A case study of the application of WEKA software to solve the problem of liver inflammation

Željko Vujović (✉ etracon@t-com.me)

University of Maribor <https://orcid.org/0000-0001-8679-3344>

Research Article

Keywords: Confusion Matrix, Machine Classifier, Bayesian Classifier, Multilayer Perceptron, Support Vector Machine, Software Metrics, (Decision Tree, Neural Network)

Posted Date: April 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-464435/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

The aim of this paper was to consider the reliability of the basic metrics of evaluation of classification models: accuracy, sensitivity, specificity and precision. The WEKA software tool was applied to the "Hepatitis C virus (HCV) for Egyptian patients dataset". The algorithms Bayesnet, Naivebayesh, Multilayer Perceptron, J48 and 10-fold cross validation were used in the study. The main results obtained are that, with all four algorithms in question, they achieved approximately the same accuracy of correctly classified specimens. BaiesNet - 22.96%, NaiveBaies - 26.14%, MultilaierPerceptron - 26.57% and J48 - 25.27%.

Binary classification metrics - sensitivity, specificity and precision show very different values, depending on the intended class. Metric specificity, for all four algorithms, shows that a value that is in most of the range of possible values [0,1]. Metric sensitivity and precision, for all four algorithms, showed values that are in the lower part of the range of possible values [0,1].

The results of this study showed that WEKA software could not yet be considered as a relevant tool for the diagnosis of Hepatitis C Virus, on whose data set it was applied.

1. Background

The aim of this paper was to compare the properties of machine learning methods based on decision tree, machine support vector (SVM), neural networks and Bayesian networks, with a specific example of a hepatitis C virus dataset for Egyptian patients. For this comparison, the reliability measures of the used algorithms were used: accuracy, sensitivity, specificity and precision. The focus was on the classification model of WEKA software, developed at the University of Waikato, New Zealand, which, as one of the results, provides a detailed analysis of the accuracy of classifier predictions by class.

The reasoning behind this research was that it was not known what the accuracy, sensitivity, specificity and precision of machine learning methods were based on decision tree, machine support vector (SVM), neural networks and Bayesian networks, which were theoretically addressed in work „The Big Data and Machine Learning“, which preceded this work .

The classification problem in WEKA software version 3.8.4 is solved by a variety of algorithms. The classifier directory contains a total of 56 algorithms, arranged in 7 folders, as follows: folder bayes – 6 algorithms, functions – 11, lazy – 3, meta – 20, misk – 2, rules – 6 and trees – 8. For the decision tree used is the J48 algorithm. For machine support vector (SVM), there is an SMO algorithm in WEKA software. This algorithm was not used in the work because of the technical disadvantages of the machine on which the research was conducted. The Multilayer Perceptron algorithm was used for neural networks and Bayes Net and Naive Bayes for Bayesian networks.

The main results are that the accuracy of the four algorithms tested is approximately the same. It amounts to about 30%. This means that they did not prove to be good enough on the dataset to which they were applied. This most likely indicates that new data is needed. The possibility that a specific problem is not foreseeable has been ruled out for the time being. A possible reason for the weak traits shown by the algorithms used is that the data in the data set was not properly processed. The specificity metric is in the upper part of the range [0,1], which is its possible range of values. This means that the specificity is very good. In contrast, sensitivity and precision are in the lower range [0,1]. This means that these metrics are not good enough. All of the above showed that the machine learning methods listed are not good enough under the conditions in which they were used. They also indicated that a possible direction for their improvement was the improvement of the preliminary processing of the data set that was analyzed and on the basis of which the prediction was made, that is, the classification model was made. This enhancement of pre-processing includes scaling techniques, feature selection, data transformation, distribution transformation, data modeling. A dataset that has 1385 instances is a small set. To solve the problem of predicting diseases caused by the hepatitis C virus, a larger data set is needed than the one discussed in this paper.

2. Material

Table 1

Hepatitis C Virus (HCV) for Egyptian patients Data Set

Data Set Characteristics:	Multivariate	Number of Instances:	1385	Area:	Life
Attribute Characteristics:	Integer, Real	Number of Attributes:	29	Date Donated	2019-09-30
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	32719

Abstract

Egyptian patients who underwent treatment dosages for HCV about 18 months. Discretization should be applied based on expert recommendations; there is an attached file shows how.

Source: Professor: Sanaa Kamal, (Professor of Medicine, Ain Shams University - Faculty of Me

dicine-Egypt), Prof. Dr. Khalid Abdelhameed ElBahnasy, (Professor of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University-Egypt), Dr. Mohamed Hamdy ElEleimy, (Associate Professor at Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University-Egypt), Dr. Doaa Hegazy, (Assistant Professor at Information Systems Department, Faculty of Computer and Information Sciences, Ain Shams University-Egypt), Mr. Mahmoud Nasr, (MSc. Faculty of computer and information sciences - Ain Shams University-Egypt). For any inquiries please contact me at: m.nasr @ cis.asu.edu.eg

3. Methods

Confusion matrix for binary and four-class classification, TP Rate, FP Rate, Precision, Recall, F-Measure, Matthews Correlation Coefficient (MCC), Receiver Operating Characteristic Curve - ROC Area, Precision-Recall Curve Area - PRC Area.

Confusion matrix for binary classifier. Actual class values are labeled True (1) and False (0), and Predicted as Positive (1) and Negative (0)

Performance estimates of classification models are derived from the terms TP, TN, FP, FN, existing in the confusion matrix.

TP (True Positive) – Data point in confusion matrix is true positive when there is predicted a positive outcome and what happened is the same.

FP (False Positive) – Data point in confusion matrix is false positive when there is predicted positive outcome and what happened is a negative outcome. This scenario is known as Type 1 error. It is like a boon in a bad prediction.

FN (False Negative) – Data point in confusion matrix is false negative when there is predicted negative outcome and what happened is positive outcome. This scenario is known as Type2 error and it is considered as much dangerous as Type 1 error.

TN (True Negative) – Data point in confusion matrix is true negative when there is predicted a negative outcome and what happened is the same.

Confusion matrix for four-class classification. 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.

Accuracy

Accuracy is calculated as the total of two correct predictions (TP + TN) divided by the total number of data sets (P + N). The best accuracy is 1.0 and the worst is 0.0.

Sensitivity (Recall or True Positive Rate – TPR)

Sensitivity is calculated as the number of correct positive predictions (TP) divided by the total number of positive (P). Also called Recall (REC) or True Positive Rate. The best sensitivity is 1.0 and the worst is 0.0.

Specificity (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 is 0.0.

False Positive Rate – FPR

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.

Precision

Precision is calculated as the number of correct positive predictions (TP) divided by the total number of positive predictions (TP + FP). The best precision is 1.0 and the worst is 0.0.

Recall

(See Figure 9)

F-Measure

The **F-score** or **F-measure** is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Precision is also known as positive predictive value, and recall is also known as sensitivity in diagnostic binary classification.

$$F1\text{-Score} = 2 \frac{\text{precision} * \text{recall}}{(\text{precision} + \text{recall})}$$

MCC

It's a correlation between predicted classes and ground truth. It can be calculated based on values from the confusion matrix:

$$MCC = \frac{tp * tn - fp * fn}{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}$$

Alternatively, you could also calculate the correlation between `y_true` and `y_pred`.

We can adjust the threshold to optimize MCC. When to use it, When working on imbalanced problems, When you want to have something easily interpretable.

ROC Area

It is a chart that visualizes the tradeoff between true positive rate (TPR) and false positive rate (FPR). Basically, for every threshold, we calculate TPR and FPR and plot it on one chart. Of course, the higher TPR and the lower FPR is for each threshold the better and so classifiers that have curves that are more top-left side are better.

We can see a healthy ROC curve, pushed towards the top-left side both for positive and negative class. It is not clear which one performs better across the board as with $FPR < \sim 0.15$ positive class is higher and starting from $FPR \sim 0.15$ the negative class is above.

ROC AUC score

In order to get one number that tells us how good our curve is, we can calculate the Area Under the ROC Curve, or ROC AUC score. The more top-left your curve is the higher the area and hence higher ROC AUC score.

Alternatively, it can be shown that ROC AUC score is equivalent to calculating the rank correlation between predictions and targets. From an interpretation standpoint, it is more useful because it tells us that this metric shows **how good at ranking predictions your model is**. It tells you what is the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance.

PRC Area

The PRC Area is the area below the Precision-Recall curve. The PRC curve was obtained by combining precision (PPV) and sensitivity (TPR). For each threshold, the PPV and TPR are calculated and the corresponding graph point is plotted.

PR AUC score | Average precision

The Area Under the Precision-Recall Curve is one number that describes model performance.

PR AUC score is the average of precision scores calculated for each recall threshold [0.0, 1.0].

4. Results

=== Run information ===

Scheme: weka.classifiers.bayes.**BayesNet** -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E

weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Relation: HCV-Egy-Data_modified

Instances: 1385

Attributes: 29

Age, Gender, BMI, Fever, Nausea,Vomting, Headache,Diarrhea, Fatigue & generalized bone ache, Jaundice,Epigastric pain,WBC,RBC,HGB,Plat,AST 1,ALT 1,ALT4,ALT 12,ALT 24 ALT 36,ALT 48,ALT after 24,RNA Base,RNA 4,RNA 12,RNA EOT,RNA EF,Baseline histologi-cal Grading, Baselinehistological staging

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Bayes Network Classifier

not using ADTree

#attributes=29 #classindex=28

Network structure (nodes followed by parents)

Age (1): Baselinehistological staging

Gender(1): Baselinehistological staging

BMI(1): Baselinehistological staging

Fever(1): Baselinehistological staging

Nausea/Vomting(1): Baselinehistological staging

Headache (1): Baselinehistological staging

Diarrhea (1): Baselinehistological staging

Fatigue & generalized bone ache (1): Baselinehistological staging

Jaundice (1): Baselinehistological staging

Epigastric pain (1): Baselinehistological staging

WBC(1): Baselinehistological staging

RBC(1): Baselinehistological staging

HGB(1): Baselinehistological staging

Plat(1): Baselinehistological staging

== Stratified cross-validation ==

=== Summary ===

Correctly Classified Instances 318 22.9603 %

Incorrectly Classified Instances 1067 77.0397 %

Kappa statistic -0.0287

Mean absolute error 0.3763

Root mean squared error 0.4393

Relative absolute error 100.3802 %

Root relative squared error 101.4822 %

Total Number of Instances 1385

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
---------	---------	-----------	--------	-----------	-----	----------	----------	-------

0.107	0.186	0.156	0.107	0.127	-0.091	0.423	0.205	1
-------	-------	-------	-------	-------	--------	-------	-------	---

0.271	0.270	0.241	0.271	0.255	0.001	0.510	0.249	2
-------	-------	-------	-------	-------	-------	-------	-------	---

0.214	0.266	0.217	0.214	0.216	-0.052	0.457	0.234	3
-------	-------	-------	-------	-------	--------	-------	-------	---

0.320	0.307	0.270	0.320	0.293	0.013	0.524	0.281	4
-------	-------	-------	-------	-------	-------	-------	-------	---

Weighted Avg.

0.230	0.258	0.222	0.230	0.224	-0.032	0.479	0.243	
-------	-------	-------	-------	-------	--------	-------	-------	--

=== Confusion Matrix ===

a b c d <-- classified as

36 94 94 112 | a = 1

61 90 85 96 | b = 2

79 94 76 106 | c = 3

55 96 95 116 | d = 4

=== Run information ===

Scheme: weka.classifiers.bayes.**NaiveBayes**

Relation: HCV-Egy-Data_modified

Instances: 1385

Attributes: 29

Age, Gender, BMI, Fever, Nausea,Vomting, Headache,Diarrhea, Fatigue & generalized bone ache, Jaundice,Epigastric pain,WBC,RBC,HGB,Plat,AST 1,ALT 1,ALT4,ALT 12,ALT 24 ALT 36,ALT 48,ALT after 24,RNA Base,RNA 4,RNA 12,RNA EOT,RNA EF,Baseline histologi-cal Grading, Baselinehistological staging

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

Attribute	Class			
	1	2	3	4
	0.24)	(0.24)	(0.26)	(0.26)

== Stratified cross-validation ==

=== Summary ===

Correctly Classified Instances 362 26.1372 %
Incorrectly Classified Instances 1023 73.8628 %

Kappa statistic 0
Mean absolute error 0.3748
Root mean squared error 0.4329
Relative absolute error 99.9999 %
Root relative squared error 100 %
Total Number of Instances 1385

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.000	0.000	?	0.000	?	0.496	0.241	1	
0.000	0.000	?	0.000	?	0.496	0.238	2	
0.000	0.000	?	0.000	?	0.496	0.255	3	
1.000	1.000	0.261	1.000	0.414	?	0.496	0.260	4
Weighted Avg.								
0.261	0.261	?	0.261	?	0.496	0.249		

=== Confusion Matrix ===

a b c d <-- classified as
0 0 0 336 | a = 1
0 0 0 332 | b = 2
0 0 0 355 | c = 3
0 0 0 362 | d = 4

=== Run information ===

Scheme: weka.classifiers.functions.**MultilayerPerceptron** -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Relation: HCV-Egy-Data_modified

Instances: 1385

Attributes: 29

Age, Gender, BMI, Fever, Nausea,Vomting, Headache,Diarrhea, Fatigue & generalized bone ache, Jaundice,Epigastric pain,WBC,RBC,HGB,Plat,AST 1,ALT 1,ALT4,ALT 12,ALT 24 ALT 36,ALT 48,ALT after 24,RNA Base,RNA 4,RNA 12,RNA EOT,RNA EF,Baseline histologi-cal Grading, Baselinehistological staging

Test mode: 10-fold cross-validation

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 368 26.5704 %
Incorrectly Classified Instances 1017 73.4296 %

Kappa statistic 0.0206
Mean absolute error 0.3718
Root mean squared error 0.5466
Relative absolute error 99.2009 %
Root relative squared error 126.2575 %
Total Number of Instances 1385

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.193	0.254	0.196	0.193	0.195	-0.060	0.453	0.220	1
0.277	0.236	0.271	0.277	0.274	0.041	0.527	0.249	2
0.282	0.247	0.282	0.282	0.282	0.035	0.521	0.278	3
0.307	0.243	0.308	0.307	0.307	0.063	0.533	0.280	4
Weighted Avg.								
0.266	0.245	0.265	0.266	0.266	0.021	0.509	0.257	

=== Confusion Matrix ===

a b c d <-- classified as
65 94 86 91 | a = 1
79 92 86 75 | b = 2
96 76 100 83 | c = 3

```

91 78 82 111 | d = 4
=== Run information ===
Scheme:   weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: HCV-Egy-Data_modified
Instances: 1385
Attributes: 29
Age, Gender, BMI, Fever, Nausea,Vomting, Headache,Diarrhea, Fatigue & generalized bone ache, Jaundice,Epigastric pain,WBC,RBC,HGB,Plat,AST 1,ALT
1,ALT4,ALT 12,ALT 24 ALT 36,ALT 48,ALT after 24,RNA Base,RNA 4,RNA 12,RNA EOT,RNA EF,Baseline histologi-cal Grading, Baselinehistological staging
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      350          25.2708 %
Incorrectly Classified Instances    1035          74.7292 %
Kappa statistic                    0.0029
Mean absolute error                 0.3751
Root mean squared error             0.5814
Relative absolute error             100.0671 %
Root relative squared error         134.2938 %
Total Number of Instances          1385
=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall   F-Measure  MCC    ROC Area  PRC Area  Class
0.250    0.236    0.253     0.250   0.251      0.014   0.501     0.249     1
0.271    0.226    0.274     0.271   0.273      0.045   0.526     0.252     2
0.231    0.245    0.246     0.231   0.238     -0.014   0.488     0.248     3
0.260    0.290    0.240     0.260   0.250     -0.030   0.476     0.255     4
Weighted Avg.
0.253    0.250    0.253     0.253   0.253      0.003   0.497     0.251
=== Confusion Matrix ===
 a  b  c  d  <-- classified as
84  82  79  91 |  a = 1
67  90  80  95 |  b = 2
80  82  82  111 |  c = 3
101  74  93  94 |  d = 4

```

Table 2. Comparison of sensitivity, specificity and precision of the observed algorithms

R1 – COMPARISON TABLES

	Osjetljivost = TP Rate				Specifičnost = 1 - FP Rate				Preciznost			
Klasa	B.N.	N.B.	M.P.	J48	B.N.	N.B.	M.P.	J48	B.N.	N.B.	M.P.	J48
Portalna fibroza (F1)	0,107	0,000	0,193	0,250	0,814	1,000	0,746	0,764	0,156	?	0,193	0,253
Malo sepse (F2)	0,271	0,000	0,277	0,271	0,790	1,000	0,746	0,774	0,241	?	0,277	0,274
Mnogo sepse (F3)	0,214	0,000	0,282	0,231	0,734	1,000	0,753	0,754	0,217	?	0,282	0,246
Ciroza (F4)	0,329	1,000	0,307	0,250	0,693	0,000	0,757	0,760	0,270	0,261	0,307	0,240

Accuracy = [BayesNet=22,96%, NaiveBayes=26,28%, MultilayerPerceptron=26,57%, J48=25,27%]

Table 3. Analysisi of the four algorithms used which is obtained on the output for testing the Experimenter

```

Tester:   weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-matrix "weka.experiment.ResultMatrixPlainText -mean-prec 2
          -stddev-prec 2 -col-name-width 0 -row-name-width 25 -mean-width 0 -stddev-width 0 -sig-width 0 -count-width 5 -print-col-names -print-row-names -enum-col-names"

Analysing: Percent_correct
Datasets:  1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:      4/4/20, 9:16 AM

Dataset              (1) bayes.Ba | (2) bayes (3) funct (4) trees
-----
HCV-Egy-Data_modified (100) 26.14 | 24.48 25.00 24.95
-----
                    (v/ /*) | (0/1/0) (0/1/0) (0/1/0)

Key:
(1) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954
(2) bayes.NaiveBayes '' 5995231201785697655
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(4) trees.J48 '-C 0.25 -M 2' -217733168393644444

```

Table 4. Analysis of the four algorithms used which is obtained on the output for testing the Experimenter

R2 - SUPPORT VECTOR MACHINE - Report on an attempt to execute the SMO algorithm

Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heapsize.Initial heap size 32MB,

- Initial heap size 32MB,
- Current memory(heap) used: 461,4MB,
- max memory (heap) available: 510 MB.

R3 - RANK AND ACCURACY OF ALGORITHMS

The experiment showed that all four algorithms used were of the same quality and that one of them could not be determined to be better than the rest. The BayesNet algorithm showed an accuracy of 22.96%, NaiveBayes 26.28%, MultilayerPerceptron 26.57% and J48 25.27%.

5. Discussion

A set of metrics is used to evaluate the classification model. The basic metrics, derived from the confusion matrix are: TP; FP; FN i TN. In addition following are used: ACC – Accuracy, ERP – Error Rate (1-ACC), TPR – True Positive Rate, FPR – False Positive Rate, PREC – Precision (PPV – Positive Predictive Value), REC – Recall (TPR, Sensitivity), TNR – True Negative Rate (SP, Specificity), F_β score, MCC – Matthews Correlation Coefficient, ROC – Receiver Operating Characteristic Curve, PRC – Precision-Recall Curve i druge.

In the results of this study, for all evaluated models, the following metrics were presented: TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area, and confusion matrix for four classes.

TP Rate - True Positive Rate is the same as the Metall metric (Sensitivity) Shows the sensitivity of the model to positive predictions Shows the percentage of positive predictions, the probability that the actual positive value will be positive In medical diagnostics, for example, sensitivity The test is the ability of the test to correctly identify those who have the disease.It is a true positive rate.It shows how many positive labels the model has identified, of all possible labels.

FP Rate - False Positive Rate, a false positive rate (percentage, probability) is a measure of the accuracy of a test, whether it is a medical diagnostic test or something else. In technical terms, a false-positive rate is defined as the probability of falsely rejecting the null hypothesis.

Precision is the same as the PPV - Positive Predictive Value metric identifies the frequency at which the model was accurate in predicting positive class. This is the share of relevant copies among the downloaded copies. unlike the Recall metric (sensitivity, reminder), which is the proportion of relevant copies that are downloaded.

Recall is the proportion of relevant cases found by a search, divided by the total number of existing relevant cases. Relevance indicates how well the downloaded document meets the user's need for information. Relevance may include concerns such as timeliness, authority, or novelty of results.

The F-Measure or F-Score provides a combination of precision and sensitivity in one measure that captures both features, giving each the same weight. It is the harmonious middle of the two fractions, precision and sensitivity. The result is a value between 0.0 for the worst F-measure and 1.0 for the perfect F-measure. A harmonic mean is defined as the reciprocal of the arithmetic mean of the reciprocal of a set of elements. In our case we have two elements, precision (P) and sensitivity (R). Based on that, it was obtained that $F\text{-Measure} = 2 \cdot (P \cdot R) / (P + R)$. The F-Measure is interesting in some cases when more attention is paid to precision. For example, when false positives are more important to minimize, and false negatives are still important. In other cases, it is interesting when more attention is paid to sensitivity. For example, when false negative results are more important to minimize, and false positive results are still important.

The MCC - Matthews Correlation Coefficient takes into account the equilibrium ratio of the four categories of the confusion matrix (TP, FP, FN, TN). It is considered a balanced measure that can and should be used even when the classes are unbalanced. It is the basic correlation coefficient between the observed and the predicted binary classification. Name value from - 1 to + 1. A value of + 1 represents perfect prediction, 0 - random prediction, and - 1 indicates a complete mismatch between prediction and observation. MCC is considered to be one of the best metrics for describing the confusion matrix of true and false positives and negatives by a single number. It does not depend on which class is positive.

ROC Area is the area under the ROC curve (AUC). Summarizes the performance of each classifier in one measure and serves to compare classifiers. It is equivalent to the probability that a randomly selected positive instance is ranked higher than a randomly selected negative instance. AUC provides a unified measure of performance for all possible classification thresholds. AUC values range from 0 to 1. A model whose predictions are 100% incorrect has an AUC of 0.0, and one whose predictions are 100% correct has an AUC of 1.0. TPR and FPR are calculated for each threshold and plotted in a single graph. The higher the TPR and FPR for each threshold, the better. Based on this, it is concluded that better classifiers have more curves on the left. ROC AUC score is a number that corresponds to the area under the ROC curve. This indicator shows how good the model is in ranking predictions. It says what is needed: what is the probability that a randomly selected positive instance is ranked higher than a randomly selected negative instance. The ROC Area is higher, and thus the ROC AUC score, when the upper left curve is larger.

ROC Curve (Receiver Operating Characteristic Curve)

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.

ROC Curve is a graphical plot used to show the diagnostic ability of binary classifiers.

A ROC Curve is constructed by plotting the true positive rate (TPR) against false positive rate (FPR). For example, in medical testing the true positive rate in which people are correctly identified to test positive for the disease in question. A discrete classifier that returns only the positive class gives a single point on the ROC space. But for probabilistic classifiers, which give a probability or score that reflects the degree to which give a probability or score that reflects the degree to which an instance belongs to one class rather than another, we can create a curve by varying the threshold for the score. Note that many discrete classifiers can be converted to a scoring classifier by 'looking inside' their instance statistics. For example, a decision tree determines the class of a leaf node from the proportion of instances at the node. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity ($1 - \text{FPR}$). Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal ($\text{FPR} = \text{TPR}$). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. Note that the ROC does not depend on the class distribution. This makes it useful for evaluating classifiers predicting rare events such as diseases or disasters. In contrast, evaluating performance using accuracy $(\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FN} + \text{FP})$ would favor classifiers that always predict a negative outcome for rare events.

Area under curve (AUC)

To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve, which is abbreviated to AUC. It is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. A classifier with high AUC can occasionally score worse in a specific region than another classifier with lower AUC. But in practice, the AUC performs well as a general measure of predictive accuracy.

AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.

The PRC area is the area below the Precision-Recall curve. The PRC curve was obtained by combining precision (PPV) and sensitivity (TPR). For each threshold, the PPV and TPR and the corresponding gaffe point are calculated. Higher sensitivity means less precision. The sensitivity value, at which precision begins to decline rapidly, is used to select a threshold and a good model. By calculating the area under the precision-sensitivity curve, a number is obtained that describes the performance of the model. PR AUC is the average accuracy of the results for each sensitivity threshold [0.0; 1.0]. It is desirable that the algorithm has high precision and high sensitivity. These two metrics are not independent. That is why a compromise is made between them. A good PR curve has a higher AUC. Research has shown that PR is graphically more informative than ROC graphs when estimating binary classifiers on unbalanced sets.

Precision-Recall Curve

It is a curve that combines precision (PPV) and Recall (TPR) in a single visualization. For every threshold, you calculate PPV and TPR and plot it. The higher on y-axis your curve is the better your model performance.

You can use this plot to make an educated decision when it comes to the classic precision/recall dilemma. Obviously, the higher the recall the lower the precision. Knowing **at which recall your precision starts to fall fast** can help you choose the threshold and deliver a better model.

We can see that for the negative class we maintain high precision and high recall almost throughout the entire range of thresholds. For the positive class precision is starting to fall as soon as we are recalling 0.2 of true positives and by the time we hit 0.8, it decreases to around 0.7.

PR Curve is desired that the algorithm should have both high precision, and high recall. However, most machine learning algorithms often involve a trade-off between the two. A good PR curve has greater AUC (area under curve). In the figure above, the classifier corresponding to the blue line has better performance than the classifier corresponding to the green line. It is important to note that the classifier that has a higher AUC on the ROC curve will always have a higher AUC on the PR curve as well. Consider an algorithm that classifies whether or not a document belongs to the category "Sports" news. Assume there are 12 documents, with the following ground truth (actual) and classifier output class labels. By setting different thresholds, we get multiple such precision, recall pairs. By plotting multiple such P-R pairs with either value ranging from 0 to 1, we get a PR curve.

5.1. CONSIDERATION OF EXPERIMENTAL RESULTS

The BayesNet, NaiveBayes, MultilayerPerceptron and J48 algorithms were used in the paper, four of the 56 algorithms embedded in WEKA software, because they were theoretically processed in the work that preceded this one. Algorithms solved the problem of four-class classification. Each instance of a classified data set could be assigned to one of four classes. 10-fold layered cross-validation was used to evaluate the model. This type of validation was chosen because it is widely known that it evaluates objectively the skill of a model, with little bias and little variance. Number 10 indicates the number of groups to which the data sample is divided. Each group has the same percentage of observations with a given categorical value. The general procedure of 10-fold cross-validation is performed as follows:

1. The data set is shuffled randomly without hesitation.
2. The mixed tax set is divided into 10 groups.
3. For each group, individually, do the following:
 - One group is taken as a hold out set (test dataset). This set provides a final assessment of the model's properties for machine learning, after training and model validation.
 - The remaining groups are taken as a training dataset.
 - The model is fitted with a training dataset and evaluated with a test dataset.
4. The resulting model rating is retained and the model rejection. Modeling skills are summarized using a sample of model evaluation results.

Each observation in the data sample is assigned to an individual group and remains in that group for the duration of the procedure. Each sample is given the opportunity to be used as the hold out set 1 time, and to train the model 9 times. An overview of the statistics for each algorithm compared shows how accurately the classifier could have predicted an instance class in the selected test mode. The values of the Kappa coefficients show that the observed algorithms are at the boundary between unacceptable and slightly acceptable quality. Mean absolute error and Root mean squared error values may be considered satisfactory. High values of Relative absolute error and Root relative squared error indicate that the observed algorithms predict well. Detaljnu analizu tačnosti predviđanja po klasama, izražena je metrikama TP Rate, FP Rate, Precision, Recall, F-Measure, ROC Area i PRC Area. These metrics provide more information about the properties of the algorithms than the accuracy itself. Based on their values and definitions of sensitivity, specificity and precision metrics, a table comparing the reliability parameters of the algorithms was made. It shows that the sensitivity, specificity and precision of the algorithms are very different.

The values compared show the following hierarchy of algorithms accuracy: MultilayerPerceptron > Naive Bayes > J48 > Bayes Net. Interestingly, the accuracy of correctly classified instances obtained by 10-fold cross-validation differs from the accuracy obtained by experiment using the Experimenter option (TABLE 2). The reason for this difference could be the subject of special research.

The sensitivity hierarchy of algorithms for individual classes is, F1: J48 > MultilayerPerceptron > Bayes-Network > NaiveBayes, F2: MultilayerPerceptron > J48 = BayesNetwork > NaiveBayesand, F3: Multilayer-Perceptron > J48 > BN > NB, and for class F4: NaiveBayes > BayesNetwork > MultulayerPerceptron > J48. The Naive Bayes algorithm has poor sensitivity for all classes except class F4-Cirrhosis, for which it has the best sensitivity. It can be seen that the sensitivity values are in the lower part of the range [0,1], in which 0 is the worst sensitivity and 1 is the best. This means that the sensitivities could be better.

The hierarchy of specificity for individual classes is, F1: BayesNetwork > J48 > MultilayerPerceptron > NaiveBayes, F2: BayesNetwork > J48 > MultilayerPerceptron > NaiveBayes, F3: J48 > Multilayer Perceptron > BayesNetwork > NaiveBayes, F4: J48 > MultilayerPerceptron > BayesNetwork > NaiveBayes. The table shows that specificity has values closer to the upper limit of the range [0,1], which contains values for specificity. This means that the specificity is very good.

The precision hierarchy for individual classes is, F: J48 > MultilayerPerceptron > Bayes-Network? NaiveBayes, F1: MultilayerPerceptron > J48 > BayesNetwork? NaiveBayes, F2: Multilayer-Perceptron > J48 > BayesNetwork ? NaiveBayes, F4: MultilayerPerceptron > BayesNetwork > NaiveBayes > J48. The table shows that the precision values are closer to the lower limit of the range [0,1], which contains the precision values. He finds that precision could be better.

The Confusion Matrix, at the output of the classifier, shows how many instances are assigned to each class. The elemental matrix shows the color of an example test whose real class is a row and the predicted class is a column.

6. Conclusion

This comparative analysis of the reliability metrics of machine learning algorithms, accuracy, sensitivity, specificity, and reliability showed:

- The observed algorithms have poor properties for classifying the data set in question. This can be seen from the accuracy values of each of these algorithms. The number of correctly classified examples is less than the number of incorrectly classified examples.
- Sensitivity, specificity and precision have, in comparison, very different values, which depend on the class being predicted. Specificity is very good and sensitivity and precision are satisfying. All this is not enough to conclude what kind of errors a classifier is making. Hepatitis C Virus (HCV) For Egyptian Patients Data Set should be increased, so that it can be used as a reliable basis for modeling and predicting diseases caused by the hepatitis C virus
- A new, larger data set needs to be pre-processed, which includes scaling techniques, feature selection, data transformation, distribution transformation, data modeling.

In particular, other metrics for evaluating classification models need to be studied, primarily F-Measure, MCC, ROC Area, PRC Area.

On the other hand, two important questions arise. First, are the results obtained in this study relevant for solving the problem of liver inflammation and second, which modern review paper on the problem of liver inflammation should be taken as basis for new directions in the study of liver inflammation?

Declarations

- **ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

Not applicable

- **CONSENT FOR PUBLICATION**

Not applicable

- **COMPETING INTERESTS**

The author declare that he have no, financial competing interests. The author declare that he have no known non-financial competitive interests .

- **FUNDING**

Not applicabele

- **AUTHOR'S CONTRIBUTION**

The conception, design, acquisition, analysis and interpretation of data are, on the whole, the contribution of the author. This is, on the whole, individual research work.

The author agrees that issues related to the accuracy or integrity of any part, even those in which the author is not personally involved, should be investigated, resolved and the resolution documented in the literature.

- **ACKNOWLEDGEMENTS**

Not applicable.

- **AVAILABILITY OF DATA AND MATERIAL**

All data generated or analysed during this study are included in this published article [and its supplementary files]

References

- [1] Nir Nissim, Yuval Shahar, Yuval Elovici, George Hripcsak, Robert Moscovitch, „Inter-labeller and intra-labeller variability of condition severity classification models using active and passive learning methods“, Artificial Intelligence in Medicine, Volume 81, Spetemeber 2017, pages 12-32.
- [2] Wong-Ren, Bi-Qian Sun, Hua Tang, Jian Huang, Hao Lin, „Identify and analysys crotonylation sites in histone by using support vector machine“, Machine Learning and Graph Analysis in Computational Biomedicine, November 2017.
- [3] Leyi Wei, Shixiang Guo, Kalvin KL Wong, „A novel hierarchical selective ensemble classifier with bioinformatics applications“, Machine Learning and Graph Analysis in Computational Biomedicine, November 2017.
- [4] Ž.Vujović, „The Big Data and Machine Learning“, Journal of information technology and multimedia systems, Vol. 19, Issue 71(2020), pp. 11-19, DOI: 10.5281/zenodo.4279238
- [5] Machine Learnong Glossary <https://developers.google.com/machine-learning/glossary#c> (Accessed, 14.04.2021.)
- [5] <https://images.app.goo.gl/TgHc5e4PXVRFmiUh6> (Prist. 29.02.2020.)
- [6] <https://code.likeagirl.io/beginning-to-weka-step-by-step-93f6564d9f2>, Nov,23.2018.(Prist. 28.02.2020.)
- [7] <https://www.efilecabinet.com/what-is-csv-file> (Pristupljeno 29.02.2020.)
- [8] <https://www.google.com./amp/s/m.wikiknow.com> (Pristupljeno 29.02.2020.)
- [9] <https://waikato.github.io/weka-wiki> (Pristupljeno 29.02.2020.)
- [10] <https://www.google.com/amp/s/fileinfo.com> (Pristupljeno 29.02.2020.)
- [11] <https://machinelearningmastery.com/load-csv-machine-learning-data-in-weka>, Jan.23,2016.(Pr.01.03.20)
- [12] <https://www.cs.waikato.ac.nz/ml/weka/arff.html>, Novembar 1st, 2008 (Pristupljeno 01.03.2020.)
- [13] What is Cross-Validation? – Definition from Techopedia, <http://www.techopedia.com/definition>(Pristupljeno 04.12.2020.)
- [14] Cross Validation in WEKA – Stack Overflow, <https://stackoverflow.com/questions/10437677/ceoss-validation-in-weka> (Pristupljeno

04.12.2020.)

- [15] Sarang Narkhede, "Understanding Confusion Matrix", May 9, 2018., <https://towardsdatascience.com> (Prist. 05.12.2020.)
- [16] Sanjay Nandacumar, "Confusion Matrix – are you confused? (Part I and Part II)", <https://sanjaysw05.medium.com> (05.12.20.)
- [17] https://en.wikipedia.org/wiki/Comma-separated_values, Okt 2005 (Pristupljeno 01.03.2020.)
- [18] <https://machinelearningmastery.com/binary-classification-tutorial-weka/> Dec.11,2019(Prist.06.03.2020.)
- [19]<https://medium.com/thalus-ai/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b> Nov. 11, 2017 (Pristupljeno 16.03.2020.)
- [20] <https://dzone.com/articles/ml-metrics-sensitivity-vs-specificity-difference> Sep. 18, 2018. (Pristupljeno 16.03.2020.)
- [21]<https://towardsdatascience.com/metrics-for-evaluating-machine-learning-classification-models-python-example-59b905e079a5> Jun7,19(16.03.20.)
- [22]<https://www.google.com/search?q=how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-3-cclassification-3-eac420ec991&rlz=1C1> Sep. 12, 2018. (Pristupljeno 16.03.2020.)
- [23] <https://machinelearningmastery.com/how-to-run-your-first-classifier-in-weka/> Aug.22,19.(Pr. 19.03.2020.)
- [24] <https://machinelearningmastery.com/design-and-run-your-first-experiment-in-weka/> Aug. 22, 2019. (Pristup. 20.03.2020.)
- [25] <https://www.cs.utexas.edu/~mooney/cs391L/hw2/Experiments.pdf> March 8, 2001. (Prist. 20.03.2020.)
- [26] <http://datamining.liacs.nl/DaMi2013-2014/index.php?t=practicum-2> (Pristupljeno 21.03.2020.)
- [27] http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf January 21, 2013. (Pristupljeno 21.03.2020.)
- [28] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. (Pr. 22.03.2020.)
- [29] <https://ieeexplore.ieee.org/document/1490524> Sept. 2005. (Pristupljeno 22.03.2020.)
- [30] <https://machinelearningmastery.com/confusion-matrix-machine-learning/> Jan. 12. 20. (Pr. 23.03.2020.)
- [31] <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62> May 9, 18. (Pr. 23.03.20.)
- [32] J.Đ.Novaković, „Rešavanje klasifikacionih problema mašinskog učenja“, Fakultet tehničkih nauka u Čačku“, 2013.
- [33] <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2> Aug.,24, 19. (Pr. 28.03.20.)
- [34] R.R.Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Sewald, D.Seuse, "WEKA Manual for Version 3-7-8", January 21, 2013. (Pristupljeno 31.03.2020.)
- [35] <https://classeval.wordpress.com/introduction/basic-evaluation-measures/> Dec.20,16.(Prist.02.04.2020.)
- [36]https://www.researchgate.net/publication/323130913_A_novel_model_based_on_non_invasive_methods_for_prediction_of_liver_fibrosis Dec. 2019. (Pristupljeno 04.04.2020.)
- [37] <https://machinelearningmastery.com/confusion-matrix-machine-learning/> Jan.20,20.(Prist, 07.04.20.)
- [38] <https://machinelearningmastery.com/k-fold-cross-validation/> Aug.8,2019. (Prist. 09.04.2020.)
- [39] Charman Chan, „What is a ROC Curve and How to Interpret it“, <https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/> *Accessed, 18.04.2021.)
- [40] Takaya Saito, Marc Rechmsmeier, „The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets“, Published online 2015 March, doi: 10.1371/journal.phone.0118432 (Accessed, 18.04.2021.)
- [41] Introduction to precession-recall plot, <https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/> (Accessed, 18.04.2021.)
- [41] Suzane Euklend, „Precesion-recall curves-what are they and how are they used“, March 2017, <https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used> (Accessed, 18.04.2021.)
- [43]Anania Uberoi, „Precision-Recall Curve|ML“, Last Updated: 19.jul 2019. <https://www.geeksforgeeks.org/precision-recall-curve-ml/> (Accessed, 18.04.2021.)

NEW REFERENCES:

1. Ahmed Iqbal¹, Shabib Aftab², Umair Ali³, Zahid Nawaz⁴, Laraib Sana⁵, Munir Ahmad⁶, Arif Husen⁷ "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
2. Majdah Alshammari¹, Mohammad Mezher², "A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020

3. Maruthi Rohit Ayyagari, "Classification of Imbalanced Datasets using One-Class SVM, k-Nearest Neighbors and CART Algorithm", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 2020
4. Abdulaziz Albahr1, Marwan Albahr2, "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
5. Hector Daniel Huapaya, Ciro Rodriguez, Doris Esenarro, "COMPARATIVE ANALYSIS OF SUPERVISED MACHINE LEARNING ALGORITHMS FOR HEART DISEASE DETECTION", 3C Tecnología. Glosas de innovación aplicadas a la pyme. ISSN: 2254 – 4143 Edición Especial Special Issue Abril 2020
6. Tom Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers", © 2004. Kluwer Accademic Publishers
7. Alaa Tharwat, "Classification assesment methods", Applied Computing and Informatics, Volume 17, Issue 1, 30. July 2020.
8. Binish Khan, Piyush Kumar Shukla, Manish Kumar Ahirwar, "Strategic Analysis in Prediction of Liver Disease Using Different Classification Algorithms," *International Journal of Computer Sciences and Engineering*, Vol.7, Issue.7, pp.71-76, 2019.
9. Singh, Jagdeep, Bagga, Sachin, Kaur, Ranjodh, "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques", *Procedia Computer Science*, Volume 167, 2020, Pages 1970-1980
10. Chicco, D.; Tötsch, N.; Jurman, G. (2021). "The Matthews correlation coefficient (MCC) is more reliable than *balanced accuracy*, *bookmaker informedness*, and *markedness* in two-class confusion matrix evaluation". *BioDataMining*. **14**. doi:10.1186/s13040-021-00244-z.
11. Chicco D., Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.

AUXILIARY REFERENCES:

1. <https://medium.com/@shubhanshi.shubh860/confusion-matrix-in-machine-learning-cd7333d72f5d> 14, 2018. (Pristupljeno 1.04.2020.)
2. <https://machinelearningmastery.com/what-is-imbalanced-classification/> 14, 2020. (Pristupljeno 15.04.2020.)
3. <https://www.quora.com/What-is-an-imbalanced-dataset> Jul 26, 2015. (Pristupljeno 15.04.2020.)

Figures

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

Figure 1

A caption was not provided with this version

Model 2		Stvarne vrijednosti →			
Predviđena vrijednost ↓	A	198	7	0	2
	B	2	1	8	3
	C	0	0	1	4
	D	0	2	1	1

Figure 2

a four-class confusion matrix.

Four outcomes of a classifier

Figure 3

Oval representation of the four binary classification results of the test dataset

Figure 4

Two ovals show how to calculate accuracy

Figure 5

Two ovals show how sensitivity is calculated

Figure 6

Two ovals show how it is calculated

Figure 7

Two ovals show how to calculate a false positive rate - FPR

Figure 8

Two ovals show how precision is calculated

Figure 9

Recall

Figure 10

ROC Curves

Figure 11

Precision-Recall Curve