Web-based Application for Classification Using Naïve Bayes and K-means Clustering
(Case Study: Tic-tac-toe Game)

Indriyani[1*] and M. Ihsan Alfani Putera[2]

[1*]STIKOM BALI Institute of Technologi and Business, Bali, Indonesia
indry.joice@gmail.com
[2]Department of Information System, Institut Teknologi Kalimantan, Indonesia
ihsanalfani@lecturer.itk.ac.id

Abstract - A database can consist of numerical and non-numerical attributes. However, several data processing algorithms, such as K-means clustering, can be used only in a dataset with numerical attributes. Data generalization by using Naïve Bayes and K-means clustering methods is usually employed WEKA (Waikato environment for knowledge analysis) application. Although the strength of WEKA lies in increasingly complete and sophisticated algorithms, the success of data mining still lies in the knowledge factor of the human implementer. The task of collecting high-quality data and knowledge of modeling and the use of appropriate algorithms is needed to guarantee the accuracy of the expected formulations. In this paper, we propose a simple web-based application that can be used like WEKA. The methodology used in this study includes several stages. The first stage is the preparation of data, which is the tic-tac-toe game dataset that is converted to CSV (comma-separated values) format. The next stage is the process of modifying data from non-numeric to numeric, specifically for clustering with the K-means algorithm. Afterward, the calculation of the distance between data is conducted and followed by data clustering. The final stage is the summary of these processes and results. From the experimental results, it was found that clustering can be done on categorical attributes that are transformed first into the numerical form using web-based applications.

***Keywords*—Naïve Bayes classifier, K-means clustering, Non-numeric to numeric, Web-based application.**

I. Introduction

ADVANCES in information technology has progressed rapidly in all fields of life. Lots of data generated by sophisticated information technology, ranging from industry, economics, science and technology, and various other fields [1]. Clustering is one of the popular data processing methods that have been used in various research fields, ranging from artificial intelligence, neural network technology, pattern recognition, and image processing. Clustering can be interpreted as a process of sorting a dataset into separate cluster groups and in which each cluster has something in common [2]. Clustering can also mean a method for collecting data in the form of unsupervised data mining, whose purpose is to separate an entire dataset into several clusters which is smaller in size. One of the popular clustering methods is the K-means algorithm [3]. K-means algorithm can be used for a dataset with numerical attributes. However, in reality, a database can consist of numerical and non-numerical attributes.

Another popular data processing method is clustering by using Naïve Bayes. Naïve Bayes is a simple probabilistic-based prediction technique that is based on the application of Bayes theorem or rule with a strong independence assumption on its features [4]. Data generalization by using Naïve Bayes and K-means clustering methods is usually employed WEKA (Waikato environment for knowledge analysis) application [5]. Although the strength of WEKA lies in increasingly complex and sophisticated algorithms, the success of data mining still lies in the knowledge factor of the human implementer. The task of collecting high-quality data and knowledge of modeling and the use of appropriate algorithms is needed to guarantee the accuracy of the expected formulations [6].

In this paper, we propose a web-based application for clustering a tic-tac-toe game dataset by using Naïve Bayes and K-means clustering. This application transforms the input non-numerical dataset into a numerical dataset, therefore it can be used for clustering by using the K-means algorithm.

II. Literature Review

A. Data Mining

Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases. According to Turban, et al. (2005), data mining can be divided into several

groups based on its task, which are [7]:

- **Description**: Sometimes researchers and analysis simply want to try to find ways to describe the patterns and trends contained in the data.
- **Estimation**: Estimation is similar to classification, except that the target variable is usually numerical.
- **Prediction**: Prediction is similar to classification and estimation, except that prediction gives a result that will be in the future.
- **Classification**: The target variable in classification is usually categorical.
- **Clustering**: Clustering is an unsupervised method for finding and grouping data that has similar characteristics.
- **Association**: The task of association in data mining is finding attributes that appear at a time. In the business world, it is more commonly called shopping basket analysis.

B. Naïve Bayes Algorithm

Naïve Bayes is a statistical classification method that can be used to predict the probability of membership in a class [8]. Naïve Bayes is based on the Bayes theorem which has classification capabilities similar to decision trees and neural networks. Naïve Bayes has proven to have high accuracy and speed when applied to databases with large data [9]. Naïve Bayes' prediction is based on the Bayes theorem formula as follows [10][11]:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

C. K-means Algorithm

The K-means algorithm is one of the clustering algorithms that is the most commonly used in various applications [12]. This algorithm is used for grouping data based on its attribute value into as many as *k* clusters. The K-means algorithm is as follows [13]:

1. Set the number of *k* which represent the number of clusters that will be formed,
2. Select *k* number of data randomly that will be used as the center of the clusters (centroid),
3. Determine the membership of a cluster by collecting data that is close to the centroid,
4. Update the centroid of each cluster by averaging the value of data that belong to the cluster,
5. Repeat steps 3-4 until the centroid of all the clusters is not changing anymore.

The distance between a data point and the centroid of a cluster is usually measured by using Euclidean distance. Euclidean distance calculates the square root of the difference of the attribute's value from a pair of data, using the following formula:

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} \quad (2)$$

The example of Euclidean distance measurement is as follows. If data A has the value of (0, 3, 4, 5) and data B has the value of (7, 6, 3, -1), then the Euclidean distance between A and B is:

$$d_{AB} = \sqrt{(0-7)^2 + (3-6)^2 + (4-3)^2 + (5-(-1))^2}$$

$$d_{AB} = \sqrt{49 + 9 + 1 + 36}$$

$$d_{AB} = 9.747$$

D. Accuracy

The accuracy of the predicted results can be calculated when the amount of data that is classified correctly or incorrectly is known [14]. The formula to calculate accuracy is as follows:

$$Accuracy = \frac{The\ number\ of\ correct\ prediction}{The\ number\ of\ prediction} \quad (3)$$

III. METHODOLOGY

A. System Design

The developed system is used to test the tic-tac-toe game dataset using the Naïve Bayes algorithm and K-means clustering. The design of the proposed system is shown in Figure 1.

Figure 1. The design of the proposed system.

B. Data Preparation

The data that will be used for this research is the tic-tac-toe game dataset. In general, analyzing the tic-tac-toe game dataset use an application such as WEKA. However, in this study, we propose a new simple web-based application that can be used like WEKA. The tic-tac-toe game dataset is usually a file with an .arff extension, such as shown in Figure 2. For this application, the .arff file must be converted into a file with the .csv extension to be uploaded to the database. The example of the converted .csv file is shown in Figure 3.

C. Binning Process

The binning process is also called the normalization process, which is the process of transforming values from non-numerical data into numerical data that can be calculated [15]. This calculation process is needed in implementing the K-means algorithm for the process of data clustering.

The used dataset is the tic-tac-toe game dataset. The tic-tac-toe game is a classic board-type game with a 3×3 board. Board-game is a game with pieces that are placed on top, moved from or moved on a special surface, which is called as board [16]. In the tic-tac-toe game, there are nine rooms (3×3) in the shape of a rectangular box (pawns). This game uses two-player symbols, 'X' or 'O', whereas if they do not contain both, a symbol B [17]. To make it easier to process

```
@relation 'Data set tic tac toe full urut'

@attribute top-left-square {x,o,b}
@attribute top-middle-square {x,o,b}
@attribute middle-left-square {x,o,b}
@attribute top-right-square {x,o,b}
@attribute middle-middle-square {o,b,x}
@attribute middle-right-square {o,b,x}
@attribute bottom-left-square {x,o,b}
@attribute bottom-middle-square {o,x,b}
@attribute bottom-right-square {o,x,b}
@attribute Type {positive,negative}

@data
x,x,x,x,o,o,x,o,o,positive
x,x,x,x,o,o,o,x,o,positive
x,x,x,x,o,o,o,o,x,positive
x,x,x,x,o,o,o,b,b,positive
x,x,x,x,o,o,b,o,b,positive
x,x,x,x,o,o,b,b,o,positive
```

Figure 2. The tic-tac-toe game dataset with .arff extension

	A	B	C	D	E	F	G	H	I	J
1	o	x	x	o	x	x	o	b	o	negative
2	o	x	o	o	o	x	x	x	x	positive
3	x	x	o	o	x	b	b	x	o	positive
4	b	x	x	o	x	o	b	x	o	positive
5	b	b	o	o	b	b	x	x	x	positive
6	x	o	o	x	b	o	x	x	b	positive
7	o	o	o	x	b	x	b	b	x	negative

Figure 3. The tic-tac-toe game dataset that has been transformed into .csv extension

the by using K-means clustering, the data must be changed with the following rules:

1. The value 'X' is converted into number 1,
2. The value 'O' is converted into number 2,
3. The value 'B' is converted into number 0.

In the binning process for the K-means clustering, the data with the .csv extension will be transformed as shown in Figure 4.

	A	B	C	D	E	F	G	H	I	J
1	2	1	1	2	1	1	2	0	2	negative
2	2	1	2	2	2	1	1	1	1	positive
3	1	1	2	2	1	0	0	1	2	positive
4	0	1	1	2	1	2	0	1	2	positive
5	0	0	2	2	0	0	1	1	1	positive

Figure 4. The tic-tac-toe game dataset that has been transformed into .csv extension and has been through the binning process

D. Data Processing

The sample data of the tic-tac-toe game dataset, which has been converted to the CSV file, will be tested in a web-based application that has been made. The stages of the data processing include importing data, the Naïve Bayes algorithm, and the K-means algorithm.

1. Importing Data

The user needs to import the CSV data into the web-based application for further processing. The user interface of the web-based application for the import process is shown in Figure 5.

2. Naïve Bayes Algorithm

The "Classify" menu that leads to the "Naïve Bayes" submenu is used to display the results of the data classification process by using Naïve Bayes Algorithm. The user interface of the web-based application for the classification process by using the Naïve Bayes algorithm is shown in Figure 6.

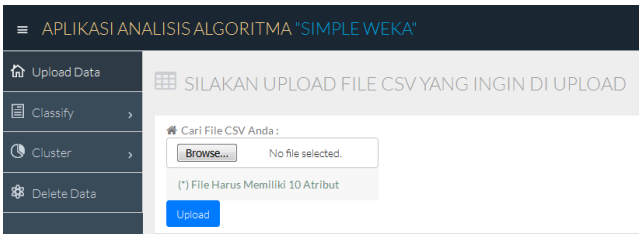

Figure 5. The user interface for importing tic-tac-toe game dataset

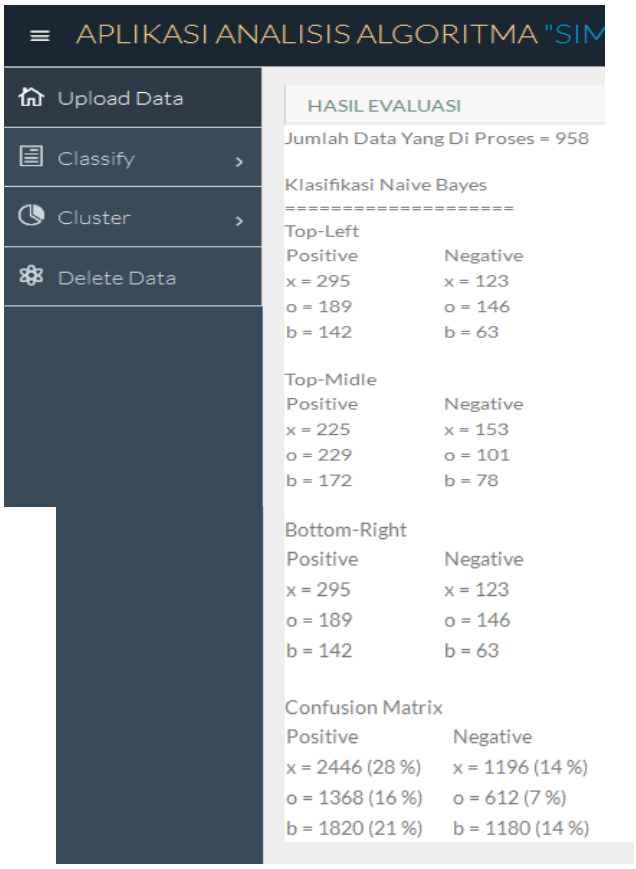

Figure 6. The user interface for the process of result of classification by using Naïve Bayes algorithm

3. K-means Clustering Algorithm

The "Classify" menu that leads to the "K-means" submenu is used to display the results of the data clustering process by using the K-means algorithm. The calculation phase for clustering is different from the classification because clustering can only be used if the data is numeric. Therefore the tic-tac-toe game dataset which initially has non-numerical attributes must be changed first to the numerical attributes in the binning process, which result can be seen as in Figure 7.

The result of the centroid selection can be seen in Figure 8. If the centroid of each cluster has been selected, the result of the K-means algorithm for the tic-tac-toe game dataset is shown in Figure 9.

DATA OBJEK

Objek	Data Setelah DiRubah Numerik
1	1,1,1,1,2,2,1,2,2
2	1,1,1,1,2,2,2,1,2
3	1,1,1,1,2,2,2,2,1
4	1,1,1,1,2,2,2,0,0
5	1,1,1,1,2,2,0,2,0
6	1,1,1,1,2,2,0,0,2

Figure 7. The result of binning process

DATA CLUSTER

Hapus Cluster | Tambah Cluster

Cluster	Centroid
1	1,1,1,1,2,2,1,2,2
2	2,2,1,1,1,2,2,1,1

Figure 8. The selected centroid of each cluster

Cluster Centroid Akhir :
Cluster 1 -> 2 1 1 1 1 1 1 1 1
Cluster 2 -> 2 1 2 2 1 1 2 1 1
Cluster 3 -> 0 1 1 2 1 1 1 1 2
Total Iterasi : 6
jumlah Cluster : 3

Total	180
Cluster 1	99 (55%)
Cluster 2	24 (13%)
Cluster 3	57 (32%)

Ulangi Lagi

Figure 9. The analisis result of the tic-tac-toe game dataset by using K-means clustering algorithm

IV. Experimental Results

The experiments will be conducted by using the 30%, 70%, 100% data in the dataset as the test set. The result of the system will be compared with the original data therefore the accuracy of the system in each experiment can be calculated. The accuracy of the system is calculated as follows:

1. Calculate the probability of positive and negative data. If the experiment is conducted by using 30% of the data, which is about 287-300 data. If we use 287 data, the number of positive data is 184, and the number of negative data is 103, then the probability of the positive and negative data can be calculated as:

$$Positive\ data = \frac{184}{287} = 0.64$$

$$Negative\ data = \frac{103}{287} = 0.36$$

2. Calculate the probability of each attribute for the positive data, for example:

$$Top - left = \frac{81}{184} = 0.44$$

$$Top - middle = \frac{77}{184} = 0.42$$

$$Middle - left = \frac{79}{184} = 0.43$$

3. Multiply the probability of each attribute to obtain positive prior.

$$Positive\ prior = 0.44 \times 0.42 \times 0.43 \times \ldots$$

4. Multiply the positive prior with the positive data to obtain the positive prediction.

$$Positive\ prediction = 1.56 \times positive\ prior$$

5. Repeat steps 2-4 for the negative data to obtain the negative prediction.
6. If the value of positive prediction is bigger than the value of negative prediction, then the final prediction is positive and vice versa.
7. If the experiments are done, for example, for 4 times with different data, then the accuracy value can be measured by using Equation (3).

A. *Experiment on Naïve Bayes*

In this experiment, only 30% of the data is used. The number of selected data is between 287 – 300 data, and the results are shown in Figure 10 – Figure 13, respectively. The experimental result by using 30% of the data records is shown in Table 1, therefore the accuracy value can be calculated as follows. The classification of 30% of tic-tac-toe dataset by using Naïve Bayes algorithm gives 100% accuracy.

$$Accuracy = \frac{4}{4} \times 100\% = 100\%$$

TABLE I
THE EXPERIMENTAL RESULT OF NAÏVE BAYES CLASSIFICATION USING 30% OF THE DATA

The number of data	Prediction
287	Positive
288	Positive
191	Positive
300	Positive

B. *Experiment on K-means Clustering*

This experiment used 30 data records for clustering the data into 3 and 4 clusters. The example of the 30 data records and the centroid of each cluster is shown in Figure 14. The analysis of the data using 3 and 4 clusters can be seen in Figure 15 and Figure 16, respectively. It can be seen from the experimental results using 30 data records that the first cluster is more dominant because it consists of 60% of data.

Jumlah Instance : 287
Klasifikasi data yg benar 184 (64.11 %)
Klasifikasi data yg salah 103 (35.89 %)

Confusion Matrix

a	b	
81	37	a=positive
49	47	b=negative
54	19	

Prediksi data ini : Positive

Figure 10. Experiment using 30% of the data with the total of 287 data

Jumlah Instance : 288
Klasifikasi data yg benar 185 (64.24 %)
Klasifikasi data yg salah 103 (35.76 %)

Confusion Matrix

a	b	
82	37	a=positive
49	47	b=negative
54	19	

Prediksi data ini : Positive

Figure 11. Experiment using 30% of the data with the total of 288 data

Jumlah Instance : 291
Klasifikasi data yg benar 187 (64.26 %)
Klasifikasi data yg salah 104 (35.74 %)

Confusion Matrix

a	b	
84	38	a=positive
49	47	b=negative
54	19	

Prediksi data ini : Positive

Figure 12. Experiment using 30% of the data with the total of 191 data

Jumlah Instance : 300
Klasifikasi data yg benar 194 (64.67 %)
Klasifikasi data yg salah 106 (35.33 %)

Confusion Matrix

a	b	
87	39	a=positive
51	48	b=negative
56	19	

Prediksi data ini : Positive

Figure 13. Experiment using 30% of the data with the total of 300 data

CLUSTERING DATA

Total Data adalah : 30

Objek	Data Setelah DiRubah Numerik
1	2,1,1,2,1,1,2,0,2
2	2,1,2,2,2,1,1,1,1
3	1,1,2,2,1,0,0,1,2
4	0,1,1,2,1,2,0,1,2
5	0,0,2,2,0,0,1,1,1
6	1,2,2,1,0,2,1,1,0
7	2,2,2,1,0,1,0,0,1
8	1,1,2,2,1,2,2,1,1

DATA CLUSTER

Hapus Cluster | Tambah Cluster

Cluster	Centroid
1	2,1,1,2,1,1,2,0,2
2	2,2,2,1,1,0,1,1,2
3	0,1,2,2,1,1,0,1,2

Proses

Figure 14. The 30 data records and the selected centroid of each cluster

Cluster Centroid Akhir :
Cluster 1 -> 1 1 1 2 1 1 2 1 1
Cluster 2 -> 2 2 1 1 1 1 0 1 1
Cluster 3 -> 0 0 2 2 1 1 0 1 2
Total Iterasi : 3
jumlah Cluster : 3

Total	90
Cluster 1	37 (41%)
Cluster 2	25 (28%)
Cluster 3	28 (31%)

Figure 15. The analysis of the 30 data records by using 3 clusters

Cluster Centroid Akhir :
Cluster 1 -> 1 1 1 2 1 1 2 1 1
Cluster 2 -> 2 1 1 1 1 1 0 1 2
Cluster 3 -> 0 1 2 2 1 1 0 1 2
Cluster 4 -> 1 1 1 0 1 2 1 1 1
Total Iterasi : 6
jumlah Cluster : 4

Total	180
Cluster 1	69 (38%)
Cluster 2	45 (25%)
Cluster 3	52 (29%)
Cluster 4	14 (8%)

Figure 16. The analysis of the 30 data records by using 4 clusters

But when the number of clusters is added, the amount of data in cluster 1 is reduced to 38%.

It can also be seen that by using 3 clusters, the program iterates 3 times, but by using the same data and with 4 cluster analysis, the program iterates 6 times. Therefore, it can be concluded that the bigger the number of the cluster used, the longer it takes for the system to analyze the data [18].

V. CONCLUSION

From the experimental results of Naïve Bayes classification by using 30%, 70%, and 100% of the dataset, we can make a comparison table of appearance rates for each object "X", "O", "B" as in Table 2. From Table 2 we can conclude that: 1) the more data the probability of the appearance of an object is lower; 2) by using 70% and 100% data, the probability value appears the same, but the percentage of the probability is different because when viewed from the 30% and 70% data the image is influenced by the negative value of the object (data mismatch) so that the greater the data the more the error value / negative objects in a data series.

TABLE II
THE COMPARISON OF APPEARANCE RATES OF OBJECT "X", "O", AND "B"

No	Object	30% Data	70% Data	100% Data
1	"X"	1145 (44%)	2446 (41%)	2446 (28%)
2	"O"	580 (22%)	1368 (23%)	1368 (16%)
3	"B"	896 (34%)	2141 (36%)	4788 (56%)

From the experimental results of K-means clustering algorithm, we can conclude that: 1) the more data that is analyzed the longer the required computational time and also the higher the needed hardware resource specification; 2) with the same amount of data, computing time is also influenced by the number of clusters formed because it affects the number of iterations that the system does to cluster.

In this research, the use of the Naïve Bayes and K-means algorithm to determine the probability value of objects in the tic-tac-toe game has been discussed. It is expected that for future work, a comparison of other methods both in terms of classification and clustering algorithm can be conducted. Therefore, the advantages of each method can be determined.

REFERENCES

[1] K. Chellapilla and D. B. Fogel, "Evolution, Neural Networks, Games, and Intelligence," in *1999 Proceedings of the IEEE*, vol. 87, no. 9.

[2] Sinharay, S. *An overview of statistics in education*. Elsevier, 2010, pp. 1-11.

[3] Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A., "What to do when K-means clustering fails: a simple yet principled alternative algorithm," *PloS one*, vol. 11, no. 9, 2016.

[4] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., and Strachan, R., "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," *Expert systems with applications*, vol. 41, no. 4, 2014, pp. 1937-1946.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, 2009, pp. 10-18.

[6] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E., "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, 2015, pp. 1.

[7] E. Turban, et al., *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset, 2005.

[8] J. F. Ulysses, "Data Mining Classification Untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan Metode Naïve Bayes," *Magister Teknik Informatika Universitas Atma Jaya Yogyakarta*.

[9] Zulfikar, W. B., Lukman, N., "Perbandingan Naive Bayes Classifier Dengan Nearest Neighbor Untuk Identifikasi Penyakit Mata," *Jurnal Online Informatika*, vol. 1, no. 2, 2016, pp. 82-86.

[10] A. Jananto, "Algoritma Naïve Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa," *Jurnal Teknologi Informasi DINAMIK*, vol. 18, no. 1, Jan 2013.

[11] M. S. Suhartinah and Ernastuti, "Graduation Prediction of Gunadarma University Students Using Algorithm Naïve Bayes C4.5 Algorithm," *Faculty of Industrial Engineering*, 2010.

[12] I. Budiman, T. Prahasto, and Y. Christyono, "Data Clustering Menggunakan Metodologi Crisp-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma," presented in 2012 Seminar Nasional Aplikasi Teknologi Informasi (SNATI 2012), Yogyakarta.

[13] J. O. Ong, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University," *Jurnal Ilmiah Teknik Industri*, vol 12, no. 1, June 2013, pp. 10-13.

[14] Indraswari, R., Herulambang, W., Rokhana, R., "Melanoma classification using automatic region growing for image segmentation," in *Proceeding ICTA 2017 UBHARA Surabaya*, pp. 165-172.

[15] Zeng, G., "A necessary condition for a good binning algorithm in credit scoring," *Applied Mathematical Sciences*, vol. 8, no. 65, 2014, pp. 3229-3242.

[16] S. Jain and N. Kera, "An Intelligent Method for Solving Tic-tac-toe Problem," presented at the 2015 International Conference on Computing, Communication, and Automation (ICCCA).

[17] Abu Dalffa, M., Abu-Nasser, B. S., Abu-Nasser, S. S., "Tic-Tac-Toe Learning Using Artificial Neural Network," *International Journal of Engineering and Information System (IJEAIS)*, vol. 3, no. 2, February 2019, pp. 9-19.

[18] Kodinariya, T. M., Makwana, P. R., "Review on determining number of Cluster in K-Means Clustering," *International Journal*, vol. 1, no. 6, 2013, pp. 90-95.