Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 5, May - 2015

Data quality in Accounting Information Systems

Comparing Several Data Mining Techniques

Erjon Zoto
Department of Statistics and Applied Informatics
Faculty of Economy, University of Tirana
Tirana, Albania
erjon.zoto@unitir.edu.al

***Abstract*—Organizations are constantly seeking for high quality data to be shared within their environment. On that purpose what they need is a proper architecture that determines the specified events affecting, evaluating and also monitoring the level of data quality in the company's data warehouse. In their job position, every employee can experience several problems regarding the quality of data he uses on a daily manner. These problems may be large enough to impede the company's staff from accomplishing the given objectives, causing considerable monetary loss. Some of the specific problems can be solved simply enough, others need a stronger effort to get over them. Instead of correcting them, experts suggest that the proper action on this case would be to understand their primary causes and then determine how to prevent them properly. This paper will try to address some of the main factors affecting data quality, by taking in consideration several methods and techniques, and applying them in the local environment. Besides showing these factors, an important part will be dedicated to the comparison between several data mining techniques used on this case.**

Keywords— data quality; performance; logistic regression; decision trees; neural networks

I. INTRODUCTION

Data quality issues have been always a major concern for business managers, and this might be explained with the following arguments.

First, managers always tend to say how necessary is for them to see the data themselves in order to better manage their own business. This necessity is growing further in the recent years, as the knowledge workers themselves believe that data is necessary to be understood in order to perform the given tasks.

Second, many organizations, as they go more and more global, seek for a highly effective integration process for their globally spread data sources.

Third, the requirements over fitting data with current standards have been increasing lately, thus using data inappropriately will not become unaware.

Speaking of standards and systems, data quality has become crucial for the accounting information systems' (AIS) success. As data processing has become a more important process inside the organization, the latter needs to address properly the issues related to the qualitative data management.

Such a management should be based on several data quality policies that get along with the factors affecting data quality, especially inside the AIS.

Data quality factors will be the main focus of this paper, and they will be evaluated through several methods, including the ones related to Artificial Intelligence.

In the beginning, the group of best factors affecting data quality in AIS will be derived from a survey made on this purpose. It comprised more than 180 professionals from the fields of Accounting and IT.

Then, the data mining techniques will be explored and the corresponding results will be interpreted. Main focus will be given to the accuracy rates when using each of the techniques.

Finally, an overall comparison between the results from the data mining techniques will serve as a starting point for future research in this topic. Results were different for each technique, even when the list of factors was reduced to less than 10.

II. DATA QUALITY PERFORMANCE INDICATORS

A. The Survey

As mentioned above, a survey was made with a large sample, chosen from the categories of professionals that are mostly related to the topic itself.

There were 182 respondents, while the total number of questionnaires spread reached 700, thus having a response rate of 26%.

It included three main sections, as mentioned below:

- General data
- Importance level
- Performance level

Throughout this paper, main focus will be on the performance section, since all necessary data from the analysis mentioned below were taken from that part.

B. The Factors

There were 24 factors identified from the literature and related to several fields, such as management (Work environment, Change management, Staff skills etc.), accounting (Internal Controls, Auditing etc.) and IT (training, AIS characteristics etc.), as in [1]. Some of the factors were associated with several sub factors, so that the respondent would understand better the factor mentioned and give a more reasonable answer.

C. The Data

Respondents were asked to answer to the preliminary questions of the first section and then all the questions related to the importance and performance level of individual factors. In most of the cases, there was the same approach of categorizing the potential answer in 5 fields, from very low to very high performance (importance).

All the fields in the questionnaire were compulsory to be filled; otherwise the questionnaire would not count at all. This enabled better response rates among the online users, whereas the printed version would ask for more time to be processed.

The results of the survey were pretty much similar with the relevant literature on the subject. Apart from evaluating between a very low and a very high level of performance, each respondent was asked also to name the three most important and then the three best performing factors, as in [2]. There were 546 total answers taken and the main results are shown in table 1 below:

TABLE I. BEST PERFORMING FACTORS, TOTAL RANKINGS

Rank	Factor	Total rankings
1	Characteristics of AIS	59
2	Internal Controls	47
3	Training	41
4	Standards and Policies related	40
5	Measurement and Reporting	40
6	Knowledge over AIS and data quality	33
7	Managerial commitment	27
8	Teamwork	26
9	AIS audit	26
10	Control over Data Quality	19
11	Change management	19
12	Strategic vision	18
13	Cost/benefit analysis	17

III. PREPROCESSING DATA

Before analyzing the results from the other techniques, the data need to be transformed for a better processing process. First, the number of categories is very high. At this moment, the evaluation has been done using the Likert scale, with 5 values ranging from “Very low” to “Very high”. The new scale will have two levels, “Low” and “High”, and it will be represented from a binary variable. A value of 1 will replace levels including “Average” up to “Very high”, while the other levels will be replaced by 0.

Another transformation is related to the number of variables to use in the following models. At the moment, there were 24 variables that have been evaluated from the respondents. The now binary variables should be reduced in a smaller number in order to be relevant.

Several methods can be pursued in order to find the best subgroup of factors. The process is called feature selection and some of the most used techniques include Principal Component Analysis (PCA), Correlation-based Feature Selection (CFS), factor analysis, and sensitivity analysis, as mentioned in [3].

All of the methods above were tested using Weka software in the available dataset, taken from the performance section of the questionnaire. In the end, the chosen one was the CFS, with its representative CfsSubsetEval, as it considers the usefulness of each feature when predicting the class label together with the correlation between them, as in [4]. The search method used was Best First, as it is a method searching for potential subsets of factors while able for a backtracking procedure, as mentioned by [5]. It can start from an empty set of factors, the full set or something in between. The direction chosen in this case was forward, with an initial empty set of factors.

The factors derived from the method chosen are mentioned below, as in [6]:

- Managerial commitment
- Training
- Strategic vision
- Standards and Policies related
- Characteristics of AIS
- Measurement and reporting
- Cost/benefit analysis

The ranking above is related to the factor number in the performance section. These factors will

be used to test each of the following Data mining techniques.

IV. Logistic Regression

In the real business world, there are many variables with two possible values. So, consumers decide on buying or not buying a product, a product may pass or fail quality control, there is high good or bad credit risk evaluation, an employee may be promoted or not etc. There are several regression techniques that can be used to analyze datasets with dependant categorical variables. When a dependant variable is categorical and all of the independent variables (or at least a part of them) are categorical, the best method to use would be logistic regression.

A. Main Concepts

Logistic regression has the ability to determine the impact of many independent variables in predicting a value among the possible outcomes of the dependent variable categories.

There are two uses for logistic regression:

- Predicting the group membership of an individual case. Results of the analysis made will be expressed as odd ratios, comparing the probability of success over failure.
- Providing knowledge over relationships between several independent variables or the relative strength for each of them.

B. Variables

The logistic regression model that will be used in this case will obviously consider explaining the dependent variable, the level of data quality in AIS, with one or more independent variables, representing the remaining factors after the preliminary analysis mentioned previously.

The seven variables are already known now, but the ranking of best factors from the logistic regression method is given below:

- Training
- Characteristics of AIS
- Cost/benefit analysis
- Measurement and reporting
- Standards and policies applied
- Strategic Vision
- Managerial staff commitment

The ranking above is based on the Wald coefficient, which is a known parameter that estimates the relevance of the specific variable in the model. The larger the Wald coefficient, the better is the variable.

C. Interpreting Results

Primary data tell that, if there is no impact from the independent variables, the accuracy of the prediction of the dependent variable's value would be 50%. Thus, if we would say that the level of data quality in AIS is high (low), we would be correct in 50% of the cases. This in fact shows a perfect division in the existing dataset between the respondents using AIS with a perceived high level of data quality and those using AIS with a perceived low level of data quality, with 91 cases for each category.

When all dependent variables are included in the model, the overall accuracy raises up to 69.2 %, as shown in table 2. In details, the table shows that the percentage of correct predictions for the cases with high level of data quality is 73.6%, which is a higher value compared to the percentage of correct predictions for the cases with low level of data quality, 64.8%. This means that the success of the event is predicted more accurately than the failure case. Else, we might conclude that the type I error rate is lower than the type II error rate.

TABLE II. Confusion Matrix - Logistic Regression

Real values		Predicted values		
		Data quality level		***Percentage Correct***
		High	***Low***	
Data quality level	***High***	67	24	73.6
	Low	32	59	64.8
Overall Percent.				69.2

Regarding the model significance, we might use the significance test based on $\chi 2$ statistics, as we are dealing with a logistic regression model.

The table below shows that the model is highly significant:

TABLE III. Model Summary

-2 LL	Cox & Snell R^2	Nagelkerke R^2	Prob
215.511	.183	.244	.000

The table shows that the Nagelkerke R2 is more than 24%, which means that the model described is able to explain 24% of the variation in the values of the dependent variable.

After showing the significance of the model, the next step is the significance and interpretation of the individual independent variables. As shown by table 3, the only significant variables are those related to the Constant value and Training:

TABLE IV. Variables In The Logistic Regression

Variable	B	Wald	Signif.	Exp(B)
Managerial	.204	.263	.608	1.226

commitment				
Training	.958	6.677	.010	2.606
Strategic Vision	.295	.535	.464	1.343
Standards and Policies	.290	.547	.460	1.337
Characteristics of AIS	.475	1.234	.267	1.608
Measurement and Reporting	.361	.663	.415	1.435
Cost/benefit Analysis	.341	.798	.372	1.407
Constant	-1.167	18.916	.000	.311

The fact that there are only two significant variables, what's more important, one out of seven factors, means that there is a more important correlation between the values of data quality in AIS and training compared with other factors.

A parameter that helps the analysis of dependent variables' impact in a logistic regression model is the Wald coefficient. In the table above, its higher values correspond again to the Constant variable and the one related to the Training factor.

When assessing the real impact of each factor in the values of the dependent variable, the values from the first and the last column are important. From the first column, the positive values bring a positive impact to the high values of data quality, while the negative value for the Constant variable shows that, with no independent variables in the model, the default value of data quality in AIS is low. Also, a value of .96 for the Training variable shows that there is a higher chance of having a high level of data quality by 96% if the value of Training is 1 (high perceived performance). The same things can be said for the other variables, while for the Constant value, it denotes that, it decreases the chance of high data quality by more than 100%.

Values from the last column are odd ratios, meaning that they show the probability for high values of the dependent variable. In details, a value of 2.6 for the Training variable shows that, when there is a high perceived value of performance for this variable, the odds for a high level of data quality is 2.6 times higher, all other values constant. At the same time, ratios between the values in the last column show the relative impact of the factors in the high level of data quality in AIS.

V. The Decision Tree

A decision tree is similar to a real tree, where each node represents a test value for an attribute, while each leaf represents a test result. The tree then tries to separate observed values into several mutually exclusive subsets.

There are several splitting techniques for decision trees.

A well-known technique for building decision trees is the C4.5 one. It uses greedy search methods, including building and pruning decision trees' structures, on the purpose of exploring all possible models.

A. Main Concepts

The C4.5 algorithm extends the variables range to categorical and numerical ones. This improvement does overestimate those attributes that split data in subsets with low class entropy, thus where most of the instances belong to one of the main classes. The algorithm then chooses as the best splitting attribute the one that offers maximum level of discrimination between classes.

It is all based in the theory of information gain, and the C4.5 algorithm chooses at each step the attribute with the highest information gain to make the subsequent split, and then continues with the attributes having second highest information gain, and so on, as shown in [7].

B. Variables

In the previous section, the logistic regression model was used to determine the ranking of the seven best factors that can contribute to the high level of data quality in AIS, according to the dataset related to the performance level in the survey. In the following section, the same factors will be analyzed through the C4.5 algorithm.

The ranking of the factors according to the C4.5 algorithm is given below:

- Training
- Characteristics of AIS
- Managerial staff commitment
- Strategic Vision
- Measurement and Reporting
- Cost/benefit analysis
- Standards and policies applied

The ranking above takes in consideration the position of each of the variables in the tree itself, from the upper levels to the lower ones. The tree is visualized below:

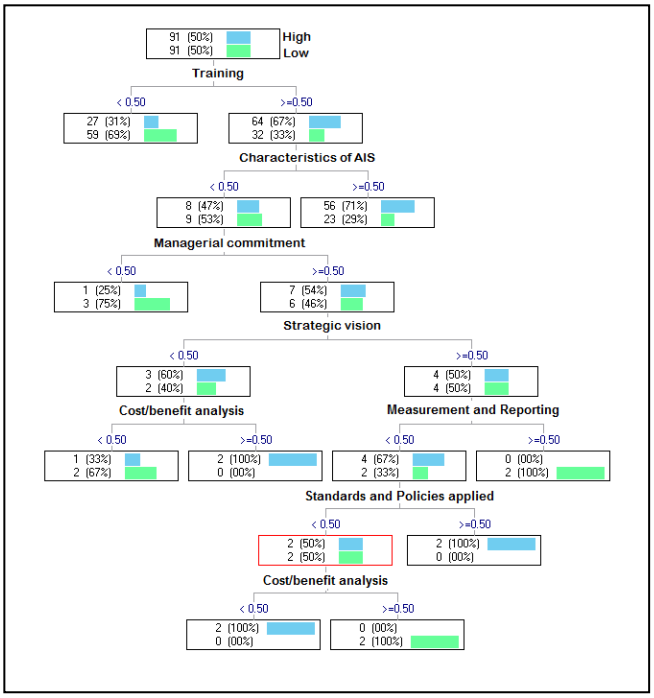

Fig. 1. *Applying C4.5 method - 7 factors*

C. *Interpreting Results*

The tree visualized above shows an overall accuracy of 71.4%, meaning that 130 out of 182 cases were predicted correctly, as shown from table 5. In details, more than 68.1% of the cases with high level of data quality were predicted accurately, while the percentage of correct predictions for a low level of data quality is at a higher rate, nearly 75%. Thus, the success of the event is predicted less accurately than the failure, which means that there is a higher value of type I error rate.

TABLE V. CONFUSION MATRIX – C4.5 ALGORITHM

Real values		Predicted values		
		Data quality level		***Percentage Correct***
		High	***Low***	
Data quality level	***High***	62	29	68.1
	Low	23	68	74.7
Overall Percent.				71.4

From fig 1, we might deduce that the best splitting attribute from the seven included is the one related to training, which is the same case with the other results so far. This factor is able to classify 67% of the successful cases while having the value of 1. Generally speaking, when the training is perceived as realized above average level, this is associated with a high level of data quality perceived in 67% of the existing cases (64 out of 96). The other factors have lower ratios. Meanwhile, a low level of perceived training is associated with a low level of perceived data quality in 69% of the existing cases (59 out of 86), which is again the best splitting ratio.

In the second level of the tree, on the right side, the next-best splitting factor is the one related to the characteristics of AIS. As shown above, when there is a high perceived level of training, a high level of perceived AIS characteristics related to data quality is associated with a high level of data quality in 71% of the cases (56 out of 79). Meanwhile, with a high level of perceived training, a low level of AIS characteristics is associated with a low level of data quality in 53% of remaining cases (9 out of 16).

The third best splitting factor is related to the managerial commitment. With a high level of perceived training and a low perceived level of AIS characteristics, a high perceived level of managerial commitment is associated with a high level of perceived data quality in 54% of the cases (7 out of 13), while a low level of perceived managerial commitment is associated with a low level of perceived data quality in 75% of the remaining cases (3 out of 4).

The same analysis can be done with the remaining attributes of the decision tree.

VI. NEURAL NETWORKS

Their architecture is based on the characteristics of a single node and the characteristics of its link with the network as a whole. The main characteristics of a node are the subset of the nodes linking to it, a summing tool and the activation function.

Typically, neural network architecture is determined from the number of inputs inside the network, the number of outputs, total number of initial nodes and the linking between them. Neural networks are usually classified inside two categories, based on the type of linkage between nodes: feedforward and recurrent (cyclical).

The network is said to be feedforward when data processing begins from the input side towards the output side continuously, without cycles or returning back. A recurrent network includes at least a node that transmits input data following a circular path inside the network, usually serving for feedback purposes.

A. *Main Concepts*

Multilayer feedforward networks are one of the most important and most popular classes of neural networks, used in real-life problems. Typically, such a network consists of several inputs in the input layer, one or more hidden layers of computing nodes and then the output layer of these nodes. Data processing is done from the input side, through the hidden layers, towards the output side. This type of network is more commonly known as a multilayer perceptron (MLP), denoting a more general case than a single perceptron, which includes a single layer.

MLPs have had a wide success in solving multiple problems, training the network in a supervised manner with a well known algorithm, entitled as

backpropagation algorithm with fault tolerance. This algorithm is based on two phases, a forward step and a backward step.

In the first step, a trained sample is applied in the input layer, and the subsequent effect is transmitted layer after layer producing a set of resulting output nodes. The backward step rearranges the specific weights for the nodes, thus reaching a better result, which is nearer to the real values.

B. Variables

Using the same dataset as the previous techniques explored, the MLP gives the following results related to the factors rankings:

- Training
- Standards and policies applied
- Cost/benefit analysis
- Strategic Vision
- Characteristics of AIS
- Measurement and Reporting
- Managerial staff commitment

The ranking above was based on the accuracy rate developed from each factor using the MLP technique.

C. Interpreting Results

The MLP technique shows an overall accuracy of 79.1%, meaning that 144 out of 182 cases were predicted correctly, as shown from table 6. In details, more than 79.1% of the cases with high level of data quality were predicted accurately, and the same rate goes for the percentage of correct predictions for a low level of data quality. Thus, the success of the event is predicted as accurately as the failure, which means that the values of type I and type II error rates are the same, around 21% each.

TABLE VI. CONFUSION MATRIX – MLP TECHNIQUE

Real values		Predicted values		
		Data quality level		***Percentage Correct***
		High	***Low***	
Data quality level	***High***	72	19	79.1
	Low	19	72	79.1
Overall Percent.				79.1

This model is composed of 7 input nodes (the factors), 2 output nodes (the classes) and a hidden layer composed of 6 intermediate nodes, which results to be the most accurate model regarding the MLP technique.

From the ranking above, we might deduce that the best predicting attribute from the ones included is related to training, and this is what we have concluded from the other results as well. When this factor is not included in the list of variables predicting the value of data quality level, the model is able to classify only 70.88% of the total cases accurately, 129 out of 182 overall.

The second-best predicting attribute, according to the MLP technique used, is related to the standards and policies applied inside the organization regarding data quality. This result is not similar with results from other techniques explored so far, but it resembles to the ranking of best performing factors, from the total rankings in Table 1. When this factor is not included in the list of seven factors, the model is able to classify only 71.43% of the total cases accurately, 130 out of 182 cases overall.

The third-best predicting attribute is related to the cost/benefit analysis done over the data quality. This result is the best ranking position for this factor, equaling the results taken from the logistic regression, where the Wald coefficient ranks this factor in the same position. When the cost/benefit analysis is not included in the list of attributes predicting the level of data quality, the model is able to classify only 130 out of 182 cases overall, or 71.43% of the total cases accurately. There are identical values between the latter two factors, but the one related to policies and standards yielded better results in continuous tests compared to the cost/benefit analysis.

The remaining attributes give a lower predicting power to the model raised with the MLP technique, thus the three attributes mentioned above are more important in this case and were explained in more details.

VII. COMPARING RESULTS

In the sections above, the author has found the best factors that can predict more accurately the level of data quality in AIS. There were several methods applied and each one of them showed different results, as shown in the table below

TABLE VII. SUMMARY OF THE RESULTS

Factor	Factor ranking		
	L. R.	***C4.5***	***MLP***
Managerial commitment	7	3	7
Training	1	1	1
Strategic Vision	6	4	4
Standards and Policies	5	7	2
Characteristics of AIS	2	2	5

Factor	Factor ranking		
	L. R.	***C4.5***	***MLP***
Measurement and Reporting	4	5	6
Cost/benefit Analysis	3	6	3

In the following part, the results above will be compared between the techniques used. Since the factors are the same between all techniques, the analysis will focus on the differences between the factor rankings in each of the techniques.

A. Testing the Differences

Once again, the analysis will focus on the differences between the factors' rankings in each pair of the techniques used and then it will determine whether this difference is statistically significant or not. For this purpose, a nonparametric method will be used that finds the rank correlation between the groups of the factors.

This method defines whether in two different situations there do exist or not statistically significant differences between the rankings of the factors included in the two groups. Differences will be evaluated based on the difference values between the new position and the old one in the ranked list of factors.

B. The Rank Correlation Method

The basic hypothesis is related to the assumption that there is no rank correlation between the studied techniques, or its value is not statistically significant. The alternative hypothesis is related to the assumption that there is a significant correlation between the techniques, which affects also in a similar factor ranking between them.

After determining the hypothesis, the next step is determining the value of the rank correlation coefficient, based on the formula below:

$$r_s = 1 - {6\sum(x_i - y_i)^2}\Big/{n(n^2-1)}, \text{ where} \qquad (1)$$

x_i is the rank of the *i*-th factor from the first method, y_i is the rank of the *i*-th factor in the second method and n is the number of factors in each of the methods.

Assuming a quasi normal distribution, the values of the average and standard deviation for the correlation coefficient can be evaluated accordingly with the whole population values, as follows:

$$\mu_{r_s} = 0 \qquad (2)$$

$$\sigma_{r_s} = \sqrt{\frac{1}{n-1}} \qquad (3)$$

The last step is comparing the value of observed z against the critical one, according to the specific criteria. The value of observed z is defined as follows:

$$z = \frac{r-\mu}{\sigma} \qquad (4)$$

A higher value of the observed *z* compared to the critical one denotes that the alternative hypothesis is true, thus there is a significant correlation between the rankings of the techniques involved. A lower value of the observed *z* denotes that the basic hypothesis is true.

C. Final Results

The first techniques that will be compared will be the logistic regression and the C4.5 algorithm. Thus, the test on this case will determine whether there does exist or not a statistically significant difference in the factors' rankings between the methods. From table 7 the needed values may be derived as follows:

TABLE VIII. DIFFERENCE BETWEEN RANKS - L.R. VS C4.5

Factor	x_i	y_i	$(x_i - y_i)^2$
Managerial commitment	7	3	16
Training	1	1	0
Strategic Vision	6	4	4
Standards and Policies	5	7	4
Characteristics of AIS	2	2	0
Measurement and Reporting	4	5	1
Cost/benefit Analysis	3	6	9
Cost/benefit Analysis	3	6	9

After substituting the values in the formulas above, more specifically from eq. (1), the rank correlation coefficient is found to be 0.39. The positive value shows that there is a potential positive relationship between the ranks in both techniques, thus a certain correlation exists between them. The observed value of the correlation is yet not large enough to predict a statistically significant value.

The standard deviation and observed *z* are given below:

$$\sigma_{r_s} = \sqrt{\frac{1}{7-1}} = \sqrt{\frac{1}{6}} = 0.41$$

$$z = \frac{r-\mu}{\sigma} = \frac{0.39}{0.41} = 0.96$$

The observed *z* is lower than the critical value of 1.96, thus the available data are not enough to reject the null hypothesis, which means that the techniques studied have no significant correlation between them and the rankings of the factors are independent between both techniques.

The next step is determining the correlation coefficient between the logistic regression and the

MLP techniques. Using the values from table 7 and substituting in eq. (1), the rank correlation coefficient is equal to 0.54. The positive value in this case is higher than in the first case, which may denote a stronger positive relationship between the ranking of the factors in both techniques. Yet, such a value should be tested accordingly.

With the same value for the standard deviation, the observed z is calculated to be around 1.31.

This value is still lower than the critical value of 1.96, thus again the available data are not enough to reject the null hypothesis. There is no statistically significant correlation between the rankings of the factors from the logistic regression and the MLP technique.

The last comparison includes the data from the C4.5 and MLP techniques.

The rank correlation coefficient is found to be -0.07. This is a negative value, which may denote a negative correlation between the ranks of the factors. Obviously, such a value is too low to be significant and this will be shown clearly from the statistical test.

The observed z is calculated to be around -0.17, which is larger enough than the critical value of -1.96 so that the null hypothesis will not be rejected. There is no statistically significant negative correlation between the rankings of the factors from C4.5 and MLP techniques.

Summarizing, all tests show that there is no statistically significant correlation between all techniques. This result was somewhat expected, since each of the techniques has its own way of determining the relevance and impact of the various factors included in the model.

The results show that there can be made no associations between the ranks of the factors, thus each of the techniques has its own rank of factors and they are independent from one another.

REFERENCES

[1] H. Xu, “Critical Success Factors for Accounting Information Systems Data Quality”, 2003

[2] E. Zoto, “Data quality and Accounting Information Systems: Actual performance in Albania”, PIEB 2014, vol. 14, issue 1, Prague, unpublished

[3] M. Omid, A. Mahmoudi, and M. H. Omid, “Development of pistachio sorting system using principal component analysis (PCA) assisted artificial neural network (ANN) of impact acoustics”. Expert Systems with Applications, 37: 7205-7212, 2010.

[4] M. A. Hall, “Correlation-based feature selection for machine learning”, The University of Waikato, 1999.

[5] I. H. Witten, E. Frank, “Data Mining: Practical machine learning tools and techniques”, Morgan Kaufmann, 2005.

[6] E. Zoto, Dh. Tole, “Analyzing data quality in Accounting Information Systems”, International Economy & Business Doctoral Students Conference, pp. 319-328, December 2014, Tirana, in press

[7] Z. Xiaoliang, W. Jian, Y. Hongcan, and W. Shangzhuo “Research and Application of the improved Algorithm C4.5 on Decision Tree”, International Conference on Test and Measurement (ICTM), Vol. 2, pp. 184-187, 2009.