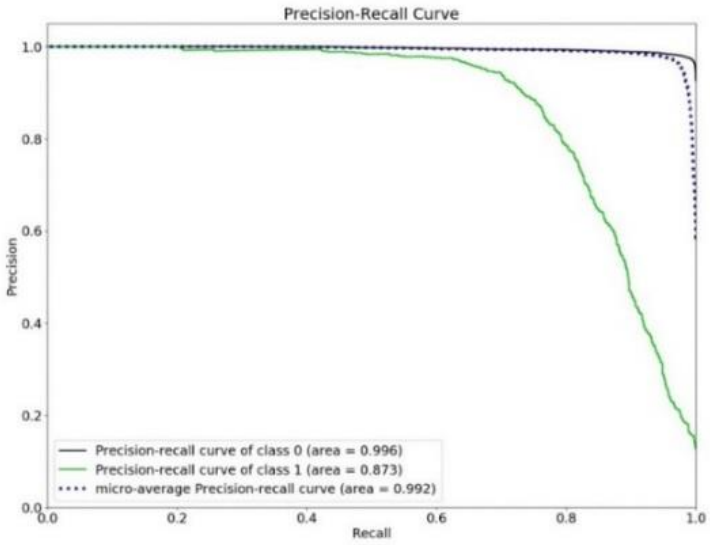


Figure 10 Precision-Recall curve [9]

III EXPERIMENTAL RESULTS

TABLE 1: Metrics Summary

	Bayes Net	Naive Bayes	Multilayer Perceptron	J48
Correctly classified instances	318	362	368	350
Incorrectly classified instances	1007	1023	1017	1035
Kappa statistic	-0,0287	0	0,0206	0,0029
Mean absolute error	0,3763	0,3748	0,3718	0,3751
Root mean squared error	0,4393	0,4329	0,5466	0,5814
Relative squared error	100,382%	99,9999%	99,2009%	100,0671%
Root relative squared error	101,4822%	100%	126,2575%	134,2938%
Total number of instancess	1385	1385	1385	1385

Summary of the accuracy of the four representative classifiers expressed by general metrics. Metrics are listed in the rows of the table, and their values, for each classifier, in the columns of the table. A special number is the total number of instances, which is the same for each classifier.

TABLE 2: Table of best and worst metric values for detailed class accuracy

	The Best	**The Worst**
TP Rate	1,0	0,0
FP Rate	0,0	1,0
Precision	1.0	0,0
Recall	1.0	0.0
F-Measure	1.0	0.0
MCC	+1.0	0.0
ROC Area	0.9	0.5
PRC Area	1.0	0.5

The best and worst values of each general metric are used to measure the accuracy of the classifier. Metrics are in rows and values are in columns of the table.

TABLE 3: Detailed Accuracy By Class

		TP Rate	**FP Rate**	**Precision**	**Recall**	**F.Measure**	**MCC**	**ROC Area**	**PRC Area**
c	BayesNet	0,107	0,186	0,156	0,107	0,127	0,091	0,423	0,205
l	NaiveBayes	0,000	0,000	?	0,000	?	?	0,496	0,241
a	M.L.Perc.	0,193	0,254	0,196	0,193	0,195	0,060	0,453	0,220
s	J48	0,250	0,236	0,253	0,250	0,251	0,014	0,501	0,247
s(1)	Weight Av.	0,230	0,250	0,222	0,230	0,224	0,032	0,473	0,243
c	BayesNet	0,271	0,270	0,241	0,271	0,255	0,001	0,510	0,249
l	NaiveBayes	0,000	0,000	?	0,000	?	?	0,496	0,238
a	M.L.Perc.	0,577	0,236	0,271	0,277	0,274	0,041	0,527	0,249
s	J48	0,271	0,226	0,274	0,271	0,273	0,045	0,526	0,252
s(2)	Weight Av.	0,261	0,261	?	0,261	?	?	0,496	0,249
c	BayesNet	0,214	0,266	0,217	0,214	0,216	-0,052	0,457	0,234
l	NaiveBayes	0,000	0,000	?	0,000	?	?	0,496	0,255
a	M.L.Perc.	0,282	0,247	0,282	0,282	0,282	0,035	0,521	0,278
s	J48	0,231	0,245	0,246	0,231	0,238	-0,0014	0,488	0,248
s(3)	Weight Av.	0,266	0,245	0,265	0,266	0,066	0,013	0,509	0,257
c	BayesNet	0,320	0,307	0,270	0,320	0,293	0,013	0,524	0,281
l	NaiveBayes	1,000	1,000	0,261	1,000	0,414	?	0,496	0,260
a	M.L.Perc.	0,307	0,243	0,308	0,307	0,307	0,063	0,533	0,280
s	J48	0,260	0,290	0,240	0,260	0,250	0,030	0,476	0,255
s(4)	Weight Av.	0,253	0,250	0,253	0,253	0,253	0,003	0,497	0,251

The detailed accuracy of each of the four representative classifiers for each of the predictions of the class is expressed by the values of eight different metrics. Metrics are in the columns of the table, the names of the classifiers in the rows of the table, separately for each class. For each class, the weighted value of each of the eight metrics is shown. This value is the average that results from multiplying each component by a factor that reflects its significance.

TABLE 4: Confusion Matrix

BayesNet				NaiveBayes				M.L.Perceptorn				J48				
a	**b**	**c**	**d**	**a**	**b**	**c**	**d**	**a**	**b**	**c**	**d**	**a**	**b**	**c**	**d**	
36	94	94	112	0	0	0	336	65	94	86	91	84	82	79	94	a= 336
61	90	85	96	0	0	0	332	79	92	86	75	67	90	80	95	b= 332
79	94	76	106	0	0	0	355	96	76	100	83	80	82	82	111	c = 335
55	96	95	111	0	0	0	362	91	78	82	111	101	74	93	94	d= 362

Comparative table of four confusion matrices for all four representative classifiers. In the rows, the number is provided for each class, and in the columns the actual value of the class.

TABLE 5: Type I Errors and Type II Errors

	BayesNet				NaiveBayes				M.L.Perceptron				J48			
	a	**b**	**c**	**d**	**a**	**b**	**c**	**d**	**a**	**b**	**c**	**d**	**a**	**b**	**c**	**d**
Error Type I	309	215	279	246	336	332	255	0	271	240	355	251	255	242	255	270
Error Type II	195	284	277	314	0	0	0	1.033	266	248	234	249	248	238	852	300

Comparative table of Type I and Type II error values for each class and each representative classifier. There are types of errors in the rows, and their size in the columns.

IV DISCUSSION

The average value of correctly classified instances is 25.24%, and incorrectly classified instances 73.68%. (Table 1)

Landis and Koch proposed the following standards for the kappa coefficient: ≤0 = poor, .01 - .20 = insignificant, .21 - .40 = fair, .41 - .60 = moderate, .61– .80 = substantial, a. 81–1 = almost perfect. [29] In line with the above proposal, BayesNet and NaiveBaies have a poor kappa coefficient, and multilayer perceptron and J48 are negligible. It is concluded that the values of the kappa coefficients show that the instances, classified by the machine learning model, do not match the data marked as the basic truth. MAE values: 0.3763 for BaiesNet, 0.3748 for NaiveBaies, 0.3718 for MultilayerPerceptron, 0.3751 for J48 are closer to the lower limit (ideal) than the upper (worst). We, therefore, appreciate that they are acceptable. (Table 1)

Anthony Ladson gave a model performance table based on the efficiency coefficient. For the case of model performance validation, the values of the efficiency coefficients describe the classification as follows: E≥0.93 - excellent, 0.8≤E<0.93 - good, 0.6≤E<0.8 - satisfactory, 0.3 ≤E<0.6 - transient, E<0.3 - bad. [30] Based on this, values of 0.4393 for BayesNet, 0.4329 for NaiveBayes, 0.5466 for MultilayerPerceptron, and 0.5814 for J48 are in the transient group. (Table 2)

Relative absolute error (RAE) can have values from 0 to infinity. Ideally, it should have a value of 0. Based on this, it is concluded that the values of 100.3802% for BaiesNet, 99.999% for NaiveBaies, 99, 2009% for MultilaierPerceptron, and 100.0671% for J48% are approximately the same as in the naive model (ZeroR classifier). The root of the relative square error (RRSE) can have a value from 0 to infinity. Ideally, it should have a value of 0. RRSE values: 101.4822% for BayesNet, 100% for NaiveBaies, 126.2775% for MultilaierPerceptron, and 134.2938% for J48 rate NaiveBayes as a naive model, and BayesNet, MultilayerPerceptron and J48 worse than naive. (Table 1)

Analysis of the detailed accuracy of the classes (Table 1 and Table 2) shows very significant results. Based on the tables of best and worst metric values for detailed class accuracy, we conclude:

1. TP Rate has extremely poor values, close to the worst, for all rated models and all classes. The exception is NaiveBaies, which has the best value of 1,000 for class 4, but the same NaiveBayes has the worst value of TP Rate, 0,000, for classes 1,2, and 3. Relatively good value of TP Rate, 0,577, showed MultilaierPerceptron for class 2. Weighted values TP Rates are consequently poor.
2. FP Rate for NaiveBayes has an optimal value of 0.000 for classes 1,2 and 3, as opposed to class 4 for which it has a maximum value of 1000. BayesNet, MultilayerPerceptron, and J48, as well as a weighted value for all four models, and all four classes are extremely bad.
3. Precision has values below a level satisfactory for all four models.
4. Recall, has the same values as TP Rate. The question is why are they separated for display in a separate column?
5. The F-Measure has values that are below levels that meet all rated models and all four classes.
6. MCC showed unsatisfactory values, which are at the level of random prediction, for all evaluated models and all classes.
7. The ROC Area showed values for all models and all classes that are on the verge of bad.
8. The value of the PRC area, for all models and all classes, is below the level that is the worst.

The metrics of detailed assessment by classes unequivocally show that the evaluated models, applied in a presented way, do not satisfy. (Table III) This means that new research is needed and the answer to the question: why do metrics of detailed accuracy give poor estimates of the models used?

By comparative analysis of the confusion matrix for all four classification models and all four classes, we see that the predictions of true positive results (TP) are not good enough. (Table 4) Type I and type II errors are relatively high. The goal of modeling is to reduce these errors to minimum values.
Separate consideration of type I and type II errors for the four applied models shows that NaiveBayes has a type I error value equal to 0, for class d, and type II errors for classes a, b, and c. (Table 5) These data further problematize the use of this model. For the other three models, the type I and type II errors are, on average, 2.5 times larger than exactly predicted.

V CONCLUSIONS

In this paper, we have considered in detail the 16 metrics for the evaluation of classification models, which exist in WEKA software, version 3.4.1., Developed at the University of Waikato, New Zealand. The consideration is in line with the initial assumption of the paper that classification models are suitable for solving the classification problem applied to a specific set of hepatitis C virus data for Egyptian patients.
In addition to the above 16 metrics, we found in the literature that there are other metrics: False discovery rate, Log Loss, Barrier score, Cumulative gain chart, Lift curve, Kolmogorov-Smirnov plot, and Kolmogorov - Smirnov statistics. We did not describe them because we were unable to demonstrate their application to the data set we selected. These metrics remain for display in a later review paper.
All metrics considered negatively evaluated the classification models, which we used. This has led to doubts because these are models that are generally accepted as standard and reliable. Why, metrics, do they rate them negatively on a selected data set? Is the number of attributes in the selected data set too large? How many attributes are needed and what are those attributes? Is it necessary to pre-process the data of the selected set?
The special significance of this paper is that it highlights the multitude of metrics used to evaluate each classification model. It emphasizes the diversity of these metrics and the parameters they measure to better understand the model and its features.
New questions and problems, which arose from this paper, are: What are the techniques for pre-processing data in a data set, and how should discretization, purification, reduction, and discussion of data be performed in a specific hepatitis C virus data set for Egyptian patients?
We suggest that the classification be performed in five classes, as provided in the latest professional literature: class a-inflammation of the liver, class b-fibrosis, class c-cirrhosis, class d – end-stage disease (ESLD), and class e-cancer.

VI ACKNOWLEDGMENT

To the editor and reviewers of IJACSA - The Science and Information (SAI) Organization. To Dejan Vujović, an engineer for the development and maintenance of application software in the Montenegrin Electriciity Transmission System Podgorica.

VII REFERENCES

[1] T.Fawcett, „ROC Graphs: Notes and Practical Considerations for Researchers.” *Kluwer Academic Publishers*, 2004.
[2] J.Sim, C.C.Wright, „The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements.” *Physical Therapy*, Volume 85, Issue 3, Pages 257 -68, 2005. https://doi.org/10.1093/ptj/85.3.257
[3] J.Đ.Novaković, „Rešavanje klasifikacionih problema mašinskog učenja.“ *Fakultet tehničkih nauka u Čačku, 2013.*
[4] R.R:Bouckaert, E. Frank, M. Hall, R. Kirkby, R.Reutmann, A. Sewald, A., D. Seuse, „*WEKA Manual for Version 3-7-8*.”, 2013.
[5] T. Saito, M. Rechmsmeier, „The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.”, PLoS ONE, 2015 *doi: 10.1371/journal.phone.0118432*
[6] M.Nasr, K. Elbahanacy, M. Hamdy, S.M.Kamal, „A novel model based on non-invasive methods for prediction of liver fibrosis.”*13th International Computer Engineering Conference (ICENCO), 2017*
[7] T. Saito, M. Rehmeismeier, „Basic evaluation measures from the confusion matrix.” *WordPress, 2017*
[8] V. Leal, „How to build a confusion matrix for a multiclass classifier?” *CrossValidated, StackExchange Inc, 2021*
[9] S. Auckland, S, „Precision-recall curves-what are they and how are they used.” *Acutecuretesting, 2017*
[10] S. Narkhede, „Understanding Confusion Matrix.” *Towards Data Science, 2018*
[11] A. Mishra, „Metrics to Evaluate your Machine Learning Algorithm.” *Towards Data Science, 2018*
[12] D. Dua and C. Graff, „UCIMachineLearning Repository [http://archive.ics.uci.edu/ml]. ”*Irvine, CA: The University of California, School of Information and Computer Science.Hepatitis C Virus (HCV) for Egyptian patients Data Set, 2019*
[13] A.Iqbal, A, Aftab2, S, Ali3, U, Nawaz4, Z, Sana5, L, Ahmad6, M, Husen7, A, „Performance Analysis of Machine Learning Techniques on Software Defect Prediction using NASA Datasets.” *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 5, 2019*
[14] R.Delgrado, X-A, Tibau „Why Cohen’s Kappa should be avoided as a performance measure in classification.” *PLoS One;14 (9), 2019*
[15] J.S.Saladi, „What Are The Stages of Liver Failure?.” *Healthline, 2019*
[16] Ž. Vujović, „The Big Data and Machine Learning.” *Journal of information technology and multimedia systems, Vol. 19, Issue 7. pp.11-19, DOI: 10.5281/zenodo.427923, 2020*
[17] S. Nandacumar, „Confusion Matrix – are you confused? (Part I and Part II).” *Medium, 2020*
[18] A. Albahr1, M. Albahar2, „An Empirical Comparison of Fake News Detection using different Machine Learning Algorithms.” *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 9, 2020*
[19] A. Tharwat, „Classification assessment methods.” *Applied Computing and Informatics, Volume 17, Issue 1, 30, 2020*
[20] M. Widmann, „COHEN’S KAPPA: What It Is, When to Use It, and How to Avoid Its Pitfalls.” *The New Stack, 2020*
[21] S. Room, „False Discovery Rate (FDR).” *In Dubitzky W., WolkenHauer O., Cho KH., Yokota H., (eds) Encyclopedia of Systems Biology, Springer New York, NY, doi:10.1007/978'1'4419'9863-7_223, 2013*
[22] G. Dembla, „Intuition behind Log-loss score.” *Towards Data Science, 2020*
[23] J. H. Orallo, P.A. Flach, C.Ferri, „Brier curves: a new cost-based visualization of classifier performance.” *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. pp. 585–592, 2011
[24] T. Jurczyk, „Gains vs ROC curves. Do you understand the difference? ” *TIBCO® Data Science, 2020*
[25] D.S. Coppock, „Data Modelling and Mining: *Why Lift?*” *DM Review and Source Media, Inc., 2006*
[26] A. Justel, D. Pena, R. Zamar, „ A multivariate Kolmogorov -Smirnov test of goodness of fit” Statistics&Probability Letters, Volume35, Issue3, Pages251-259, 1997, doi: 10.1016/S0167-7152(97)00020-5,

[27] S.Glen, „Mean Squared Error. Definition and Example."*From StatisticHowTo.com:Elementary Statistics for the rest of us! https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean/squared-error/, 2021*
[28] „Root Relative Squared Error."*GeneXproTools Online Guide, Gepsoft. Ltd., 2000-214*
[29] L.Hartling, M.Hamm, A.Milne, et al. „Interpretation of Fliess'kappa (k) (from Landis and Koch 1977)." *Valiability and Inter-Rater Reliability Testing of Quality Assessment Instruments [Internet], Rockvile: Agency for Healthcare Research and Quality (US), 2012*
[30] A.Ladson, „Model performance based on the coefficient of efficiency." *Hidrology, Natural Resources, and R, (2019)*