Predictions of COVID-19 Spread by Using Supervised Data Mining Techniques

Predictions of COVID-19 Spread by Using Supervised Data Mining Techniques

Wid Akeel Awadh[1], Ali Salah Alasady[2] and Hadeel Ismail Mustafa[3]

[1, 3] Dep. of Computer Information System, College of Computer Science and Information Technology
[2] Dep. of Computer science, College of Computer Science and Information Technology
University of Basrah - Iraq

E-mail: umzainali@gmail.com

Abstract. In the wake of the outbreak of the new coronavirus, the countries in the world have fought to combat the spread of infection and imposed preventive measures to compel the population to social distancing, which led to a global crisis. Important strategies must be studied and identified to prevent and control the spread of coronavirus COVID-19 disease 2019. In this paper, the effect of preventive strategies on COVID-19 spread was studied, a model based on supervised data mining algorithms was presented and the best algorithm was suggested on the basis of accuracy. In this model, three classifiers (Naive Bayes, Multilayer Perceptron and J48) depended on the questionnaires filled out by Basra City respondents. The questionnaires consisted of 25 questions that covered fields most related to and that affect the prevention of COVID-19 spread, including demographic, psychological, health management, cognitive, awareness and preventive factors. A total of 1017 respondents were collected. This model was developed using Weka 3.8 tool. Results showed that quarantine played an important role in controlling the spread of the disease. By comparing the accuracy of the algorithms used, the best algorithm was found to be J48.

Key words: Supervised Data Mining, COVID-19, Predictions, Accuracy, NB, MLP, J48.

1. Introduction

A new type of coronavirus has been discovered recently after it has been identified as the cause of the spread of one of the diseases that started in China from December 2019. The resulting disease is called coronavirus 2019 (COVID-19). In March 2020, the World Health Organization declared COVID-19 as a pandemic [1]. Evidence that men are more likely to die from COVID-19 than women is growing. People who suffer from aging diseases, weak immunity and lung-related diseases are the most affected [2]. The signs and symptoms of the disease may appear 2–14 days after exposure and may include the following: fever, coughing, shortness of breath, runny nose, headache, diarrhoea, vomiting and loss of sense of smell and taste. For protection from this epidemic, people must avoid contact with others and surfaces; avoid touching eyes, nose or mouth; wash hands frequently with soap and water; and maintain social distancing by 1–3 m [3]. The two important stages of COVID-19 are stages 2 and 3. Stage 2 involves infection from one person to another, whilst stage 3 involves infection from society. In accordance with COVID-19 stages,

the action plan could be determined by different countries. In Basra, the first case of COVID-19 was documented on 24 February 2020, originating from Najaf-Iraq. After 3 months, this disease spread in almost all areas of the city. A total of 22,926 cases have been recorded thus far. Among them, 16,817 recovered and 630 died. In this research, the following questions were answered: first, what are the most important factors that reduce the spread of COVID-19? Second, among the supervised data mining (DM) algorithms used, which could best predict the prevention of the spread of the disease? These questions were answered using DM techniques.

2. DM Technique

DM is a computational process applied in many areas; it aims to obtain useful and hidden predictive knowledge. DM techniques are used to construct a model where new information could be identified by unknown information [4]. One popular feature of all DM techniques is automated learning that recognises new patterns in the observed datasets [5]. DM specialists have devoted their careers to enhance the understanding on how to process and draw conclusions from huge quantities of knowledge depending on the techniques from the intersection of database management, machine learning and statistics. Thus, they divided the algorithms into two simple groups: supervised and unsupervised algorithms [6] [7]. Supervised DM is used to build models by using training data with a familiar class to which the data belong. The training data are composed of a series of training examples. Each example is a pair consisting of a vector of attribute value and a desired output class. A supervised learning algorithm analyses the training data and generates a feature inferred, which could be used to map new examples. This algorithm allows to determine the classes labels correctly for unseen instances. Classification methods belong to this group. The most popular classification methods are decision trees, classification rules, Bayesian networks and neural networks [8] [9]. The role of unsupervised learning is more complex than that of supervised learning because its only aims at searching the data for interesting connections and attempts to group elements by postulating class descriptions for sufficient numbers of classes to cover all objects in the database [10]. These tasks range from the identification of potentially useful regularities among the data couched in the language of description given to the discovery of concepts by conceptual clustering and constructive inference and further discovery of empirical laws relating to concepts developed by the method. Clustering and association rule methods belong to this group [8] [11]. Several different classifier models are available in the literature and determining the best method is not possible as they vary from each other in several aspects, such as learning rate, amount of training data, speed of classification and accuracy. In this paper, the effect of three algorithms, namely, Naive Bayes (NB), Multilayer Perceptron (MLP) and J48, on data analysis was discussed. These three classifier algorithms were used to predict the final class (prevention of COVID-19 spread) that some new unlabelled attributes belong to. The selected algorithms were also used to find the most convenient method to predict the final class.

2.1 Naive Bayes

NB algorithm is one of the most popular machine learning methods, specifically data analysis and classification. This method relies on the statistical concept of Bayes' theorem, which calculates the probability of a specific result by using available information [12]. This classifier is called naive because it relies on the principle of independence assumptions, that is, the relationship among all attributes and features is considered independent of one another [13]. The NB classifier model is characterised by the ease of construction and development and the ability to process large data, outperforming a number of sophisticated and sophisticated algorithms. The model is trained with the data and its available properties in databases, determines the type of new records and then classifies them on the basis of data and statistics previously available [14]. It is used in many systems, such as for identifying harmful messages; classifying documents, such as in news sites, to anticipate the type of document (e.g. politics, sports and technology);

recognising the views and feelings in the text content (negative, positive or optimistic) and in face recognition in pictures [15].

2.2 Multilayer Perceptron

MLP algorithm is one of the most popular neural networks algorithms. It consists of a perceptron, an input layer for receiving the signal, an output layer for making a decision or prediction about the data and an infinite number of hidden layers, which are MLP's true computational engine between the two layers [16]. MLP is also applied to supervised machine learning problems; they train on a collection of pairs of input–output and learn to model the correlation between those pairs. Network training requires adjusting the model's parameters (weights) to reduce error [17] [18]. This network could also be used for unsupervised machine learning by using auto-associative structure, where similar values are set for the network inputs and outputs. This method is intensive in computational terms. For any fair representation, the MLP network must have at least three hidden layers, thus requiring a long time [19] [20].

2.3 J48

J48 algorithm is an extension of ID3 algorithm. It was developed by Ross Quinlan in 1993 to generate a tree-like hierarchy and construct classification trees that represent a flowchart structure, in which the non-terminal nodes indicate the attribute tests and the terminal nodes represent the decision outcomes [21] [22]. This classifier is one of the most widely used machine learning language processing domains. The main advantages of this algorithm are easy construction of graphical classification and low-cost formal generation. However, this algorithm does not generate multiple redundant attributes and modules and it is quite susceptible to noise in the data [23].

3. Model

3.1 Data Discerption

Quantitative method relies on the use of questionnaire survey. In this work, questionnaires were built on Google Docs to collect data online. These questionnaires were used to collect data of the real direction and challenges of the respondents in Basra City on the basis of on their different experiences, allowing for collection of more accurate coded data. The research data, a total of 1017 samples, were collected in March of 2020. The six main parts of the questionnaire were demographic, psychological, health management, cognitive, awareness and preventive factors. Table 1 shows the questionnaire parts, question number and question description and the respondents' answers. These answers were shortened and converted from nominal to numeric type for ease of use and understanding. The answers were classified on the basis of Likert scale as follows: 1 for strongly agree, 2 for agree, 3 for not applicable, 4 for disagree and 5 strongly disagree. The initial step in data pre-processing involved preparing the data and converting them for evaluation and processing using MS Excel. In the second step, the collected responses were converted carefully using Weka 3.8 tool.

Table 1. Questionnaire Description

Part	No.	Distribution	Rank
Demographic Data	Q1	Gender	Male
			Female
	Q2	Age (years)	15–24
			25–34
			35–44
			45–54
			> 54
	Q3	Job	Healthcare field
			Engineering field
			Teaching field
			Student
			Earner
			Unemployed
	Q4	Address	City centre
			City outskirts
	Q5	Level of Education	Ph. D.
			MS. C.
			BS. C.
			Diploma
			Secondary
			Middle
			Primary
Psychological Factors	Q6	Social media is the most influencing factor of my psychological state.	1, 2, 3, 4, 5
	Q7	I suffer from nervous tension and anxiety during this period.	1, 2, 3, 4, 5
	Q8	I consider storing goods and food items a necessity in this period.	1, 2, 3, 4, 5
	Q9	Seeking help from psychologists is necessary during this period.	1, 2, 3, 4, 5
Health Management Factors	Q10	Basra is one of the healthiest cities to fight COVID-19.	1, 2, 3, 4, 5
	Q11	The medical personnel from the province are ready to deal with the virus.	1, 2, 3, 4, 5
	Q12	The crisis could be overcome by connecting health institutions with international scientific research centres.	1, 2, 3, 4, 5
	Q13	I support receiving medical personnel from outside Iraq, such as China.	1, 2, 3, 4, 5
Cognitive Factors	Q14	One of the most important transmission routes of infection is inhalation of the droplets produced when an infected person sneezes or coughs.	1, 2, 3, 4, 5
	Q15	Contact with surfaces contaminated with droplets from an infected person is a typical way to catch the disease.	1, 2, 3, 4, 5
	Q16	COVID-19 one of the most dangerous infectious viruses in recent times.	1, 2, 3, 4, 5
	Q17	Taking nonsteroidal anti-inflammatory drugs, such as Provin and Punestan, reduce the severity of symptoms.	1, 2, 3, 4, 5
Awareness Factors	Q18	Staying away from markets, restaurants and gatherings is a good way to prevent the spread of the virus.	1, 2, 3, 4, 5-
	Q19	I support the compulsory quarantine of all expats from abroad.	1, 2, 3, 4, 5
	Q20	I support curfews and disruption of circuits and schools and I encourage e-learning.	1, 2, 3, 4, 5
	Q21	I support isolating infected people in hospitals far from residential areas.	1, 2, 3, 4, 5
Preventive Factors	Q22	Meat and eggs should be cooked well.	1, 2, 3, 4, 5
	Q23	It is necessary to wear masks and paws when leaving the house.	1, 2, 3, 4, 5
	Q24	The immune system is strengthened by taking high doses of zinc and vitamin C.	1, 2, 3, 4, 5
	Q25	Washing hands with soap and water for at least 20 seconds and using sterilisers containing 70% alcohol are the basic information that I know.	1, 2, 3, 4, 5
Total of all items			25

3.2 Questionnaire Reliability

Reliability is the extent to which an instrument would produce the same results, whether the measurement under the same conditions was to be taken again. For example, measuring people's height and weight is also extremely reliable [24]. Cronbach's alpha is statistically one of most effective methods to measure internal consistency and reliability; it is used to calculate the extent to which the questions in the survey all measure the same underlying construct. When Cronbach's alpha > 0.7, its internal reliability performance is satisfactory [25]. Table 2 shows that the Cronbach's alpha was 0.8 for the scaled variables, which consisted of 21 items (parts 2–5) and 1017 respondents.

Table 2. Questionnaire Reliability

Reliability Statistics			
Cronbach's alpha	***No. of items***	***No. of respondents***	***% of respondents***
0.8	21	1017	100%

4. Results and Discussions

A Weka software package developed at Waikato University in New Zealand was used to implement this study. This package is run on Java, and it stands out as perhaps the most professional and extensive package of machine learning algorithms [26]. Analysing the effect of attributes is necessary to obtain an enhanced insight into the importance of the attributes. Thus, the effect of certain attributes of the model on the final class was analysed in the present study. This analysis discussed the most correlated attributes (respondents) to the final class (prevention of COVID-19 spread) and how they may affect the final class. This stage revealed the average correlation of the attributes with the final class. The respondents with high correlation could be regarded as points of advice for interested parties in this field. Weka provides several filters that could be used to clean up the data before invoking a classifier. In this stage, a filter (CorrelationAttributeEval) was used to find the correlation between attributes and the final class. This filter evaluated the value of an attribute by calculating the correlation (Pearson's) between it and the final class.

Table 3. Correlation Rate of Responses

Sequence	**Respondent**	**Average**	**Sequence**	**Respondent**	**Average**
1	Q19	0.15624	13	Q3	0.04214
2	Q16	0.14841	14	Q6	0.03534
3	Q24	0.13699	15	Q8	0.03149
4	Q22	0.10039	16	Q5	0.03136
5	Q25	0.0961	17	Q17	0.02950
6	Q15	0.09046	18	Q14	0.01748
7	Q20	0.08735	19	Q1	0.0159
8	Q11	0.07532	20	Q2	0.01316
9	Q4	0.0587	21	Q9	0.00743
10	Q10	0.05505	22	Q7	0.00483
11	Q23	0.04537	23	Q12	0.00414
12	Q18	0.04358	24	Q13	0.00352

Table 3 displays the correlation rate with the evaluation mode between the attributes and the final class (10-fold cross validation) to ensure accuracy. This analysis aimed to determine the importance of each attribute

individually. The responses were ranked from the highest to the lowest correlation rate with the final class. Those with the highest correlation rates represent the value most correlated with the final class. The results showed that Q19 affected the final class the most, followed by Q16, Q24, Q22 and Q25, whereas Q9, Q7, Q12 and Q13 had the smallest output effect. Some experiments were also conducted to evaluate the performance and usefulness of the different classification models of the prevention of COVID-19 spread. The results are summarized in Tables 4–6.

Table 4. Predictive Results of Classifier Models

Evaluation Criteria	Classifier Models		
	NB	***MLP***	***J48***
Timing to Build Model (in seconds)	0.05	25.49	0.05
Correctly Classified Instances	906	934	953
Incorrectly Classified Instances	110	82	64
Prediction Accuracy	89.17%	91.92%	93.79%

Table 5. Comparison of Performance

Evaluation Criteria	Classifier Models		
	NB	***MLP***	***J48***
Kappa Statistics	0.2017	0.1542	0.1368
Mean Absolute Error	0.122	0.0835	0.099
Root Mean Squared Error	0.2975	0.2676	0.2412
Relative Absolute Error	114.27%	78.23%	92.73%
Root Relative Squared Error	192.29%	116.26%	104.80%

Table 6. Comparison of Evaluation Measures

Classifier Model	TP	FP	Precision	Recall
NB	0.925	0.667	0.959	0.925
MLP	0.964	0.825	0.952	0.175
J48	0.987	0.895	0.0.949	0.987

The performance of the three models was measured on the basis of three criteria: prediction accuracy, building time and error average, as shown in Figures 1–3, respectively.

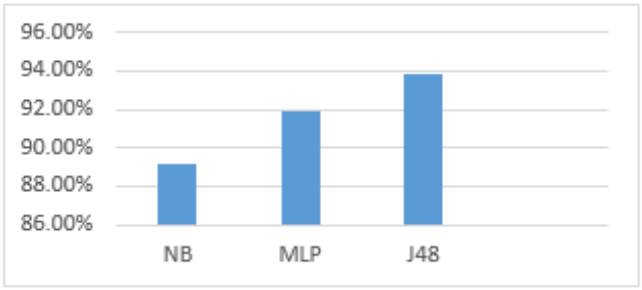

Figure 1. Prediction Accuracy

Figure 1 illustrates the classifier models used in the experiment; the accuracy average of NB algorithm was the lowest.

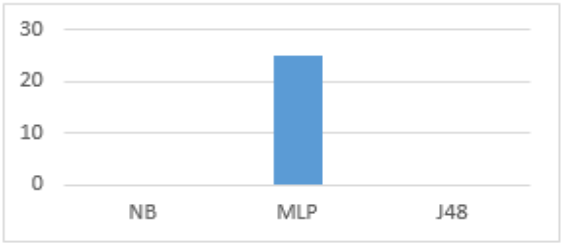

Figure 2. Building Time of Classifier Models

Figure 2 illustrates the building time of the three models under consideration. Amongst them, the MLP classifier took the longest time to build a model, whereas the NB and J48 classifiers learned to construct a model quickly for the given data. Figure 3 displays the correctly and incorrectly classified instances.

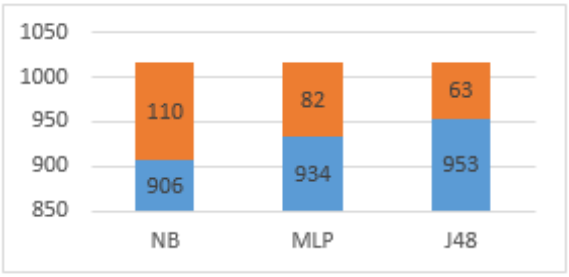

Figure 3. Error average

The performance of machine learning techniques is highly dependent upon the quality of the training data. Confusion matrices are extremely useful for evaluating the classifier models. A confusion matrix is a table used to define a classification model's output on a collection of training data for which the true values are known. In table 7 there are two possible predicted classes: "A" and "B". "A = yes" would mean they predicating the prevention of COVID-19 spread, and "B = no" would mean they do not . For example, The classifier made a total of 1017 predictions . Out of those 1017 cases, the classifier "NB" predicted "A" 981 times, and "B" 36 times. In reality, 998 respondents in the sample have the prevention, and 19 respondents do not. Briefly, the columns in table7 represent the predictions, whilst the rows indicate the final class.

.

Table 7. Confusion Matrices

Classifier Model	A	B	
NB	969	29	A
	12	7	B
MLP	992	6	A
	14	5	B
J48	983	15	A
	10	9	B

In general, the experiments proved that cross validation was statistically good in the classifier performance. Good results lead to large numbers and small down the main diagonal, perfectly zero off-diagonal elements. As shown in Table 7, the classifier algorithms (MLP, NB and J48) demonstrated good results, indicating that DM methods could effectively help in the prevention of COVID-19 spread.

Figure 4 demonstrates the tree of the J48 classifier algorithm, which was the best method to show the most connected questions to the final class. Each node represent a question and its branches are drawn in accordance with the answers.

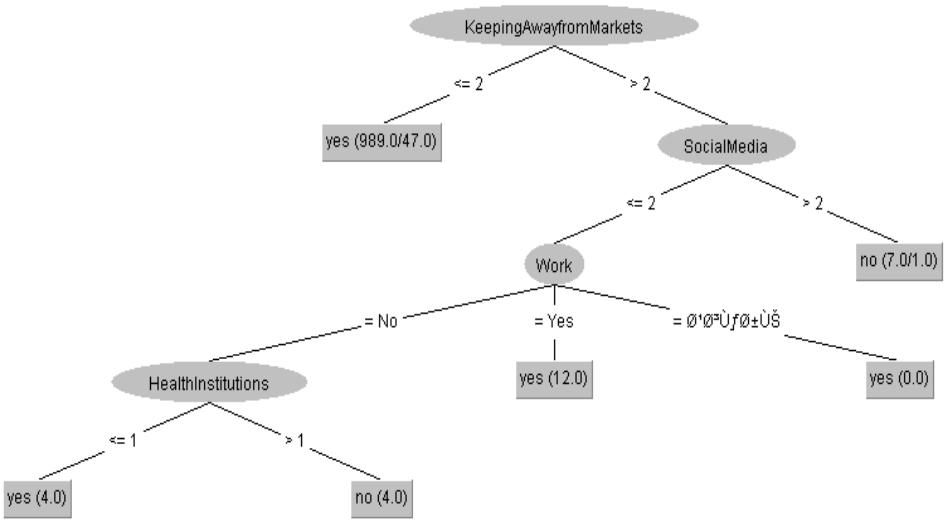

Figure 4. J48 Classifier Algorithm

5. Conclusion

In this paper, the problem of COVID-19 spread in Basra City was studied by applying three supervised DM algorithms to predict the prevention of COVID-19 spread. The performance of the learning methods was evaluated on the basis of their prediction accuracy. The results indicated that the J48 classifier outperformed the NB and MLP methods. A good classifier model must be accurate and comprehensible. This study was based on the questionnaires filled out by respondents from Basra City and the DM techniques were applied after the data were collected. This method could help medical departments by limiting the spread of the disease and reducing infection by taking necessary precautions in a timely manner and improve the quality of health institutions. For future works, the experiment could be expanded with more special attributes to obtain more accurate results. Many specific technologies could also be used when using different factors.

References

[1] Lauer, S.A., et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Annals of internal medicine, 2020. 172(9): p. 577-582.
[2] Arti, M. and K. Bhatnagar, Modeling and Predictions for COVID 19 Spread in India. ResearchGate, DOI: DOI. 10.
[3] Yu, H., et al., Reverse logistics network design for effective management of medical waste in epidemic outbreaks: Insights from the coronavirus disease 2019 (COVID-19) outbreak in Wuhan (China). International Journal of Environmental Research and Public Health, 2020. 17(5): p. 1770.
[4] Wu, H., et al., Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 2018. 10: p. 100-107.
[5] Amin, M.S., Y.K. Chiam, and K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease. Telematics and Informatics, 2019. 36: p. 82-93.
[6] Kostopoulos, G., et al. Predicting student performance in distance higher education using active learning. in International Conference on Engineering Applications of Neural Networks. 2017. Springer.
[7] Awadh, W.A., A.S. Hashim, and A.K. Hamoud, A REVIEW ON INTERNET OF THINGS ARCHITECTURE FOR BIG DATA PROCESSING. Iraqi Journal for Computers and Informatics, 2020. 46(1): p. 11-19.
[8] Cattral, R., F. Oppacher, and D. Deugo. Supervised and unsupervised data mining with an evolutionary algorithm. in Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546). 2001. IEEE.
[9] Murphy, K.P., Machine learning: a probabilistic perspective. 2012: MIT press.
[10] Abd Ulkareem, M., W.A. Awadh, and A.S. Alasady. A comparative study to obtain an adequate model in prediction of electricity requirements for a given future period. in 2018 International Conference on Engineering Technology and their Applications (IICETA). 2018. IEEE.
[11] Zhu, X. and A.B. Goldberg, Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, 2009. 3(1): p. 1-130.
[12]. Hamoud, A., et al., Students' success prediction based on Bayes algorithms. International Journal of Computer Applications, 2017. 178(7): p. 6-12.
[13] Wildani, I. and I. Yulita. Classifying Botnet Attack on Internet of Things Device Using Random Forest. in IOP Conference Series: Earth and Environmental Science. 2019. IOP Publishing.
[14] Panda, M., Developing an Efficient Text Pre-Processing Method with Sparse Generative Naive Bayes for Text Mining. International Journal of Modern Education & Computer Science, 2018. 10(9).
[15] Hassan, M.K., et al., EoT-driven hybrid ambient assisted living framework with naïve Bayes–firefly algorithm. Neural Computing and Applications, 2019. 31(5): p. 1275-1300.
[16] Heidari, A.A., et al., An efficient hybrid multilayer perceptron neural network with grasshopper optimization. Soft Computing, 2019. 23(17): p. 7941-7958.

[17]. Ploj, B., R. Harb, and M. Zorman, Border Pairs Method—constructive MLP learning classification algorithm. Neurocomputing, 2014. 126: p. 180-187.
[18] Janani, V., et al., Dengue Prediction Using (MLP) Multilayer Perceptron-A Machine Learning Approach. 2020, EasyChair.
[19] Jain, A., S. Sharma, and M.S. Sisodia, Network intrusion detection by using supervised and unsupervised machine learning techniques: a survey. International Journal of Computer Technology and Electronics Engineering, 2011. 1.
[20] Khan, A.N., et al. Learning from Privacy Preserved Encrypted Data on Cloud Through Supervised and Unsupervised Machine Learning. in 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). 2019. IEEE.
[21] Hong, H., et al., Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). Catena, 2018. 163: p. 399-413.
[22] Hamoud, A., A.S. Hashim, and W.A. Awadh, Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence, 2018. 5: p. 26-31.
[23] Kaur, R. and R. Gangwar, A Review on Naive Baye's (NB), J48 and K-Means Based Mining Algorithms for Medical Data Mining. Int. Res. J. Eng. Technol, 2017. 4: p. 1664-1668.
[24] Carson, B., The transformative power of action learning. Chief Learning Officer. Retrieved, 2017.
[25] Sekaran, U. and R. Bougie, Research methods for business: A skill building approach. 2016: John Wiley & Sons.
[26] Hashima, A.S., A.K. Hamoud, and W.A. Awadh, Analyzing students' answers using association rule mining based on feature selection. Journal of Southwest Jiaotong University, 2018. 53(5).