Scientific Journal of Informatics
Vol. 5, No. 2, Nov 2018
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

The Comparison Combination of Naïve Bayes Classification Algorithm with Fuzzy C-Means and K-Means for Determining Beef Cattle Quality in Semarang Regency

Feroza Rosalina Devi[1], Endang Sugiharti[2], Riza Arifudin[3]

[1,2,3]Computer Science Departement, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia
Email: [1]ferozarosalina24@gmail,com, [2]endangsugiharti@mail.unnes.ac.id, [3]riza.arifudin@gmail.com

Abstract

The beef cattle quality certainly affects the quality of meat to be consumed. This research performs data processing to do the classification of beef cattle quality. The data used are 196 data record taken from data in 2016 and 2017. The data have 3 variables for determining the quality of beef cattle in Semarang regency namely age (month), Weight (Kg), and Body Condition Score (BCS) . In this research, used the combination of Naïve Bayes Classification and Fuzzy C-Means algorithm also Naïve Bayes Classification and K-Means. After doing the combinations, then conducted analysis of the results of which type of combination that has a high accuracy. The results of this research indicate that the accuracy of combination Naïve Bayes Classification and K-Means has a higher accuracy than the combination of Naïve Bayes Classification and Fuzzy C-Means. This can be seen from the combination accuracy of Fuzzy C-Means algorithm and Naïve Bayes Classifier of 96,67 while combination of K Means Clustering and Naïve Bayes Classifier algorithm is 98,33%, so it can be concluded that combination of K Means Clustering algorithm and Naïve Bayes Classifier is more recommended for determining the quality of beef cattle in Semarang regency.

Keywords: Beef Cattle Quality, Combination, Naïve Bayes Classification, Fuzzy C-Means, K-Means

1. INTRODUCTION

Currently the concept of data mining is increasingly recognized as an important tool in information management because of the increasing amount of information. One of the data mining techniques is clustering [1]. Clustering is an unattended classification and is a process of partitioning a set of data objects from one set into several appropriate classes or clusters [2]. There are several methods used for grouping, including K-Means, possibilistic C-Means (PCM) and Fuzzy C-Means (FCM) [3].

Fuzzy C-Means algorithm is a data clustering technique where the existence of each data point of a cluster is determined by the membership value [4]. While K-Means is a data clustering method that partitions data into clusters / groups [5].

The use of the K-Means algorithm is limited to numerical data only [6]. One of the classification algorithm is Naive Bayesian Classification (NBC). NBC algorithm aims to classify data in a particular class. For classifier work is measured by predictive accuracy [7]. Manuscripts arranged with the following order of topics.

Beef cattle breeding is an activity that is not foreign to the broader community in Indonesia [8]. Beef cattle are a special kind of cow that is kept to be fattened because of its characteristics, such as the growth rate and meat quality is quite good [9]. Department of Agriculture, Fisheries and Food is one of the offices that located in Semarang regency which has one of the programs that is always to record and observe the good development and quality of cattle in various regions in Semarang regency. Some components affect the quality of beef cattle are age, weight, and BCS (Body Condition Score) [10].

Based on the problem that is still used manual observation in determining the quality of beef cattle, hence required a data processing able for determine the quality of beef cattle with more effective and efficient. Therefore, the authors try to combine Naïve Bayes Classification with Fuzzy C-Means and K-Means for determine the quality of beef cattle and compare the accuracy of the two combined application methods in order to obtain more recommended combination for determining the quality of beef cattle in Semarang regency.

2. METHODS

Data processing is done by combining Naïve Bayes Classification and Fuzzy C-Means also Naïve Bayes Classification and K-Means. From two combinations will be compared each accuracy to obtain the type of combination with a high accuracy and more recommended for determination of beef cattle quality in Semarang regency. Flow diagram the combination of Naïve Bayes Classification and Fuzzy C-Means shown in Figure 1.

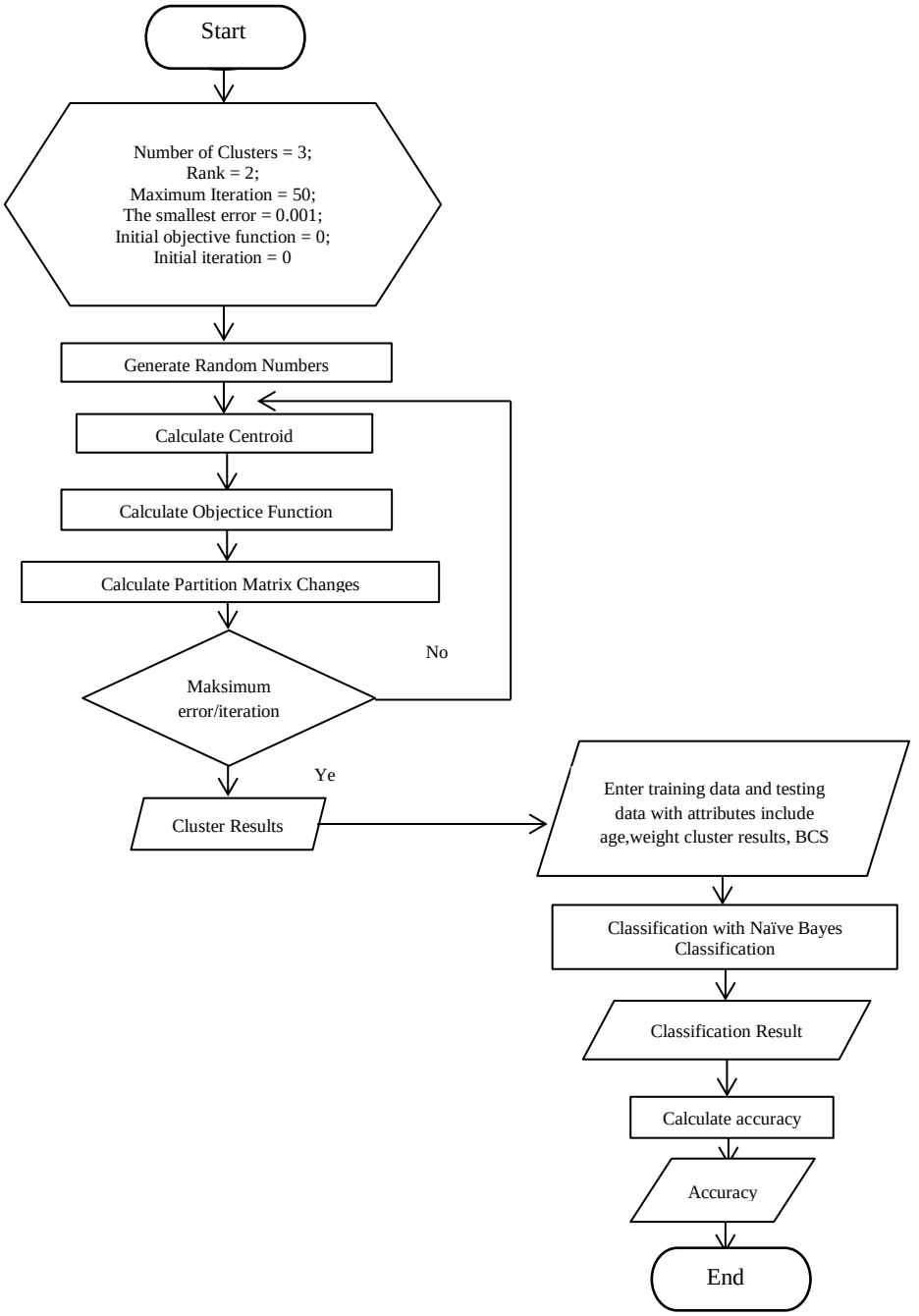

Figure 1. Flowchart of naïve bayes classification and fuzzy C-Means algorithm

The flowchart combination of algorithms Naïve Bayes Classification and K-Means shown in Figure 2.

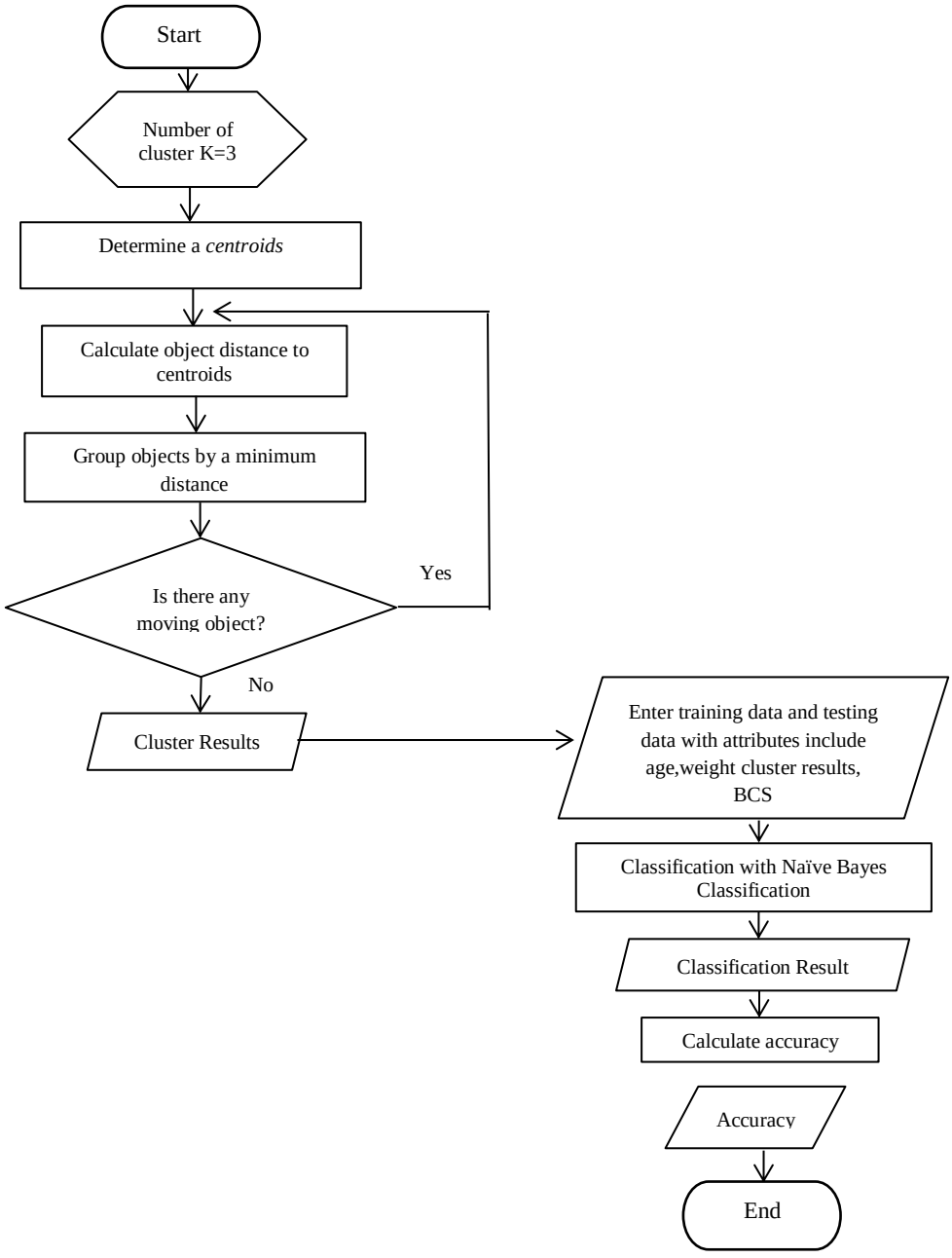

Figure 2. Flowchart of naïve bayes classification and K-Means algorithm

3. RESULTS AND DISCUSSION

This research uses dataset of beef cattle quality in 2016 and 2017 which amounted to 196 records that the process of taking it by plunging directly into the field. The attributes used in the process of determining the quality of beef cattle include age (month), weight (Kg) and BCS (Body Condition Score). This data will then be classified into Good or Bad quality. All the data type is continuous. In the process of calculation, used data training of 136 records data while data testing a number of 60 data records. The algorithm that used in this research are Fuzzy C-Means, K-Means, and Naïve Bayes Classification.

The first step in determining the quality of beef cattle was grouping the weight attribute into 3 classes using either Fuzzy C-Means algorithm or K-Means. The data type that originally is continuous, after the grouping process will turn into discrete data. The clustering result of weight attribute by using Fuzzy C-Means algorithm whose grouping technique that the existence of each data point in a cluster is determined by the degree of membership [11], shown in Table 1.

Table 1. Clustering results of weight attribute using fuzzy C-Means

No	Beef Cattle ID	Weight	C1	C2	C3	The Data Tend to enter a Cluster
1	B006	300	0,7827233612488	0,028201096157	0,189075542593	C1
2	B007	500	0,0056537772760	0,992530573248	0,001815649475	C2
3	B008	450	0,2297989807482	0,718730812543	0,051470206707	C2
4	B009	530	0,0088338707204	0,987896594959	0,003269534320	C2
5	B010	310	0,8843450242147	0,019822845034	0,095832130750	C1
6	L004	160	0,0918592412369	0,024079895491	0,884060863272	C3
7	L014	190	0,0335546765573	0,007303400083	0,959141923359	C3
8	L019	380	0,8689021610255	0,078721287726	0,052376551248	C1
9	M006	370	0,9240350829009	0,040522529347	0,035442387751	C1
10	M007	550	0,0312912083194	0,956211557313	0,012497234367	C2
	:	:	:	:	:	:
196	PO046	170	0,0725983699448	0,017985052992	0,909416577062	C3

From the clustering results using Fuzzy C-Means in Table 1, it can be seen that in C1 there is weight data between 280 to 410. In C2 there is data weight between 430 to 670 and on C3 there is data weight between 150 to 260. The advantages of Fuzzy C-Means are it has a high level of accuracy and fast computation time [12].

The clustering result of weight attribute by using K-Means algorithm shown in Table 2.

Table 2. Clustering results of weight attribute using K-Means

No	Beef Cattle ID	Weight	*Cluster*
1	B006	300	C1
2	B007	500	C2
3	B008	450	C3
4	B009	530	C2
5	B010	310	C1
6	L004	160	C1
7	L014	190	C1
8	L019	380	C3
9	M006	370	C3
10	M007	550	C2
:	:	:	:
196	PO046	170	C1

The results using K-Means in Table 2 can be seen that in C1 there is data weight between 150 to 370. At C2 there is weight data between 380 to 500 and at C3 there is weight data between 530 to 670. The results of K-Means are strongly influenced by the k parameter and centroid initialization. Generally K-Means initializes the centroid randomly [13]. After performing clustering process of weight attribute neither using Fuzzy C-Means algorithm and K-Means, do the process of beef cattle quality classification using Naïve Bayes Classification. The data with the clustering result of weight attribute using Fuzzy C-Means algorithm is shown in Table 3.

Table 3. Data with the clustering result of weight attribute using fuzzy C-Means

No	Beef Cattle ID	Age	Weight Cluster	BCS	Quality
1	B006	6	C1	4,5	Good
2	B007	60	C2	4	Good
3	B008	36	C2	4	Good
4	B009	36	C2	4,5	Good
5	B010	15	C1	3,5	Good
6	L004	72	C3	2	Bad
7	L014	7	C3	3	Good
8	L019	42	C1	4	Good
9	M006	24	C1	3	Good
10	M007	60	C2	4,5	Good
:	:	:	:	:	:
196	PO046	15	C3	2	Bad

The data with clustering of weight attribute by using K-Means algorithm is shown in Table 4.

Table 4. Data with the clustering result of weight attribute with K-Means

No	Beef Cattle ID	Age	Weight Cluster	BCS	Quality
1	B006	6	C1	4,5	Good
2	B007	60	C2	4	Good
3	B008	36	C3	4	Good
4	B009	36	C2	4,5	Good
5	B010	15	C1	3,5	Good
6	L004	72	C1	2	Bad
7	L014	7	C1	3	Good
8	L019	42	C3	4	Good
9	M006	24	C3	3	Good
10	M007	60	C2	4,5	Good
:	:	:	:	:	:
196	PO046	15	C1	2	Bad

The next step, do the process of classification with Naïve Bayes Classification, Naïve Bayes Classification algorithm will produce better results if using more training data [14]. In this process is done the division of data types are continuous and discrete. In this case it is known the data type of age attribute and BCS is continuous while the data type of attribute cluster weight is discrete. For continuous data type, calculate mean value (μ) and Standard Deviation (S) to then calculated probability value with Gauss Density function. While for discrete data type directly calculated probability value.

The results of the classification of beef quality using combination of Naïve Bayes Classification and Fuzzy C-Means are shown in Table 5.

Table 5. The results of naïve bayes classification and fuzzy C-Means combination

No	Beef Cattle ID	Age	*Weight Cluster*	BCS	Quality	Prediction
1	B001	72	C2	4	Good	Good
2	B002	84	C2	3	Good	Good
3	B003	30	C1	3	Good	Good
4	B004	84	C2	4	Good	Good
5	B005	72	C2	4,5	Good	Good
6	BS001	6	C1	3	Good	Good
7	L001	42	C1	3	Good	Good
8	L002	84	C2	3	Good	Good
9	L003	72	C1	2,5	Bad	Good
10	L005	72	C2	4	Good	Good
:	:	:	:	:	:	:
60	PO046	15	C3	2	Bad	Bad

After the classification results are known, then determine the accuracy used confusion matrix [15]. The Confussion Matrix is used to display the number of correct and false predictions made by the model compared to the actual classification in the test data [16], the use of confussion matrix allows better analysis of various types of errors [17]. The Confusion matrix of Naïve Bayes Classification and Fuzzy C-Means algorithms can be seen in Table 6.

Table 6. Confusion matrix of combination naïve bayes classification and fuzzy C-means

		Actual	
Prediction		Yes	No
	Yes	54	1
	No	1	4

From table 6 it is known that the value of TP is 54, TN is 4, FP is 1 and FN is 1. For calculating the level of accuracy can be used Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x100\% \quad (1)$$

$$Accuracy = \frac{54+4}{54+4+1+1} x100\% = 96{,}67\ \%$$

While the classification of beef quality using combination of Naïve Bayes Classification and K-Means algorithm result is shown in Table 7.

Table 7. The result of naïve bayes classification and K-Means combination

No	Beef Cattle ID	Age	Weight Cluster	BCS	Quality	Prediction
1	B001	72	C3	4	Good	Good
2	B002	84	C2	3	Good	Good
3	B003	30	C1	3	Good	Good
4	B004	84	C3	4	Good	Good
5	B005	72	C3	4,5	Good	Good
6	BS001	6	C1	3	Good	Good
7	L001	42	C2	3	Good	Good
8	L002	84	C2	3	Good	Good
9	L003	72	C1	2,5	Bad	Bad
10	L005	72	C3	4	Good	Good
:	:	:	:	:	:	:
60	PO046	15	C1	2	Bad	Bad

The Confusion matrix of Naïve Bayes Classification and K-Means combination can be seen in Table 8.

Table 8. The confusion matrix of naïve bayes classification and K-Means

		Actual	
Prediction		Yes	No
	Yes	54	0
	No	1	5

Accuracy can be calculated by Equation 1 as produce the following calculation.
$Accuracy = \frac{54+5}{54+5+0+1} x100\% = 98{,}33\ \%$

From that results, it can be seen that the accuracy of the Naïve Bayes Classification and K-Means algorithms is higher than the combination of Naïve Bayes Classification and Fuzzy C-Means algorithms. The accuracy comparison of the two combinations can be seen in Table 9.

Table 9. The accuracy comparison of the two combinations

Naïve Bayes Classification and Fuzzy C-Means	Naïve Bayes Classification and K-Means
96,67%	98,33%

In the process of clustering weights attributes using both Fuzzy C-Means and K-Means algorithm on the classification of beef cattle quality, proved equally optimized accuracy to the classification of beef cattle quality using Naïve Bayes Classification only. However, on the the combination of Naïve Bayes Classification and K-Means, the accuracy is 98.33%, it was higher than the combination of Naïve Bayes Classification and Fuzzy C-Means which has accuracy 96.67%.

The accuracy comparison of these two combinations is influenced by the result of cluster attribute weights performed with each clustering algorithm, i.e Fuzzy C-Means and K-Means. It can be seen from the cluster members of each different clustering algorithm, the number of cluster members weights each clustering algorithms will then yield result in different probability values in the classification process using Naïve Bayes Classification.

4. CONCLUSION

The accuracy of combination of the Naïve Bayes Classification and Fuzzy C-Means algorithm is 96,67%, while the combination of Naïve Bayes Classification and K-Means is 98.33%. The results showed that the accuracy of Naïve Bayes Classification and K-Means algorithm was higher than the Naïve Bayes Classification and Fuzzy C-Means algorithms combination accuracy with the difference of 1,66% so the combination of the algorithm that more recommended

for determine beef cattle quality is the combination of Naïve Bayes Classification and K-Means.

5. REFERENCES

[1] Selviana, N. I., & Mustakim. (2016). Analisis Perbandingan K-Means dan Fuzzy C-Means untuk Pemetaan Motivasi Balajar Mahasiswa. *Proceeding of Seminar Nasional Teknologi Informasi, Komunikasi dan Industri*.

[2] Nurzahputra, A., Muslim, M. A., & Khusniati, M. (2017). Penerapan Algoritma KMeans untuk Clustering Penilaian Dosen Berdasarkan Indeks Kepuasan Mahasiswa. *Techno.COM*, *16*(1), 17–24.

[3] Oliveira, J.V.D. & Pedrycz, W. (2007). *Advances in fuzzy clustering and its applications*. London: Wiley.

[4] Megawati, N., Mukid, M. A., & Rahmawati, R. (2013). Segmentasi Pasar pada Pusat Perbelanjaan Menggunakan Fuzzy C-Means (Studi Kasus: Rita Pasaraya Cilacap). *Jurnal Gaussian*, *2*(4), 343–350.

[5] Agusta, Y. (2007). K-Means Penerapan Permasalahan dan Metode Terkait. *Jurnal Sistem dan Informatika, 3(1), 47-60.*

[6] Bai, L., Liang, J., & Dang, C. (2011). An Initialization Method to Simultaneously Find Initial Cluster Centers and The Number of Clusters for Clustering Categorical Data. *Knowledge-Based Systems*, *24*(6), 785–795.

[7] Zhang, H., & Su, J. (2007). *Naive Bayesian Classifiers for Ranking*. Springer, Berlin, Heidelberg.

[8] Alam, A., Dwijatmiko, S. & Sumekar, W. (2014). Motivasi Peternak Terhadap Aktivitas Budidaya Ternak Sapi Potong di Kabupaten Buru Provinsi Maluku Farmers. *Agromedia*, *32*(2), 75-89.

[9] Abidin, Z. (2002). *Penggemukan Sapi Potong*. Jakarta : Agro Media Pustaka.

[10] Josaputri, C. A., Sugiharti, E., Arifudin, R. (2016). Decision Support Systems for The Determination of Cattle with Superior Seeds using AHP and SAW Method. *Scientific Journal of Informatics*, *3*(2), 21-30.

[11] Dunham, M. H. (2003). *Data Mining Introductory and Advance Topics*, New Jersey: Prentice Hall.

[12] Sutoyo, M. N., & Sumpala, A. T. (2015). Penerapan Fuzzy C- Means untuk Deteksi Dini Kemampuan Penalaran Matematis. *Scientific Journal of Informatics*, *2*(2), 129–136.

[13] Somantri, O., Wiyono, S., & Dairoh. (2016). Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM). *Scientific Journal of Informatics*, *3*(1), 34–45.

[14] Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive Evaluation of Performance of Computer Science Students of Unnes using Data Mining Based on Naïve Bayes Classifier (NBC). *Journal of Theoretical and Applied Information Technology*, *95*(4), 902–911.

[15] Safri, Y., Arifudin, R., & Muslim, M. A. (2018). K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining the Classification of Healthy Card Indonesia Giving to The Poor. *Scientific Journal of Informatics*. 5(1), 9-18.

[16] B. Venkatalakshmi & M.V Shivsankar. (2014). Heart Disease Diagnosis Using Predictive Data Mining. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 1873-1877.
[17] Novakovic, J. D. Veljovic, A. Ilic, S. S. Papic, Z. & Tmovic, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39–46.