# **Computational Data Analysis - Classification**

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

<sup>1</sup> UBT - Higher Education Institution, Faculty of Computer Science and Engineering, Prishtine, Kosove

<sup>2</sup> 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*).

X   6	<b>*) •</b> (* ~   <del>•</del>								Fa	iturat - P	lew - Microsoft	Excel			
File	Home	Insert Page Layout Form	· A^ A = =	v View	Developer Acr	General		•		Norr	nal Bi	ad	Good N	eutral	Calculation
Paste	Format Pair	B I U - 🖽 - 🔮	•• <u>∧</u> •≡≡≡	课课	🗚 Merge & Center =	s - %	, 14	& Condition	al Format	Chec	k Cell 🗾 🕖	xplanatory	Input Li	nked Cell	Note
	Clipboard	G Font	0	Alignme	nt G	N	umber	Formatting	* as lable *			Style	15		
	M8	• (= fx													
	Α	В	С		D			E	F		G	н	1		J
1	Nr. Fat 🕞	Bleresi	- Qyteti -	Sek	tori (Publik / Pri	vat) 🕞	Ari	kulli 👻	Sasia		Vlera 🗸	Pagues -	Sasia mesat	are -	Vlera mesatare 🗸
2	1	Blero	Peje	Privat			Blloka		1	.00	283.20	I mire	(1-100)	(1	-1000)
3	2	KS SigKos	Prishtine	Privat			Polisa TP	L	5,0	000	1,350.00	I mire	(1001-5000)	(1	.001-5000)
4	2	KS SigKos	Prishtine	Privat			Polisa Ko	s Giro	5,	500	649.00	I mire	(5001-10000)	(1	-1000)
5	3	KS Kosova e Re	Prishtine	Privat			Poisa TPL		10,0	000	2,700.00	l keq	(5001-10000)	(1	001-5000)
6	3	KS Kosova e Re	Prishtine	Privat			Polisa Ko	s Giro	10,0	000	1,180.00	l keq	(5001-10000)	(1	.001-5000)
7	3	KS Kosova e Re	Prishtine	Privat			Polisa TP	Plus	2,000	.00	180.00	l keq	(1001-5000)	(1	-1000)
8	4	Spitali Amerikan	Prishtine	Privat			Kartele		500	.00	324.50	I mire	(101-500)	(1	-1000)
9	5	ProCredit Bank	Prishtine	Privat			Blloka		15	.00	38.10	1 mire	(1-100)	(1	-1000)
10	e	5 Stone Castle	Rahovec	Privat			Kutija		16,044	.00	6,368.31	I mire	(10001-50000)	(5	001-10000)
11	7	KS Sigma	Prishtine	Privat			Raporte I	vropiane	7,0	000	3,500.00	I mire	(5001-10000)	(1	.001-5000)
12	8	KS Prisig	Prishtine	Privat			Raporte I	vropiane	5,0	000	2,500.00	1 mire	(1001-5000)	(1	001-5000)
13	9	KS Scardian	Prishtine	Privat			Poisa TPL		10,0	000	2,700.00	I mire	(5001-10000)	(1	001-5000)
14	9	KS Scardian	Prishtine	Privat			Polisa Ko	s Giro	11,000	.00	1,298.00	1 mire	(5001-10000)	(1	001-5000)
15	g	KS Scardian	Prishtine	Privat			Polisa TP	Plus	3,000	.00	270.00	I mire	(5001-10000)	(1	-1000)
16	10	KS Sigma	Prishtine	Privat			Raporte I	vropiane	8,0	000	4,000.00	I mire	(5001-10000)	(1	001-5000)
17	11	KS Elsig	Prishtine	Privat			Polisa TP	Plus	3,000	.00	270.00	1 mire	(1001-5000)	(1	-1000)
18	12	Graniti	Istog	Privat			Blloka		80	.00	302.08	I mire	(1-100)	(1	-1000)
19	13	Birra Peja	Peje	Privat			Kartele		3,000	.00	220.00	1 keq	(1001-5000)	(1	-1000)
20	14	KS Sigal	Prishtine	Privat			Raporte I	vropiane	2,000	.00	1,000.00	I mire	(1001-5000)	(1	-1000)
21	15	jeta e Re	Istog	Privat			Etiketa		50,000	.00	708.00	I mire	(10001-50000)	(1	-1000)
22	16	KS Dukagjini	Peje	Privat			Raporte I	vropiane	7,0	000	3,500.00	1 mire	(5001-10000)	(1	001-5000)
23	17	Spitali Onix	Peje	Privat			Blloka		59	.00	398.00	Imire	(1-100)	(1	-1000)
24	18	Kfori	Peje	Privat			Liber		280	.00	1,624.00	1 mire	(101-500)	(1	001-5000)
25	19	KS Insig	Prishtine	Privat			Polisa		500	.00	188.80	I mire	(101-500)	(1	-1000)
26	20	) Devolli	Peje	Privat			Kutija		9,9	927	2,680.29	I mire	(5001-10000)	(1	001-5000)
27	21	KS Eurosig	Prishtine	Privat			Poisa TPL		5,0	000	1,350.00	I mire	(1001-5000)	(1	.001-5000)
28	21	KS Eurosig	Prishtine	Privat			Polisa Ko	s Giro	5,0	000	590.00	I mire	(1001-5000)	(1	-1000)
29	21	KS Eurosig	Prishtine	Privat			Polisa TP	Plus	1,000	.00	98.00	I mire	(501-1000)	(1	-1000)
30	22	KS Illyria	Prishtine	Privat			Polisa TP	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.

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

**Table 1.** Attributes and values of billing data 2017.

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 through 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

Preprocess Classify Cluster Associate Select attributes Visualize	
Classifier	
Choose NatveBayes	
Test options Classifier output	
Use training set	
Supplied test set Set	
© Cross-validation Folds 10 Correctly Classified Instances 352 78.2222 %	
O Percentage split % 66 Incorrectly Classified Instances 98 21.7778 %	
Kappa statistic 0.5282	
More options 0.1009	
Relative absolute error 50.0642 %	
(Nom) Vieramesatare Root relative squared error 80.4137 %	
Total Number of Instances 450	
Start Stop	
Result list (right-click for options) ==== Detailed Accuracy By Class ===	
21:20:25 - bayes.NalveBayes IP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class	
0.942 0.445 0.772 0.942 0.849 0.559 0.883 0.913 (1-1000)	
0.513 0.061 0.816 0.513 0.650 0.521 0.872 0.748 (1001-5000)	
1.000 0.002 0.714 0.500 0.585 0.590 0.916 0.555 (500-10000)	
0.000 0.000 ? 0.000 ? ? 0.989 0.167 (100001-1000000)	
Weighted Avg. 0.782 0.295 ? 0.782 ? ? 0.882 0.850	
=== Confusion Matrix ===	
a h a d a constantified as	
261 16 0 0 0 = a = (1-1000)	
74 80 2 0 0   b = (1001-5000)	
3 2 5 0 0 1 c = (5001-10000)	
0 0 0 1 0 1 0 1 0 (10000-100000)	
	*

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

📿 Weka Explorer									
Preprocess Classify Cluster Associate	Select attributes Visualize								
Classifier									
Choose SMO -C 1.0 -L 0.001 -P 1.0E-12	-N 0 -V -1 -W 1 -K "weka.classifiers.fu	nctions.supportVector.Po	ilyKernel -E	1.0 -C 25000	" -calibrato	or "weka.class	ifiers.functio	ns.Logistic -R 1.0E-8 -M	-1 -num-de
Test options	Classifier output								
O Use training set	Suractified Cross-valid	acion ===							ŕ
Cross-validation Folds 10	Correctly Classified Insta Incorrectly Classified Ins	nces 361 tances 89		80.2222 19.7778	8				- 1
O Percentage split % 66	Kappa statistic Mean absolute error	0.60	23 88						- 1
More options	Root mean squared error Relative absolute error Root relative squared erro	0.32 123.42 r 104.24	98 92 % 7 %						- 1
(Nom) Vieramesatare	Total Number of Instances Ignored Class Unknown Inst	450 ances	1						- 1
Start Stop Result list (right-click for options)	Detailed Accuracy By C	lass ===							- 1
21:20:25 - bayes NalveBayes 21:24:01 - functions.SMO	TP Rate 0.845 0.744 0.600 0.833 0.000	FP Rate Precision 0.249 0.845 0.150 0.725 0.002 0.857 0.002 0.833 0.000 2	Recall 0.845 0.744 0.600 0.833 0.000	F-Measure 0.845 0.734 0.706 0.833 2	MCC 0.596 0.590 0.712 0.831	ROC Area 0.805 0.812 0.788 0.913 0.496	PRC Area 0.812 0.633 0.523 0.697 0.002	Class (1-1000) (1001-5000) (5001-10000) (10001-100000)	
	Weighted Avg. 0.802	0.205 ?	0.802	?	?	0.808	0.740		- 1
	=== Confusion Matrix ===								- 1
	a b c d e < 234 43 0 0 0   a 3916 1 0 0 0 b 3 1 6 0 0   c 1 0 0 5 0   d 0 0 0 1 0   e	classified as = (1-1000) = (1001-5000) = (5001-10000) = (10001-100000) = (100001-1000000)							
	•								7 +
Status									
ок								Log	

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.





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

📿 Weka Explorer		
Preprocess Classify Cluster Associate	Select attributes Visualize	
Classifier		
Choose J48 -C 0.25 -M 2		
Test options	Classifier output	
O Use training set O Supplied test set Set	=== Stratified cross-validation ===	ŕ
Cross-validation Folds 10     Percentage split % 66     More options	Correctly Classified Instances         445         90.889 %           Incorrectly Classified Instances         5         1.111 %           Apps statistic         0.9977         1.111 %           Mean Abbolute error         0.008         0.0668           Moor mean pupared error         0.0668         1.111 %	
(Nom) Vieramesatare	Melafive absolute error 3.9499 % Root relative aquared error 21.1015 % Total Number of Instances 450 Ignored Class Unknown Instances 1 mem Detailed Accuracy By Class mem	
212025 - bayes NalveBayes 212401 - functions SMO 2120-44 - trees J48	TP Rate FP Rate Precision Recall F-Messure NCC ROC 1.000 0.017 0.989 1.000 0.995 0.986 0.1 0.981 0.000 1.000 0.981 0.199 0.985 0.946 0.5 0.980 0.000 1.000 0.981 0.197 0.948 0.5 0.000 0.000 0.900 0.957 0.144 0.5 0.000 0.000 0.900 0.97 0.144 Weighted Avg. 0.989 0.011 7 0.989 7 7 0.5 === Confusion Matrix === a b c d e < classified as 27 0.0 0 0 1 a = (1-1000)	2 Area FRC Area Class 185 0.980 (1-1000) 186 0.981 (1001-8000) 186 0.902 (1001-10000) 180 0.902 (1001-100000) 197 0.002 (10001-100000) 1983 0.976
Status OK	3 133 0 0 0 0   P = (1001-3000) 0 0 0 0 0 0 0 0   P = (1001-3000) 0 0 0 0 1 0   e = (10000-150000) 0 0 0 1 0   e = (100001-1000000)	LogX0

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.

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

Table 5. Results from KStar Classifier

(r 109 voise)       Vestorie       Vestorie       Vestorie         (Choose       KSlar #2 20 · M a         Test options       Classifier output         • Use training set       • • • • Statified cross-validation ===         • • • Statified cross-validation ===       • • • • • • • • • • • • • • • • • • •	Weka Explorer	Palact attributes Mousting							(	
Classifier output            Use training set         Suppled test set         Suppled testest         Suppled test set         Suppled test set	Classifier	e Seleci attributes visualize								
Choose KNar-820-M a         Test options            Use training set         Sumary ==         Sumary ==         Correctly Classified Instances         Sist 0         More options.             O Coss-waldadon Folds 10         More options.             More options.             More options.             More options.             Choose National Folds 10             More options.             More options.             Stat            Stati (ight-click for options)              Stati (ight-clic										
Testified cross-validation ===         • Use training set       • • • • • • • • • • • • • • • • • • •	Choose KStar -B 20 -M a									
Use training set       === Stratified cross-validation ===         © Cross-validation Folds 0       0         © Cross-validation Folds 0       0         @ More options.       0.66         More options.       0.65         Result if (ght.Cick for options)       0.1441         Procentage split % 06       0.551 %         Result if (ght.Cick for options)       1         212025 - Days Na/wBayes       1         212025 - Days Na/wBayes       1         212026 - Days Na/wBayes       1         21206 - Days Na/wBayes       1         0.671 0.058 0.051 0.050 0.058 0.950 0.952 0.956 (1-0000)       0.050 1.000 0.000 /         0 b c d e < classified as       26 7 7 1 0 0   a = (1-1000)         16 139 1 0 0   i a = (1-1000)       1 0 0   i a = (1-1000)         16 139 1 0 0   i a	Test options	Classifier output								
Supplied test set       Set            • Cross-validation Folds         • 0         • 0         • 0	<ul> <li>Use training set</li> </ul>	a								Ä
Cross-validation Folds       0         O'construction Folds       0         Percentage split       % 66         More options.       0.66         More options.       0.258         Nome options.       0.268         Nome options.       0.268         Stat       0.0218         Stat       State         21222-0-supe NalweBayes       1         2122401-functions SM00       0.971       0.980       0.958       0.957       0.956       0.957       10.900         2122401-functions SM00       0.971       0.969       0.931       0.966       0.857       0.956       100001-100000)         1:2004-teres_188       0.041       0.921       0.966       0.937       0.966       0.921       100001-100000)         0:000       0.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.0000       1.0000       1.000       1.0000       1.00000       1.00000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000       1.0000 <th>O Supplied test set Set</th> <th> Summary</th> <th>ildation</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>	O Supplied test set Set	Summary	ildation							
Percentage split % 66       Incorrectly Classified Instances       31       6.889 %         More options       0.0258       0.0258         More options       0.0258       0.0258         More options       0.1411       1.100         (Nom) Vieramessiare       0.1411       1.179 %         Stat       Stop       1         212025 - Dayse NalwBayes       1         2120401 - functions SNO       0.591       0.592       0.591       0.596         2120401 - functions SNO       0.591       0.591       0.596       0.827       0.593       0.596         212004 - functions SNO       0.591       0.591       0.596       0.592       0.596       (10000-10000)         212004 - functions SNO       0.591       0.592       0.596       0.592       0.596       (10000-10000)         212004 - functions SNO       0.591       0.597       0.596       (10000-10000)       0.000       0.00	Cross-validation Folds 10	Correctly Classified In	stances 419		93.1111	•				
Kaps staintic       0.66         More options       Root mean squared error       0.0259         Root mean squared error       12.8179 8         Root mean squared error       45.511 4         Tock main squared error       45.511 4         Tock Transness       1         #== Detailed Accuracy By Class ===         212025-bays NaveBayse         212024-bays NaveBayse         212040-bay KSlar         0.631       0.610         0.632       0.973       0.956         (1-000)       0.621       0.950       0.857         212040-bays NaveBayse       1          21205-bays NaveBayse       1.000       0.000       1.000         212040-bay KSlar       0.951       0.956       0.487       (1000-10000)         1.000       0.000       7       0.000       1.000       1.000       1.000       1.000         1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000         1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000       1.000 <th>O Percentage split % 66</th> <th>Incorrectly Classified</th> <th>Instances 31</th> <th></th> <th>6.8889</th> <th>\$</th> <th></th> <th></th> <th></th> <th></th>	O Percentage split % 66	Incorrectly Classified	Instances 31		6.8889	\$				
More options.       mean squared error       0.1433         (Nom) Vieramesatare       0.1413         Stat       Stop         Z12264: Vest Jaw       45.041         Z12264: Vest Jaw       Accuracy By Class ==         212264: Vest Jaw       Total Number of Instances       1         === Detailed Accuracy By Class ==       Total Number of Instances       1         212264: Vest Jaw       Total Number of Instances       1         === Detailed Accuracy By Class ==       Total Number of Instances       1         === Detailed Accuracy By Class ==       0.931       0.956       0.957       0.956       0.957         212264: Vest Jaw       0.031       0.001       1.000       1.000       1.000       1.000       1.000       1.000         213206: Vark Star       0.931       0.075       0.931       7       0.966       0.924         === Centration Matrix ===       a       b       c       d       c= (10000)       1.00001.100000)         16 139       1       0       1       e (100001-100000)       1       0.924       0.924         State         State         A b       c       c       c.0331       (1000001-1000000) </th <th></th> <th>Kappa statistic</th> <th>0</th> <th>.86</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th>		Kappa statistic	0	.86						
Relative absolute error       12.179 %         (Nom) Vieramesatare       45.541 %         Total Number of Instances       450         Ignored Class Unknown Instances       1         === Detailed Accuracy By Class       1         21220.25 - Nyms NahwBayes       0.931 0.041 0.921 0.931 0.956 0.482 0.973 0.956 (1-1000)         2122.40 - Nundtons SMO       0.931 0.041 0.921 0.931 0.956 0.482 0.973 0.956 (1-1000)         2120.44 - twees_J48       0.931 0.001 0.021 0.000 1.000 1.000 1.000 1.000 1.000         2120.05 - NaxeKStar       0.931 0.001 0.221 0.931 0.936 0.922 0.468 (5001-5000)         0.000 0.000 7 0.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 0.000 0.000 2 0.931 0.001 1.000 1.000 0.000 0.000 2 0.932 0.456 (10001-100000)         Weighted Avg. 0.931 0.075 7 0.931 7 7 0.986 0.333 (100001-1000000)         weighted Avg. 0.931 0.075 7 0.931 7 7 0.986 0.924 (10001-1000000)         === Confusion Matrix ===         a b c d e c classified as         265 7 1 0 0   b = (10001-100000)         16 139 1 0 0   b = (100001-100000)         0 0 0 0   e = (10001-100000)         1 0 0 0 0   e = (100001-100000)         1 0 0 0 0   e = (100001-100000)	More options	Root mean squared error		.0258						
Gloon Vieramessare       Poot relative aguated error       45,541 %         Start       Stop         Procent Class Unknown Instances       1		Relative absolute error	12	.8179 %						
Stat         Stop           Total Number of Instances         450           Total Number of Instances         1           Tesutified (right.clck for options)         Detailed Accuracy By Class            212025-bayes.NaWeBayes         Detailed Accuracy By Class            2120401-tunctions SM00          Detailed Accuracy By Class	(Nom) Vleramesatare	Root relative squared e	rror 45	.541 %						
Stat         Stop         Investe Cases Guidenta Instantants         I           Result (right).cbc for options)         Detailed Accuracy By Class         Detailed Accuracy By Class         Detailed Accuracy By Class           2120.25 - bayes AlaveBayes         Detailed Accuracy By Class         Detailed Accuracy By Class         Detailed Accuracy By Class           2120.24 - trees_J48         Detailed Accuracy By Class         Detailed Accuracy By Class         Detailed Accuracy By Class           2120.25 - bayes AlaveBayes         Detailed Accuracy By Class         Detailed Accuracy By Class         Detailed Accuracy By Class           2120.25 - bayes AlaveBayes		J Total Number of Instanc	es 450							
Result (right-click for options)       Detailed Accuracy By Class         1212225 - bays NaiveBayses       Detailed Accuracy By Class         21226.44 - twees_J48	Start Stop	Ignored Class Onknown I	IIBCAILCEB	-						
212025-bayesNakeBayes       T7 Bate FP Bate Presion Becall F-Means MCC BOC Area Class Class         2120401-bundons SM00       0.931       0.961       0.931       0.962       0.932       0.931       0.962       0.934       0.931       0.962       0.934       0.931       0.961       0.934       0.931       0.961       0.952       0.934       0.931       0.962       0.934       0.931       0.961       0.951       0.965       0.957       0.954       0.954       0.934       0.951       0.960       0.953       0.957       0.954       0.954       0.951       0.966       0.1000       0.900       0.966       0.333       (100001-1000000)         Weighted Avg.       0.931       0.075       7       0.936       0.924       0.924       0.924       0.926       0.924       0.924       0.926       0.924       0.924       0.926       0.924       0.926       0.924       0.926       0.924       0.926       0.924       0.926       0.924       0.	Result list (right-click for options)	Detailed Accuracy B	y Class ===							
212401-burdtons SMO       0.971       0.098       0.971       0.071       0.986       0.812       0.973       0.986       (1-1000)         212404-burgs_J48       0.951       0.619       0.010       0.921       0.619       0.011-5000)         213205-burgkStar       0.966       0.0197       0.966       0.0197       0.986       0.812       0.971       0.0100         213205-burgkStar       0.960       0.005       0.714       0.500       0.059       0.011-5000)         0.000       0.000       1.00	21:20:25 - bayes.NaiveBayes	TP Rat	e FP Rate Precisi	on Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
2128.44.*ress_J48       0.931       0.041       0.921       0.051       0.957       0.956       0.957       (1001-5000)         12005       1.000       1.00	21:24:01 - functions.SMO	0.971	0.098 0.941	0.971	0.956	0.882	0.973	0.956	(1-1000)	
213205-laxyKStar       0.800       0.005       0.714       0.800       0.000       1.000 <th>21:26:44 - trees.J48</th> <th>0.891</th> <th>0.041 0.921</th> <th>0.891</th> <th>0.906</th> <th>0.857</th> <th>0.956</th> <th>0.897</th> <th>(1001-5000)</th> <th></th>	21:26:44 - trees.J48	0.891	0.041 0.921	0.891	0.906	0.857	0.956	0.897	(1001-5000)	
0.000         0.000         1.0000         1.000         1.000 <t< th=""><th>21:32:05 - lazy.KStar</th><th>0.500</th><th>0.005 0.714</th><th>1.000</th><th>1.000</th><th>1 000</th><th>1.000</th><th>1.000</th><th>(5001-10000)</th><th></th></t<>	21:32:05 - lazy.KStar	0.500	0.005 0.714	1.000	1.000	1 000	1.000	1.000	(5001-10000)	
Weighted Avg.         0.931         0.075         2         0.931         2         2         0.966         0.924           Confusion Matrix         a         b         c         d         c          c         a         b         c         d         a         c         d         a         b         c         d         a         a         b         c         d         a         a         a         b         c         d         a <th></th> <th>0.000</th> <th>0.000 ?</th> <th>0.000</th> <th>?</th> <th>?</th> <th>0.998</th> <th>0.333</th> <th>(100001-1000000)</th> <th></th>		0.000	0.000 ?	0.000	?	?	0.998	0.333	(100001-1000000)	
a         b         c         d         e         < classified as           269         7         1         0         i         a = (1-100)         iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii		Weighted Avg. 0.931	0.075 ?	0.931	?	2	0.966	0.924		
a b c d e <classified as<="" td="">           269 7 1 0 0   a = (1-1000)           16 139 1 0 0   b = (1000-5000)           0 5 5 0 0   c = (5001-10000)           0 0 0 6 0   d = (10001-100000)           1 0 0 0 0   e = (10001-100000)           xatus           OK</classified>		Confusion Matrix	-							
265         7         1         0         0         0         1         0         1         0         1         0         1         0         1         0         0         1         0         0         1         0         0         0         1         0         0         0         1         0		a h c d a	<pre><pre><pre><pre><pre><pre><pre><pre></pre></pre></pre></pre></pre></pre></pre></pre>							
16 139 1         0 0   b = (1001-5000)           0 5 5 0 0   c = (5001-10000)           0 0 0 6 6 0   d = (10001-100000)           1 0 0 0 0   b = (10001-100000)           0 K		269 7 1 0 0 1	a = (1-1000)							
0         5         5         0         0         - (\$001-10000)           0         0         0         0         1         - (\$001-10000)           1         0         0         0         + (\$0001-100000)           1         0         0         0         + (\$0001-100000)           1         0         0         0         + (\$0001-100000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)           1         0         0         0         + (\$0001-1000000)		16 139 1 0 0	b = (1001 - 5000)							
Status         0K         Log         Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx		0 5 5 0 0 1	c = (5001 - 10000)							
Status OK		1 0 0 0 0 0	a = (10001 - 100000 c = (100001 - 100000	) 00)						
Status OK			. (100001 10000							
Status OK Log 💉 X0										•
	Status									
									100	
	ок								Log	

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.

Algorithms	<b>Correctly Classified</b>	Карра	Mean Absolute		
Aigoritimis	Instances (%)	Statistics	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		

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

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
- Leblanc, M., Tibshirani, R., "Combining Estimates in Regression and Classification", Journal of the American Statistical Association, 1641-1650, 1994
- Abney, Steven P. "Semisupervised Learning for Computational Linguistics", London, 2008
- Ian H. Witten, Eibe Frank, Mark A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", Amsterdam, Netherlands, 2011.
- T. Hastie, R.Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition), Springer, 2009.
- 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.
- 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)
- 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