Computational Data Analysis - Classification

Lavdim Beqiri[1][lavdim.beqiri@ubt-uni.net], Greta Ahma[2][greta.ahma@ubt-uni.net] and Edmond Hajrizi
3[ehajrizi@ubt-uni.net]

[1] UBT - Higher Education Institution, Faculty of Computer Science and Engineering, Prishtine, Kosove
[2] UBT - Higher Education Institution, Faculty of Computer Science and Engineering, Prishtine, Kosove
3 UBT - Higher Education Institution, Faculty of Computer Science and Engineering, Prishtine, Kosove

Abstract. The research discusses Computational Data Analysis Classification, a short summary of the classification models (regression, classification and clustering), the particular focus is on classification. For a case study, for data analysis, it has been selected classification method. The comparisons were made by classifying and analyzing the processes from the datasets between algorithms: Naïve Bayes, Support Vector Machine (SVM), J48 Decision Tree, and KStar.

Keywords: Machine learning, algorithm, text, machine learning algorithms, weka, classification.

1 Introduction

In many applicative sciences such as medicine, engineering and finance, and many others, data modeling and analysis are of particular importance. Many different distributions have to model this data. The quality of the procedure used in the statistical analysis depends heavily on the distribution that is chosen to model this data. However, there are still many real details that do not fit any classical distribution or standard model. Therefore, there is a need for generating new distributions.

2 Computational Data Analysis Classification

There are many types of algorithms that are used in different fields, such as statistics, machine learning, etc. For the selection of the algorithm, we must think carefully because it is not an easy task *(Henk A. L. Kiers . Jean-Paul Rasson Patrick J. F. Groenen . Martin Schader (Eds.), 2000, p. 119*).

Machine learning encompasses many different algorithms, some with a wide range of applicability, while others may be suited for specific applications. These algorithms can be divided in two main categories: supervised and unsupervised. Supervised machine learning algorithms are the most commonly used machine learning

algorithms for predictive analytics. These algorithms rely on data sets that have been processed by human experts (hence the word "supervision"). The algorithms then learn how to perform the same processing tasks autonomously on new data sets. In particular, supervised methods are used to solve regression and classification problems:

Regression problems. This makes the evaluation of mathematical relationships between changes or many changes that are made continuously. This mathematical relationship is used to calculate values taking into account other cognitive values. A typical example of regression is GPS, which estimates the position and speed of the car, or another example is weather forecasting, or predicting the future value of a stock using historical data and other sources of information. To mentally visualize the simplest example of regression, imagine two variables, whose values are visualized as points in 2D plot similar to the image on the right side of Figure 2. Performing regression means finding the line that best interpolates the values. The line can take a lot of different shapes and is expressed as a regression function *(Leblanc, 1991*).

Classification problems. It is used when the unknown variable is discrete, usually the problem involves estimating which, from a set of predefined classes, belongs to the specific instance. Examples of classification methods are image recognition, identification faces in picture, or identification of emotional expressions. The interpretation of the classifier can be seen in two dimensions, where the points belonging to different classes have different symbols, the example shown in figure 1. The algorithm learns examples of location and shape between classes. This boundary line can then be used to classify new examples (*Henk A. L. Kiers . Jean-Paul Rasson Patrick J. F. Groenen . Martin Schader (Eds.), 2000, p. 211*).

Unsupervised. A learning method that uses unlabeled data is known as an unsupervised learning method, in contrast to supervised learning methods, which use labeled data *(Abney, 2008, p. 3*). Examples of problems solved with unsupervised methods are clustering and association.

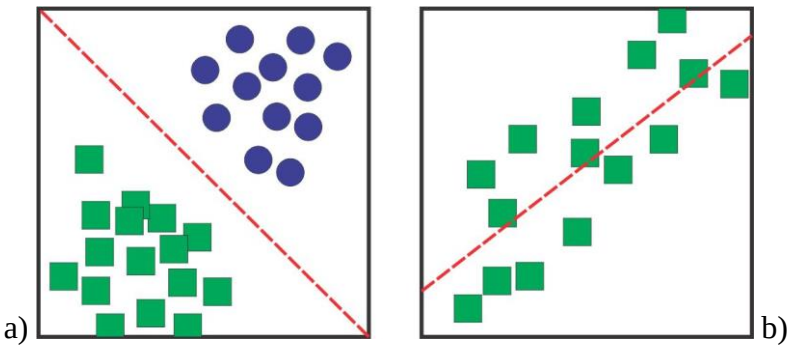


Fig. 1. a) Classification, b) Regression

Clustering methods. A simple procedure for choosing the base of the function centers is to place them equally with a random or stratified sample. As an improvement, clustering techniques can be used to find the center that more accurately represents the data distribution in the predictive space. Moody and Darken (1999) have proposed the k-means clustering algorithm, and this is used very often.

These can be seen as the automatic discovery of groups of samples that have similar characteristics, which can possibly point to the fact that a member of the cluster belongs to a well-defined class (*Henk A. L. Kiers . Jean-Paul Rasson Patrick J. F. Groenen . Martin Schader (Eds.), 2000, p. 213*). Clustering algorithms are used to identify groups of users based on their on-line purchasing history, and then send to each member targeted ads. In Figure 3, the clustering algorithm has automatically assigned a different color to group of observations that are "close" to each other.

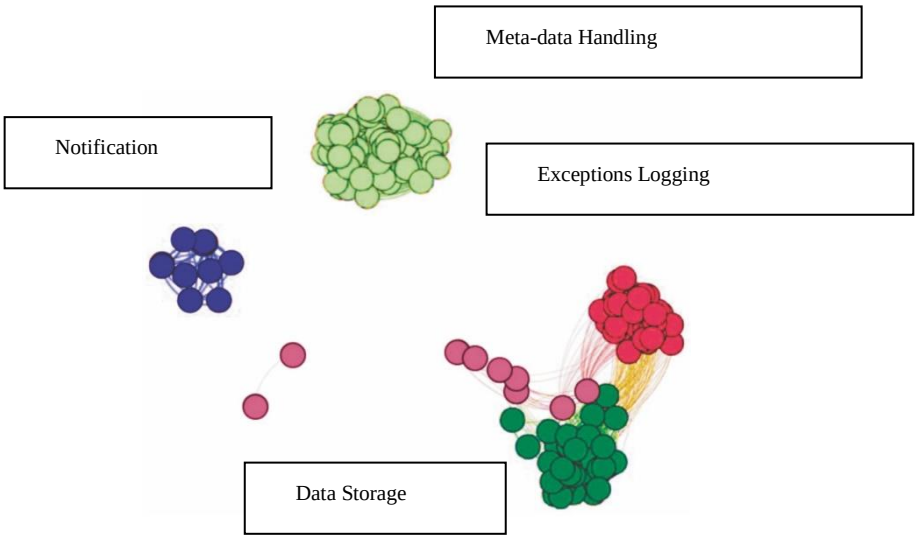


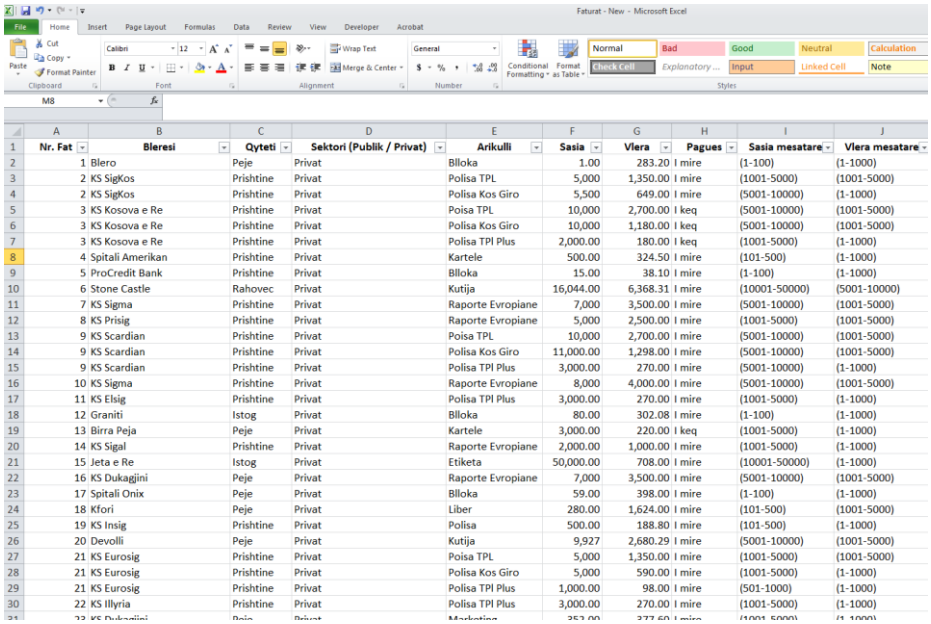
Fig. 2. Clustering algorithm

3 Data Analysis based on Data Mining Algorithms (Case Study)

3.1 Introduction

Machine learning is all about learning rules from the data set. In this paper, is used classification and analysis processes on the data collected from bills during 2017 in weka application.

Naïve Bayes classifier, Support Vector Machine (SVM), Decision tree, and KStar, have been used in order to analyze the results (*Ian H. Witten, 2011*).



Nr. Fat	Bleresi	Qyteti	Sektori (Publik / Privat)	Arikulli	Sasia	Vlera	Pagues	Sasia mesatare	Vlera mesatare
1	Blero	Peje	Privat	Blloka	1.00	283.20	I mire	(1-100)	(1-1000)
2	KS SigKos	Prishtine	Privat	Polisa TPL	5,000	1,350.00	I mire	(1001-5000)	(1001-5000)
2	KS SigKos	Prishtine	Privat	Polisa Kos Giro	5,500	649.00	I mire	(5001-10000)	(1-1000)
3	KS Kosova e Re	Prishtine	Privat	Poisa TPL	10,000	2,700.00	I keq	(5001-10000)	(1001-5000)
3	KS Kosova e Re	Prishtine	Privat	Polisa Kos Giro	10,000	1,180.00	I keq	(5001-10000)	(1001-5000)
3	KS Kosova e Re	Prishtine	Privat	Polisa TPl Plus	2,000.00	180.00	I keq	(1001-5000)	(1-1000)
4	Spitali Amerikan	Prishtine	Privat	Kartele	500.00	324.50	I mire	(101-500)	(1-1000)
5	ProCredit Bank	Prishtine	Privat	Blloka	15.00	38.10	I mire	(1-100)	(1-1000)
6	Stone Castle	Rahovec	Privat	Kutija	16,044.00	6,368.31	I mire	(10001-50000)	(5001-10000)
7	KS Sigma	Prishtine	Privat	Raporte Evropiane	7,000	3,500.00	I mire	(5001-10000)	(1001-5000)
8	KS Prisig	Prishtine	Privat	Raporte Evropiane	5,000	2,500.00	I mire	(1001-5000)	(1001-5000)
9	KS Scardian	Prishtine	Privat	Poisa TPL	10,000	2,700.00	I mire	(5001-10000)	(1001-5000)
9	KS Scardian	Prishtine	Privat	Polisa Kos Giro	11,000.00	1,298.00	I mire	(5001-10000)	(1001-5000)
9	KS Scardian	Prishtine	Privat	Polisa TPl Plus	3,000.00	270.00	I mire	(5001-10000)	(1-1000)
10	KS Sigma	Prishtine	Privat	Raporte Evropiane	8,000	4,000.00	I mire	(5001-10000)	(1001-5000)
11	KS Elsig	Prishtine	Privat	Polisa TPl Plus	3,000.00	270.00	I mire	(1001-5000)	(1-1000)
12	Graniti	Istog	Privat	Blloka	80.00	302.08	I mire	(1-100)	(1-1000)
13	Birra Peja	Peje	Privat	Kartele	3,000.00	220.00	I keq	(1001-5000)	(1-1000)
14	KS Sigal	Prishtine	Privat	Raporte Evropiane	2,000.00	1,000.00	I mire	(1001-5000)	(1-1000)
15	Jeta e Re	Istog	Privat	Etiketa	50,000.00	708.00	I mire	(10001-50000)	(1-1000)
16	KS Dukagjini	Peje	Privat	Raporte Evropiane	7,000	3,500.00	I mire	(5001-10000)	(1001-5000)
17	Spitali Onix	Peje	Privat	Blloka	59.00	398.00	I mire	(1-100)	(1-1000)
18	Kfori	Peje	Privat	Liber	280.00	1,624.00	I mire	(101-500)	(1001-5000)
19	KS Insig	Prishtine	Privat	Polisa	500.00	188.80	I mire	(101-500)	(1-1000)
20	Devolli	Peje	Privat	Kutija	9,927	2,680.29	I mire	(5001-10000)	(1001-5000)
21	KS Eurosig	Prishtine	Privat	Poisa TPL	5,000	1,350.00	I mire	(1001-5000)	(1001-5000)
21	KS Eurosig	Prishtine	Privat	Polisa Kos Giro	5,000	590.00	I mire	(1001-5000)	(1-1000)
21	KS Eurosig	Prishtine	Privat	Polisa TPl Plus	1,000.00	98.00	I mire	(501-1000)	(1-1000)
22	KS Illyria	Prishtine	Privat	Polisa TPl Plus	3,000.00	270.00	I mire	(1001-5000)	(1-1000)

Fig. 3. Data from 2017

3.2 Methodology

In this paper, is made a research to analyze the data to compile the sales strategy. Analysis are based on billing data during 2017 year, for classification the data by dividing cities, the public or private sector, the amount and value. These data serve to build a marketing strategy in which private sector, or public, or in which region to invest.

These dates contain around all 2017 bills, 357 registred invoices. The instances are described by 5 attributes, and one class attribute, some of which are linear and some are nominal.

The weka application is used for machine learning. It contains tools for, classification, regression, clustering, association rules mining, and visualization (*https://www.cs.waikato.ac.nz/ml/weka/, 2022*).

The data set description with attributes, and their values are given below.

Table 1. Attributes and values of billing data 2017.

Attributes	Values
Qyteti/City	Decan, Gjakove, Gjilan, Istog, Mitrovice, Peje, Podujeve, Prishtine, Prizren, Rahovec, Suhareke
Sektori/Sector	Privat, Publik / Private, Public
Pagues/Payer	I mire, I keq / Good, Bad
Vlera mesatare/Average Value	(1-1000), (1001-5000), (5001-10000), (10001-100000), (100001-1000000)
Saisa mesatare/Average Amount	(1-100), (101-500), (501-1000), (1001-5000), (5001-10000), (10001-50000), (50001-100000), (100000-1000000)

The visualization of attributes can be shown at the Figure 10 below.

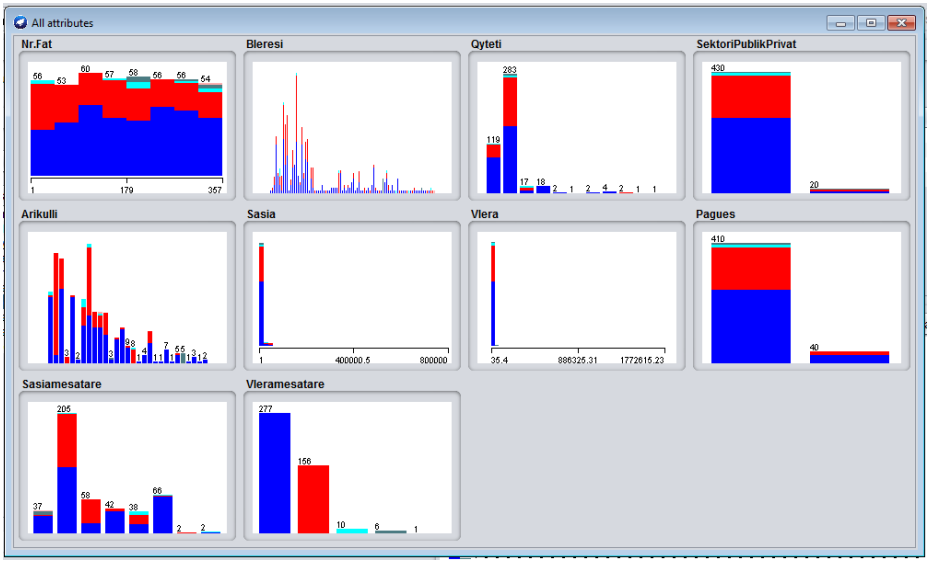


Fig. 4. Attributes on data set

3.3 Machine Learning Algorithms and results

Data mining is process of sorting large data sets to identify relations and models that can help solve problems throguht data.

The main purpose of the date process is to extract a meaningful form and this is done with the help of the algorithm.

For data mining development, three algorithms were used and tested, to determine, analyze and extract the results. The purpose of the results is the comparison between the algorithms which is the most appropriate for comparing this type of datas. The algorithms used are: Naïve Bayes classifier, Support Vector Machine (SVM), J48 Decision Tree, and KStar. Thus, we can easily realize the difference between the algorithm results. To sum up, the best classification on the data set is understandable.

3.4 Classification

Classification is a common machine learning problem. In recent years, great improvements have been made, especially in the field of image recognition. Classifiers can be seen as regression problems where the target variable is discrete, and represents the class that the domain expert has classified as a given example. It is common, in classification problems, to provide not only a set of examples data points each class, but to also establish which are the features of each data point that are more useful to estimate the corresponding class. These features can be readily available from the sensors, but more often need to be computed (or extracted) from the raw data before being fed to the learning algorithm. The definition of relevant features is a crucial step that, with the exception of very advanced algorithms such as Deep Learning, relies on human expert knowledge.

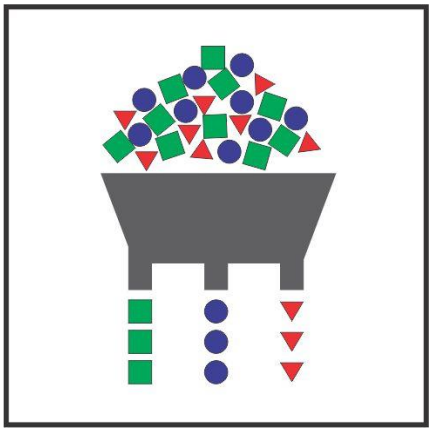


Fig. 5. Classification

There are numerous classifier algorithms that are popular for various purposes. We will briefly discuss some of them and compare results.

Naïve Bayes Classifier. Naïve Bayes shown in figure 6, is a statistical learning algorithm that applies a simplified version of Bayes rule in order to compute the posterior probability of a category given the input attribute values of an example situation. Prior probabilities for categories and attribute values conditioned on categories are estimated from frequency counts computed from the training data. Naïve Bayes is a simple and fast learning algorithm that often outperforms more sophisticated methods. The Bayesian classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

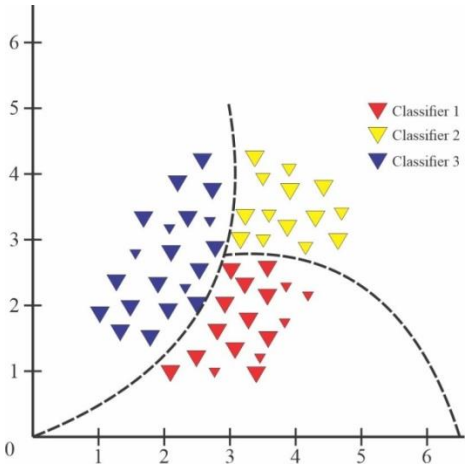


Fig. 6. Naïve Bayes Classifier

In table 2, are the results obtained from the Naive Bayes Classifier algorithm. The correctly classified instances are 78.22% and incorrectly classified instances are 21.78%, the Kappa statistics is 0.5282, and the mean absolute error is 0.1009.

Table 2. Results from Naïve Bayes Classifier

Correctly Classified Instances (%)	78.22
Incorrectly classified instances (%)	21.78
Kappa Statistics	0.5282
Mean Absolute Error	0.1009

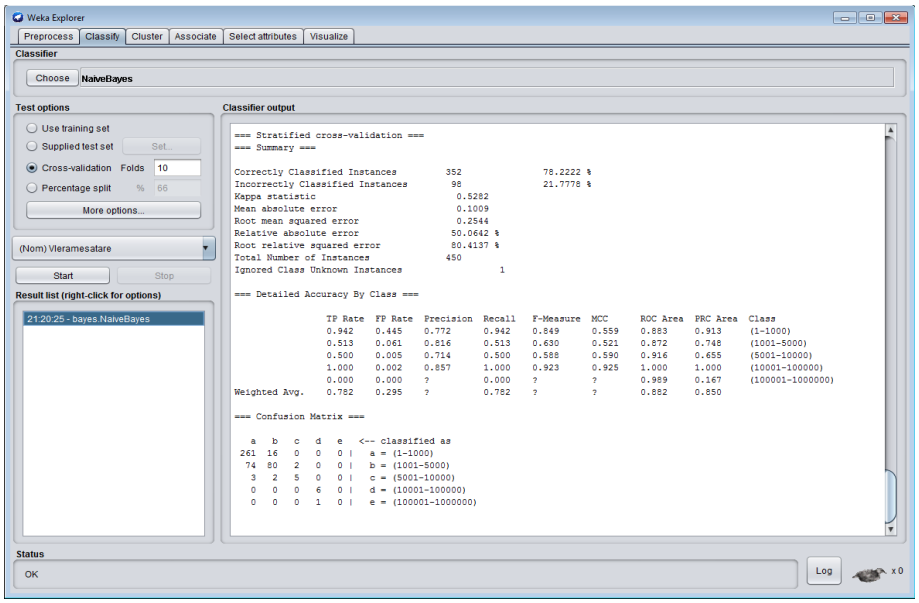


Fig. 7. Naïve Bayes classifier results on Weka

Support Vector Machine (SVM). Figure 8 shows the SVM algorithm in an illustrated way. H3 maximizes the distance between the training point and other classes. When the new date comes in, it will be classified based on one or the other side of H3. This algorithm implements minimal optimization for training a vector classifier using polynomial kernels. This implement replaces all values and transformations of nominal attributes into a binary ones, also normalizing all attributes. This is done because the output coefficients are based on the normalized dates, not on the original dates, this is very important for the interpretation of the classifier.

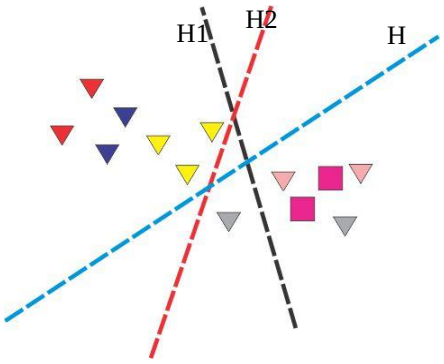


Fig. 8. Support Vector Maxhine

In table 3, are the results obtained from the Support vector machines (SVM) algorithm. The correctly classified instances are 80.22% and incorrectly classified instances are 19.78%, the Kappa statistics is 0.6023, and the mean absolute error is 0.2488.

Table 3. Results from Support Vector Machine Classifier

Correctly Classified Instances (%)	80.22
Incorrectly classified instances (%)	19.78
Kappa Statistics	0.6023
Mean Absolute Error	0.2488

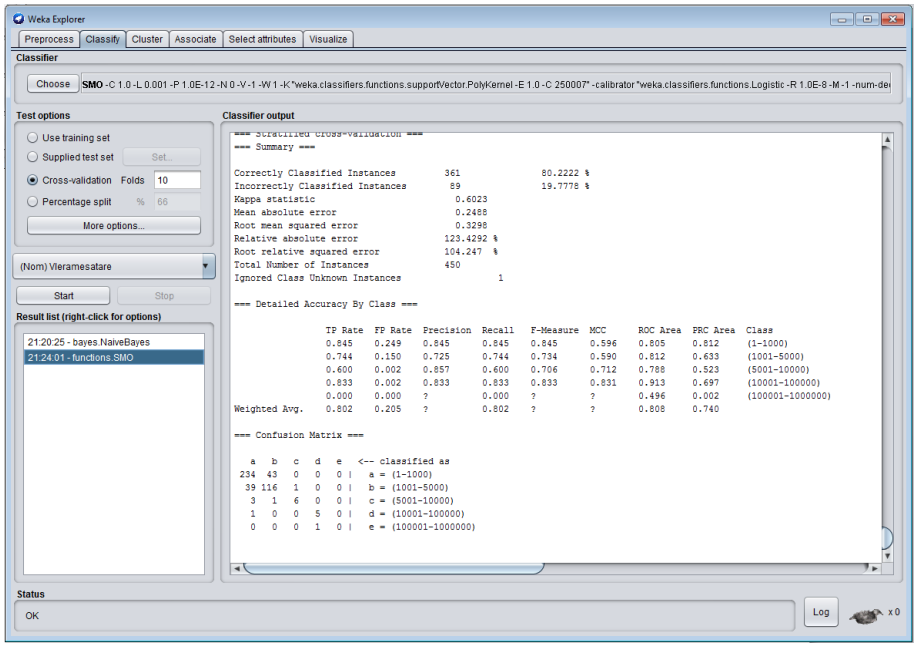


Fig. 9. Support Vector Machine classifier results on Weka

J48 Decision Tree. Is a classifier expressed as a recursive partitioning of space. It consists of the tree with roots that make up the node. The node with the output edge is called an internal node, all other nodes are called leaves, or decisive terminal nodes. Each internal node divides the space into two or more subspaces according to a certain discrete function, based on the input attributes. A simplified view of a binary decision tree and the types of nodes is shown in figure 10.

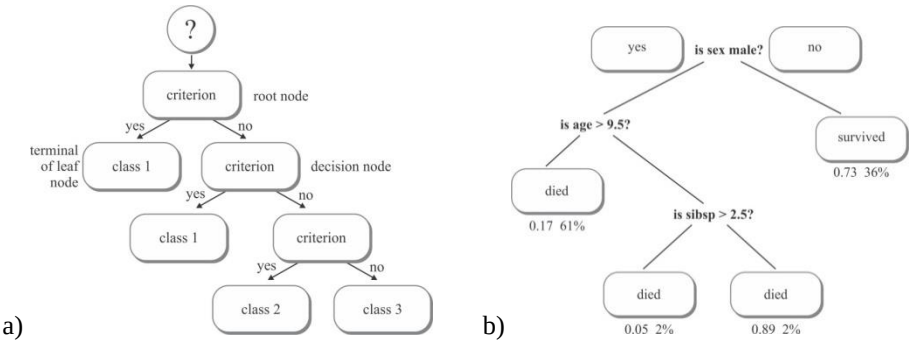


Fig. 10. a) Binary Decision Tree, b) Simple Binary Decision Tree Visualization

In table 4, are the results obtained from the J48 Decision Tree algorithm. The correctly classified instances are 98.89% and incorrectly classified instances are 1.11%, the Kappa statistics is 0.9777, and the mean absolute error is 0.008.

Table 4. Results from J48 Decision Tree

Correctly Classified Instances (%)	98.89
Incorrectly classified instances (%)	1.11
Kappa Statistics	0.9777
Mean Absolute Error	0.008

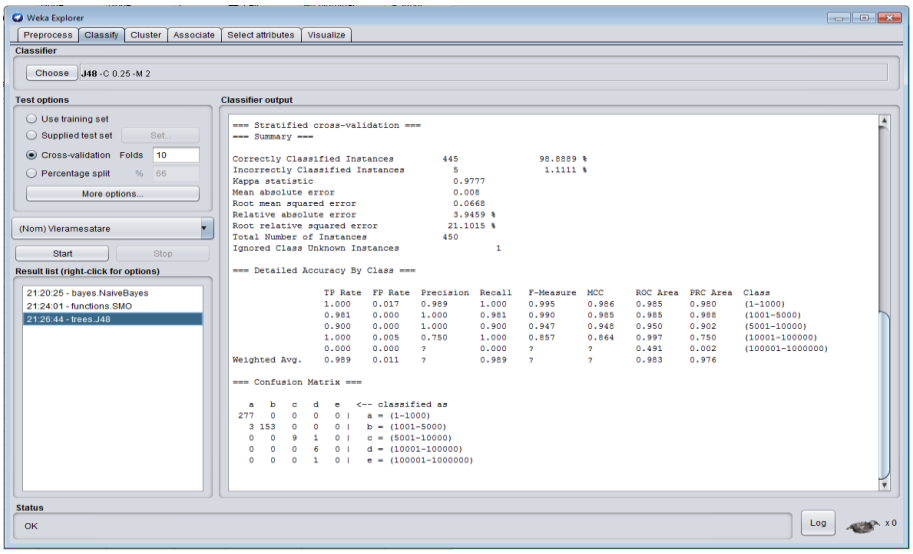


Fig. 11. J48 Algorithm For Decision Tree classifier classifier results on Weka

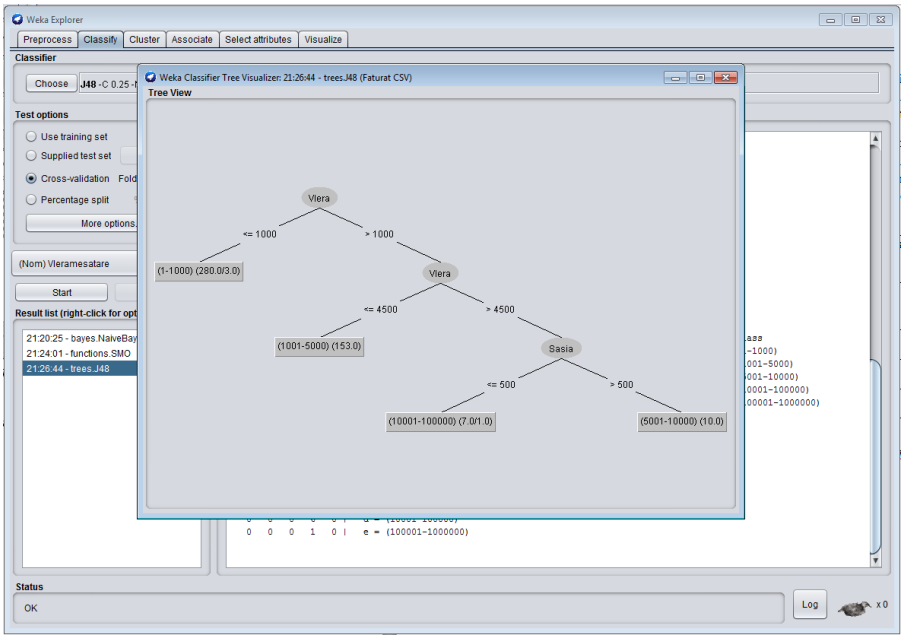


Fig. 12. Decision tree visualization

KStar. An instance-based classifier is the class of a test example that is based on the class of similar training examples, as determined by some similarity function.

In table 5, are the results obtained from the Kstar is an instance-based classifier. The correctly classified instances are 93.11% and incorrectly classified instances are 6.89%, the Kappa statistics is 0.86, and the mean absolute error is 0.0258.

Table 5. Results from KStar Classifier

Correctly Classified Instances (%)	93.11
Incorrectly classified instances (%)	6.89
Kappa Statistics	0.86
Mean Absolute Error	0.0258

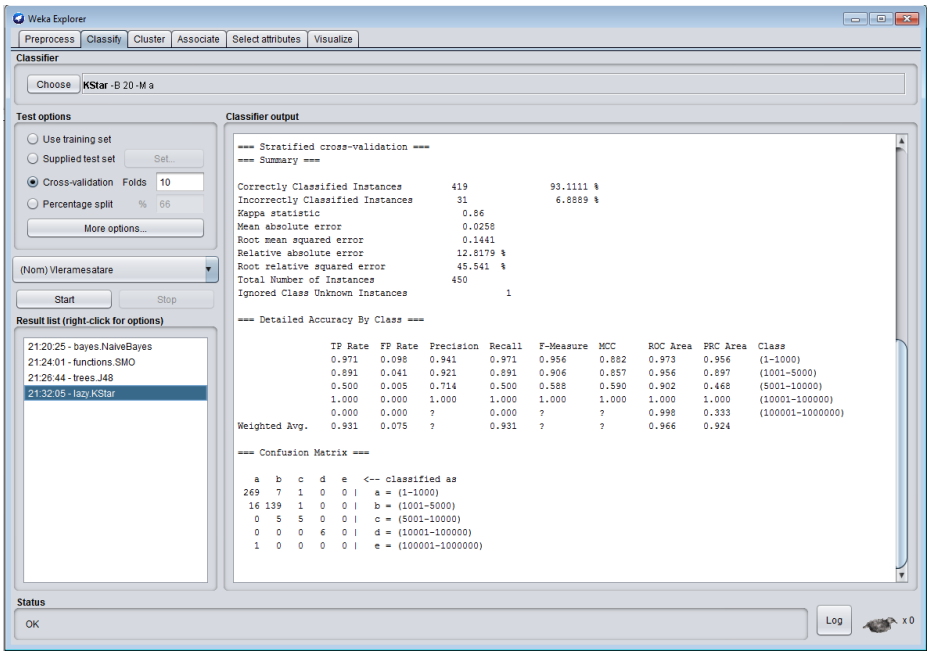


Fig. 13. KStar classifier results on Weka

4 Conclusion / Study results

From the obtained results, are obtained the best and worst conditions. The KStar algorithm has realized the best model with the value of Kappa Statistics=0.86. While the worst realized model is Naïve Bayes Classifier algorithm=0.5282.

J48 algorithm with % 98.89 accuracy classification percent rate realized to best classifications on data set. Naïve Bayes Classifier algorithm with % 78.22 accuracy classification percent rate realized to worst classifications on data set.

Mean absolute error of J48 algorithm has fixed to maximum value (0.008). Mean absolute error of SVM algorithm has fixed to minimum value (0.2488).

In this paper, Naïve Bayes classifier, SVM (support vector machine), decision tree and KStar (Instance-based classifier) have been analyzed on the Weka workbench.

Here, using machine learning algorithms, analysis made with obtained experimental results from classifications on the billing 2017 data set. Every machine learning algorithms have been applied separately on all data set, and classification results have denoted in Table 2.

The kappa statistic is used as a means of classifying agreement in categorical data. KS (Kappa Statistic) is used as a means of classifying agreement in categorical data. A kappa coefficient of 1 means a statistically perfect modeling whereas a 0 means every model value was different from the actual value. KS values for each algorithm have been calculated separately with the help of Weka functions.

Table 6. Performance analysis results of machine learning algorithms on 2017 data set

Algorithms	Correctly Classified Instances (%)	Kappa Statistics	Mean Absolute Error
Naïve Bayes Classifier	78.22	0.5282	0.1009
Support Vector Machine	80.22	0.6023	0.2488
J48 Decision Tree	98.89	0.9777	0.008
KStar	93.11	0.86	0.0258

While some of the algorithms show high performance, and some of them show poor performance. While some of the algorithms make the best modeling and some of them make the worst modeling.

References

1. Henk A. L. Kiers . Jean-Paul Rasson Patrick J. F. Groenen . Martin Schader (Eds.), "Data Analysis, Classification, and Related Methods", Springer-Verlag Berlin· Heidelberg 2000
2. Leblanc, M., Tibshirani, R., "Combining Estimates in Regression and Classification", Journal of the American Statistical Association, 1641-1650, 1994
3. Abney, Steven P. "Semisupervised Learning for Computational Linguistics", London, 2008
4. Ian H. Witten, Eibe Frank, Mark A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", Amsterdam, Netherlands, 2011.
5. T. Hastie, R.Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition), Springer, 2009.
6. Duda, R.O., Hart, P.E., and Stork D.G Pattern classification, John Wiley and Sons, New York, NY, 2001.
7. Han J. and Kamber M. Data Mining Concept and Techniques. London: Morgan Kaufmann Publishers, 2001.
8. E. Alpaydin. Introduction to Machine Learning (2nd ed.). Cambridge, MA: MIT Press, 2011.
9. J. R. Quinlan. Induction of decision trees, Machine Learning, f. 81–106, 1986.
10. J. R. Quinlan. Programs for Machine Learning. Morgan Kaufmann, 1993.
11. Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994. Machine Learning, Neural Statistical Classification, Ellis Horwood.
12. Witten, I.H. & Frank, E. 2000. Weka Machine Learning Algorithms in Java, in Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, pp. 260-320.
13. Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. Data Mining in Bioinformatics using Weka, Bioinformatics Applications Note, pp. 2475-2481, (2004)
14. Tan, M. & Eshelman, L. 1988. Using Weighted Networks to Represent Classification Knowledge in Noisy Domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.
15. https://www.cs.waikato.ac.nz/ml/weka/, access on 14.10.2022