International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 3584~3590
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp3584-3590

Performance evaluation of different classification techniques using different datasets

Abdulkadir Özdemir, Uğur Yavuz, Fares Abdulhafidh Dael
Department of Management Information Systems, Ataturk University, Turkey

Article Info

Article history:

Received Nov 27, 2018
Revised Apr 5, 2019
Accepted Apr 14, 2019

Keywords:

BayesNet
Classification
Data mining
J48
Multilayer perceptron
NaiveBayes
Weka

ABSTRACT

Nowadays data mining become one of the technologies that paly major effect on business intelligence. However, to be able to use the data mining outcome the user should go through many processes such as classified data. Classification of data is processing data and organize them in specific categorize to be use in most effective and efficient use. In data mining one technique is not applicable to be applied to all the datasets. Many data users wasting a lot of time trying many classification techniques in order to find the most an appropriate technique to be used. This paper showing the difference result of applying different techniques on the same data. This paper evaluates the performance of different classification techniques using different datasets. In this study four data classification techniques have chosen. They are as follow, BayesNet, NaiveBayes, Multilayer perceptron and J48. The selected data classification techniques performance tested under two parameters, the time taken to build the model of the dataset and the percentage of accuracy to classify the dataset in the correct classification. The experiments are carried out using Weka 3.8 software. The results in the paper demonstrate that the efficiency of Multilayer Perceptron classifier in overall the best accuracy performance to classify the instances, and NaiveBayes classifiers were the worst outcome of accuracy to classifying the instance for each dataset.

Corresponding Author:

Fares Abdulhafidh Dael,
Department of Management Information Systems,
Ataturk University,
Üniversite Mahallesi, Atatürk Üniversitesi Kampüsü, 25030, Erzurum, Turkey.
Email: faresalariqi@gmail.com

1. INTRODUCTION

Data mining is way of extract useful knowledge out of huge volume of data. The discovering knowledge come through many mining to be useful knowledge. To convert the raw data to knowledge the data should be integrated, cleaned, classified or clustered and so on. To utilize the process of extracting data mining, many techniques and standards should be used. One of the process of data mining is how to classify the data and organize them in the correct categories [1, 2]. Every data has its own characteristics, some of them nominal data and other are numerical data and so forth. According to the data characteristics the data should be classified. However, it is not resealable to be check the data contain line by line to classify the data. In data mining there are several techniques to be used to classify the data. However not all the techniques can classify correctly the data give the same result or outcome of the data. Every technique has its own model and algorithm and the way of how to classify the data [3-6].

To find out the differences of the classification techniques and the reasons of the differences, four classification techniques have selected. They are as follow BayesNet, NaiveBayes, Multilayer perceptron and J48. To test the differences of the four selected classification techniques, three different datasets have collected, Congressional voting records, Car evaluation and Contraceptive method choice. To measure the effectiveness of the techniques two parameters have tested, the time taking of each technique to build the model for the selected dataset, and the accuracy of classifying the data by each technique. Weka software has selected as platform of applying the classification techniques and the datasets.

The paper's flow is organized as follows. Section I as introduction of the paper. Section II covers literature review of data mining and classification techniques. As well about the Weka software and methodology. In the section III Results and Discussion were illustrated. In the last section IV Conclusion and summarizing the comparative results.

2. BACKGROUND

2.1. Data mining

Data mining is a processing of the raw data to get of the useful information, or to discover the knowledge from huge databases. The output of the data mining is the pattern which is to identify potentially useful, valid, ultimately understandable and novel pattern in the mining data. Mostly data mining applying in business, so the companies can make effective marketing strategies by knowing what their customers want to buy. Data mining outcome depends on the way of collecting data and how the data are processed [1, 3, 7].

2.2. Data warehousing

Data Warehousing is a center of many data collection from many places. The companies or sectors collect their data from different places and branches in one place called data warehousing. The data in data warehousing are integrated from all places, cleaned from missing data and noise data. The data in data warehousing are organized and prepared for future use or in demand from the users. The data warehouse used to support the decision of management making process [1].

2.3. Classification

Classification in data mining are some techniques use to predict, classify and organized the data in their suitable categories [8]. Each class has its own rules and algorithms. Some of the classification techniques are follow decision tree rules such as J48, some other classes are following Bayesian Network such as BayesNet and NaiveBayes, and other are following Artificial intelligence and Neural Network. Classification techniques have different applications and which dataset should be applied on. In addition, all classification techniques will not be able to predict correctly the classification of data compare to other classification techniques. To find out the best classification techniques for the testing data, the data should be a compatible with the selected classification technique rules, algorithms etc [9].

2.4. J48

J48 classifier is an optimize version of C4.5. The J48 is based on Decision tree. J48 is one of the data classification techniques used in data processing and data mining. The J48 rules and algorithms are using decision tree techniques which contains of main leaf and branches. Each of the branch or leaf contain a decision that lead to different outcome. Some of the datasets have very big tree model which contains many leafs leads to different result comparing to few leafs of decision tree when applying J48 classification technique [10, 11].

2.5. Multilayer perceptron

Multilayer Perceptron classifier is based on Artificial intelligence and Neural Network without qualification. A Multi-Layer Perceptron (MLP) has as minimum as three layers. One layer as input, the second as hidden layer and the last as the output layer. MLP is a feedforward neural network, the hidden layer can be one layer or more. In MLP each node in each layer are connected to all layer's nodes. Multilayer perceptron is one of the data classification techniques used in neural network, deep learning and other applications of data processing [12].

2.6. BayesNet

BayesNet classifier one of the classifiers in Weka software. The BayesNet is based on Bayesian Network which is based on Bayes theorem. The Bayesian network is mostly working when there might be a probability of uncertainty, or complexity and (even more importantly) causality situation. Bayesian network consist of two parts: A set of conditional probability distributions and a directed acyclic graph DAG. Bayesian networks, each node represents a Variable. A variable might be discrete or might be continuous. BayesNet classifier one of the data classification techniques applied in many areas of probability or uncertainty conditions [13].

2.7. NaiveBayes

NaiveBayes classifiers is a collection of algorithms that share common principles based on Bayes Theorem. In NaïveBayes classifiers each pair of features classified is independent from other pairs. NaïveBayes classifiers is one of the data classification techniques used in Weka software or can be used in other areas of processing data using different software [14, 15].

2.8. Data mining process

The data mining process breaks in many stages. The first stage it's called the integration stage which is collect the data from many sources as raw data with different format. The second stage it's called data cleaning in this stage after receiving the data from the first stage some of the data are incompatible or inconsistency and other data are missing value and other data are illogical entered. So, it the data cleaning stage will clean all these data. The third stage of data mining it is to collect the cleaning data in one place call Data Warehousing. In the data warehousing mostly, the data ready to be used and analyzed. However, the amount of the data in data warehousing is huge size of data to deal with it and analyze the whole data at once, for this reason next stage is presented. The fourth stage is the selection stage, which is to select the relevant data from data warehousing that will work on it. The last stage of data mining is applying the algorithms and techniques of data mining to get the pattern, that the user looking for. The outcome of the applying data mining techniques will be represented as graph or table or other format of output representation [1, 16].

3. RELATED WORKS

There are various relative studies of the different classification techniques, yet it has not been discovered that one single method is superior compared to others. Issues like accuracy, training time, scalability and many others contribute to choosing the best technique to classify data for mining. The search for best technique for classification remains a research subject. Classification is a data mining technique used to predict group membership for data instances. There are numerous traditional classification methods like decision tree (DT) induction, k-nearest neighbor classifier, Bayesian networks, support vector machines, rule-based classification, case-based reasoning, genetic algorithm, fuzzy logic techniques, rough set approach and others. The basic difference between the algorithms depends on whether they are lazy learners or eager learners [17].

A predictive KNIME model was developed and three data mining algorithms; the Naïve Bayes, PNN Predictor and Decision Tree were trained using 70% of the total samples which were randomly selected. The knowledge acquired from the training was applied in predicting the type of supply that produced the remaining 30% of the motor operational data samples. The predictive accuracy achieved in the paper is indicative of the suitability of data mining approach for motor performance monitoring [18].

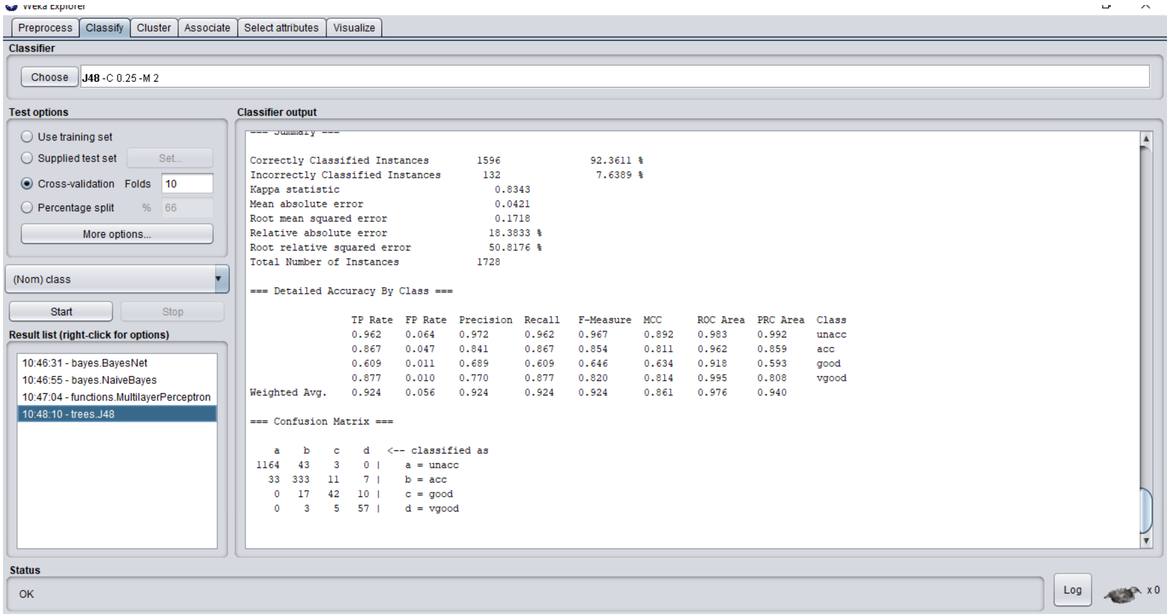
In [8], the author points out about Decision Tree (DT) or J48, that advantages of DT are easy to understand, easy to generate and reduce problem capacity. The limitations of DT are: Required separate test set, training time is so expensive, does not handle continuous variable and suffer from overfitting. The applications that fit DT are: Text Categorization and Image Classification.

In [19], MSSQL 2005 database was utilized to gather through surveys or Internet and to store information ordered under 31 criteria in four main groups contains a total of 100 students receiving vocational training in various energy application fields, who are also in the process of vocational guidance. This paper applied algorithms used in many classification techniques to a group of individuals who are in the process of vocational guidance and concluded that the most appropriate algorithm to be used for studies in this area is the Naive Bayes algorithm derived from a statistical estimation model that is called the Bayes' theorem. Since using machine learning (ML) techniques in classification studies results in accurate outcomes with a significant saving in terms of time and cost, it is recommended to make use of those algorithms used in data mining and machine learning techniques for the software to be developed in this field.

In [20], the data set used in this research is the training data set of the KDD Cup 2009 orange small data set. The data set contains 190 numeric features and 40 nominal features. Out of these 190 numeric features, 16 are empty and 132 are sparse with higher than 90% missing rate. The authors use four classification technigues, J48, NaiveBayes, SVM and KNN. The authors pointed, proposed feature selection method resolves the real-world CRM classification problems with noisy and highly imbalanced data set. The various classifiers are used for classification. As a result, the SVM has highest accuracy and sensitivity, Naïve Bayes has highest ROC and Specificity.

4. BACKGROUND
4.1. WEKA

The Weka software is a machine-learning platform for applying machine learning. Weka is abbreviation of Waikato Environment for Knowledge Analysis (WEKA). The Weka's name also refers to name of bird in New Zealand. Weka is machine learning which its collection of machine learning algorithms and standards for processing data mining. In Weka the algorithms and techniques can be applied from input file such as Excel files, Java format and others or can be applied directly from the software itself. As shows in Figure 1. The Weka Explorer window divided in to 6 tabs, each tab has different tasks. Tab 1 calls Preprocess: Per process's function is to load a dataset from different sources and manipulate the data into a desire form. Tab 2 calls Classify: is to select a classifier to process the dataset that has selected in preprocess stage. Tab 3 calls Cluster: is to select a cluster to process the dataset that has selected in preprocess stage. Tab 4 calls Associate: is to run association algorithms rules to extract insights from dataset. Tab 5 calls Select Attributes: it to run attribute selection algorithms on dataset to select those attributes that are relevant to the desire feature to predict. Tab 6 calls Visualize: is to visualize the relationship between attributes [21].



J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

10:46:31 - bayes.BayesNet
10:46:55 - bayes.NaiveBayes
10:47:04 - functions.MultilayerPerceptron
10:48:10 - trees.J48

```
=== Summary ===

Correctly Classified Instances        1596               92.3611 %
Incorrectly Classified Instances       132                7.6389 %
Kappa statistic                          0.8343
Mean absolute error                      0.0421
Root mean squared error                  0.1718
Relative absolute error                 18.3833 %
Root relative squared error             50.8176 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.064    0.972      0.962    0.967      0.892    0.983     0.992     unacc
                 0.867    0.047    0.841      0.867    0.854      0.811    0.962     0.859     acc
                 0.609    0.011    0.689      0.609    0.646      0.634    0.918     0.593     good
                 0.877    0.010    0.770      0.877    0.820      0.814    0.995     0.808     vgood
Weighted Avg.    0.924    0.056    0.924      0.924    0.924      0.861    0.976     0.940

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1164   43    3    0 |    a = unacc
   33  333   11    7 |    b = acc
    0   17   42   10 |    c = good
    0    3    5   57 |    d = vgood
```

OK

Figure 1. Weka interface explorer

4.2. Datasets information

For this paper, three data set have selected. Every data set has its own characteristics and the parameters that differentiate it from other two data sets. The Table 1 illustrate the differences between each of the dataset. The datasets have tested one by one with same settings in Weka software for all datasets. Each one of datasets has tested against the four classifiers. The output determines by the parameters of the time taken to build the model of the dataset and the percentage of accuracy to classify the target datasets. Each dataset has tested four times for the four classifiers.

Table 1. Illustrate the differences between each of the dataset

Name of datasets/ Parameters	Congressional Voting Records	Car Evaluation	Contraceptive Method Choice
Data Set Characteristics	Multivariate	Multivariate	Multivariate
Attribute Characteristics	Categorical	Categorical	Categorical
Associated Tasks	Classification	Classification	Classification
Number of Instances	435	1728	1473
Number of Attributes	16	6	9
Missing Values	Yes	No	No

5. RESULTS AND ANALYSIS

A comparison of evaluation performance of classifiers for different datasets based on the accuracy of each classifier and time taken to build the model. Accuracy is defined as the number of instances classified correctly. The Table 2 summarize the output of the classification data techniques for the three datasets based on the time taken to build the model. It is observed for the first dataset of Car Evaluation the J48 classifier give the best result of the time taken to build the model. In the second and the third datasets Contraceptive Method Choice and Congressional Voting Records respectively shows NaiveBayes classifiers give the best outcome. However, the Multilayer Perceptron classifier is the longest time taken to build the model for each dataset.

Table 2. Comparison of time taken for various classifiers

Name of datasets/ classification techniques	Congressional Voting Records	Car Evaluation	Contraceptive Method Choice
BayesNet	0 Sec	0.03 Sec	0.03 Sec
NaiveBayes	0 Sec	0.03 Sec	0 Sec
Multilayer Perceptron	0.69 Sec	4.49 Sec	3.56 Sec
J48	0.03 Sec	0.01 Sec	0.09 Sec

In the Table 3 show the Comparison of Accuracy of classifiers for different datasets. From Figure 2 and Table 2 It is observed that, for the first and the second dataset of Car Evaluation and Contraceptive Method Choice respectively, the Multilayer Perceptron classifier give the best result of the Accuracy compare to other classifiers. In the third datasets Congressional Voting Records shows J48 classifier give the best outcome of accuracy to classifying the instance. However, the NaiveBayes classifiers were the worst outcome of accuracy to classifying the instance for each dataset.

Table 3. Comparison of time taken for various classifiers

Name of datasets/ classification techniques	Congressional Voting Records	Car Evaluation	Contraceptive Method Choice
BayesNet	90.1149 %	85.706 %	51.0523 %
NaiveBayes	90.1149%	85.5324 %	50.7807 %
Multilayer Perceptron	94.7126%	99.537 %	52.3422 %
J48	96.3218%	92.3611 %	52.1385 %

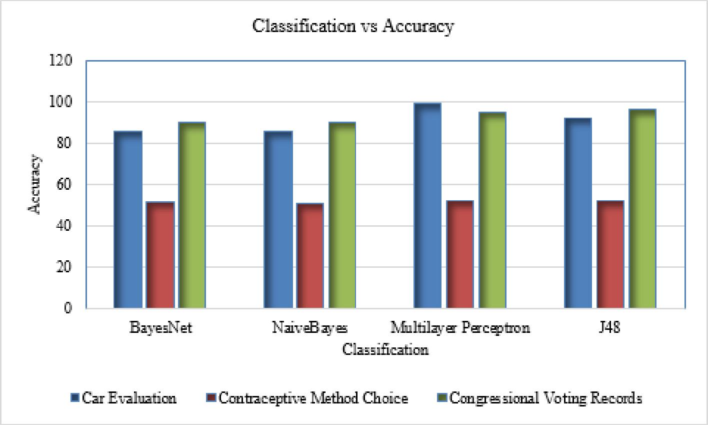


Figure 2. Graphical view of accuracy for different classifiers on different datasets

6. CONCLUSION

This paper showed the performance evaluation of different data classifiers techniques on different datasets. It found that the outcome of the data tested are different from dataset to another. There are reasons for the different output because the datasets chrematistics are different from each another dataset. Factors that may affect the classifier's performance as follow 1. Data set, 2. Number of instance and attributes, 3. Compatibility of the data with the classifier, 4. Type of attributes, 5. Missing data and data instructions, 6.

System configuration. Multilayer Perceptron classifier shows in overall the best accuracy performance to classify the instances, and NaiveBayes classifiers were the worst outcome of accuracy to classifying the instance for each dataset. Future work may focus on specific datasets that working in harmony with classifiers should be selected. The future work may focus on improving the performance of each classifiers by analyzing their algorithms and the rules.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Weka software developers and make it as an open source software without it our work would be much difficult to achieve the outcome of this paper. We also would like to thank the creator of “Contraceptive Method Choice” datasets Mr. Tjen-Sien Lim (limt@stat.wisc.edu) and the Donor: Tjen-Sien Lim (limt@stat.wisc.edu) with respect to the original source of the 1987 National Indonesia Contraceptive Prevalence Survey for sharing the datasets and makes it available for everyone. We also would like to thank the creator of “Car Evaluation Databases” datasets Mr. Marko Bohanec and the Donors: Marko Bohanec (marko.bohanec@ijs.si), and Blaz Zupan (blaz.zupan@ijs.si) for sharing the datasets and makes it available for everyone.

REFERENCES

[1] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann, 2011.

[2] Saouabi Mohamed, Abdellah Ezzati, "A data mining process using classification techniques for employability prediction," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 14, no. 2, pp. 1025-1029, 2019.

[3] Akibu Mahmoud Abdullahi, Mokhairi Makhtar, Suhailan Safie, "The patterns of accessing learning management system among students," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 13, no. 1, pp. 15-21, 2019.

[4] Longadge R., Dongre S., Latesh Malik, "Class imbalance problem in data mining: review," *International Journal of Computer Science and Network (IJCSN)*, vol. 2, no. 1, Feb 2013.

[5] Almohammedi Akram A., Nor K. Noordin and Sabri Saeed, "Evaluating the Impact of Transmission Range on the Performance of VANET," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 6, no. 2, pp. 800-809, 2016.

[6] Jabbar Waheb A., *et al*. "Design and Implementation of IoT-Based Automation System for Smart Home," 2018 *International Symposium on Networks, Computers and Communications (ISNCC), IEEE,* 2018.

[7] Almohammedi Akram A., *et al.*, "An accurate performance analysis of hybrid efficient and reliable MAC protocol in VANET under non-saturated conditions," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 7, no. 2, pp. 999-1011, 2017.

[8] Mustakim Mustakim, Novia Kumala Sari, Jasril Jasril, Ismu Kusumanto, Nurul Gayatri Indah Reza, "Eigenvalue of Analytic Hierarchy Process as The Determinant for Class Target on Classification Algorithm," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 12, no. 3, pp. 1257-1264, 2018.

[9] Deulkar M. D. S. and Deshmukh R. R., "Data Mining Classification," *Imperial Journal of Interdisciplinary Research*, vol. 2, no. 4, 2016.

[10] Kaur G., Chhabra A., "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.

[11] Koturwar Praful, Sheetal Girase and Debajyoti Mukhopadhyay, "A survey of classification techniques in the area of big data," *IJAFRC,* vol. 1, no. 11, Nov 2014.

[12] Gardner M. W. and Dorling S. R., "Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627-2636, 1998.

[13] Vaithiyanathan V., Rajeswari K., Tajane K. and Pitale R., "Comparison of different classification techniques using different datasets," *International Journal of Advances in Engineering & Technology*, vol. 6, no. 2, pp. 764, 2013.

[14] Sharma A. and Kaur B., "A research review on comparative analysis of data mining tools, techniques and parameters," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 7, 2017.

[15] Nur’Ain Maulat Samsudin, Cik Feresa binti Mohd Foozy, Nabilah Alias, Palaniappan Shamala, Nur Fadzilah Othman, Wan Isni Sofiah Wan Din, "Youtube spam detection framework using naïve bayes and logistic regression," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 14, no. 3, pp. 1508-1517, 2019.

[16] Alalwan S. A. D., "Diabetic analytics: Proposed conceptual data mining approaches in type 2 diabetes dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS).* vol. 14, no. 1, pp. 92-99, 2019.

[17] K. Saranya, "Survey on Classification Techniques Used in Data Mining and their Recent Advancements, *IJSETR*, vol. 3, pp. 2380-2385, 2014.

[18] Adekitan A. I., Adewale Adeyinka and Alashiri Olaitan, "Determining the operational status of a Three Phase Induction Motor using a predictive data mining model, *International Journal of Power Electronics and Drive Systems (IJPEDS),* vol. 10, no. 1, pp. 93-103, 2019.

[19] Bulbul Halil Ibrahim and Özkan Unsal, "Comparison of classification techniques used in machine learning as applied on vocational guidance data," *2011 10th International Conference on Machine Learning and Applications and Workshops,* IEEE, 2011, vol. 2.

[20] Natchiar S. Ummugulthum and S. Baulkani, "Customer relationship management classification using data mining techniques," *2014 International Conference on Science Engineering and Management Research (ICSEMR),* IEEE, 2014.

[21] Jagtap S. B., "Census Data mining and data analysis using WEKA," *International Conference in Emerging Trends in Science, Technology and Management-2013*, Singapore, 2013.

BIOGRAPHIES OF AUTHORS

Assoc. Prof. Dr Abdulkadir ÖZDEMIR, Lecturer at Ataturk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Turkey. Area of research interests Computer Hardware, Virtualization and Cloud Computing, Computer Networks and Management, Database Management Systems, Management Information Systems. https://www.atauni.edu.tr/abdulkadir-ozdemir, abdulkadir@atauni.edu.tr, / +90 442 231 1961.

Prof. Dr Uğur YAVUZ, Lecturer at Ataturk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Turkey. Area of research interests Information Technologies, Management Science, Web Design, Management Information Systems, Graphic Design and Animation, Artificial Intelligence. http://www.atauni.edu.tr/#personel=ugur-yavuz, ugur@atauni.edu.tr / +90 442 231 5752.

Fares Abdulhafidh Dael, PhD candidate at Ataturk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, master's degree in communication and computer Engineering from UKM University, Malaysia. Bachelor's degree in communication engineering from IIUM university Malaysia. Area of research interests Management Information Systems, Artificial intelligence, Radio communications Wireless communication, Data Network and Computer. faresalariqi@gmail.com