ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse118952020.pdf
https://doi.org/10.30534/ijatcse/2020/118952020

Prediction of Number of Participants in The Online Examination System at Universitas Terbuka

Iwan Susanto[1], Gede Putra Kusuma[2]
[1]Computer Science Department, BINUS Graduate Program - Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia, 11480, iwan.susanto@binus.ac.id
[2]Computer Science Department, BINUS Graduate Program - Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia,11480,inegara@binus.edu

ABSTRACT

The Online Examination System (SUO) conducted by the Universitas Terbuka (UT) is now one of the exams that is increasingly in demand by Universitas Terbuka students. The Online Examination System is implemented in all branches of the Open University in major cities in Indonesia. The Amount of Interest Students make UT need to predict how much interest is the registration of SUO, so that UT can prepare facilities and infrastructure so that the exam can run smoothly. Regression is a method that can be used in analyzing predictions, several regression models such as Multiple Linear Regression, Decision Tree, Support Vector Regression and Artificial Neural Network are used in this study. In addition to obtaining accurate predictions, in machine learning a feature selection is added to eliminate unrelated data and also duplicate data. Cities were grouped using the K-Means method. Decision tree model without feature selection has good performance with RMSE = 169.047 +/- 0.000 and R2 = 0.749. And for the ANN model combined with optimize selection (evolutionary) it has a good performance value of RMSE = 115.023 +/- 38,183 (micro average: 120,592 +/- 0.000) and R2 = 0.841 + / - 0.082 (micro average: 0.824) with several attributes related to the prediction process.

Key words: Number of Participants, Online Examination System, Feature Selection, Regression Model, Data Mining.

1. INTRODUCTION

At this time, IT technology has become more advanced, one of the proofs is that access to communications and information is easy with portable devices such as smartphones, tablets, or laptops. Current advances in IT technology have also penetrated the field of education, especially tertiary education, making it easy for students to study anywhere and anytime. Forums and discussions between students are also so easy with the presence of social media applications such as WhatsApp, Facebook or Instagram so that the awareness of students to get to know and learn new technology is higher if they do not want to be left behind with the technology itself.

Universitas Terbuka (UT) as one of the state universities in Indonesia whose student distribution is located throughout Indonesia both in big cities and in remote areas. UT as a higher education institution has conducted the Final Examination (UAS) in the form of written examinations since 1984 and has been carried out in all districts / cities throughout Indonesia. Over time, there are still some students who cannot take the UAS due to obstacles on the exam date, so students cannot take the exam, while the UAS is a major component in the grading requirements which has a weighting of 50% -100%, so that if students do not do the UAS , then automatically students will not pass the course and will add to their study period[1].

With the advancement of technology, UT is trying to provide opportunities for students to be more flexible in carrying out the final semester exams, so the Online Exam was developed under the name of the Online Exam System (SUO). SUO is held in classes that have provided computers for exams as many as 20 units of Computers. At each UT branch office, there are generally at least 1 Online Exam class with a maximum number of participants of 20 participants per session, in 1 day there are a maximum of 5 exam sessions.

With students getting accustomed to learning with the help of IT-based technology, for example with the help of computers and smartphones, the more open the desire of students to conduct IT technology-based online exams, SUO as a means of student examinations with the same weight value as the weighted scores on the UAS becomes an option alternative students in conducting exams, the advantage of SUO is the time that can be flexible to follow the choice of student time, the choice of time is not fully flexible, but the time range has been determined by UT in conducting the test, usually the time range provided by UT is 2-3 weeks, at That time range students can choose the date and time of the exam they can take.

Over time, the interest of students to take Online Exams increases, so that the capacity of online exam classes is insufficient, each UT branch office determines quota for online exam registration, this quota becomes a struggle between students and often many students do not get a quota.

This quota limitation is due to the lack of infrastructure prepared by the large amount of students' interest in conducting computer-based online examinations, coupled with the issue of limited exam time where exam time must be adjusted to the academic calendar of lectures, the maximum of online examinations is 1 month before grading grades. This limitation was finally known as the online exam registration quota. The need for additional SUO facility and infrastructure capacity is one of the management priorities in meeting the SUO registration quota, but this has not been decided by management, given the absence of data related to analysis and prediction of how many online exam registrants to match facilities and infrastructure which must be prepared by UT.

In this study, researchers will analyze online exam registration data from registration data 5 years earlier, namely the even semester of 2015 to odd semester of 2019, these data will be analyzed using Data Mining, namely Analysis of data to obtain new information by looking for certain patterns or rules from data that has been known before in large enough quantities**.**

2. RELATED WORKS

In determining predictions in data mining, several methods can be used to carry out the process of data analysis, in this study several data mining methods will be used that produce how large the prediction is successful and accurate, namely Linear Regression, Decision Tree, Support Vector Regression and Artificial Neural Network.In several studies, a comparison was made of how accurately the regression algorithm can predict the required results[2],[3]. Some of the use of linear regression as one of the most commonly used methods in making predictions is the study of Prediction of Stock Using Fuzzy Linear Regression[4] and also sales predictions for a bookstore using Simple Linear Regression[5]. In Linear Regression there are several types, namely Simple Linear Regression, Multiple Linear Regression, Logistic Regression, Ordinal Regression, Multinominal Regression, and Linear Discriminant Analysis. Multiple Linear Regression is a linear regression analysis used to analyze the relationship between a single response variable (the dependent variable) with two or more controlled variables (the independent variable)[6]. Multiple Linear Regression when used to predict trade balance estimates, it is found that there are several influential factors, namely palm oil exports, crude oil exports, imports of petroleum products, car imports, total landfill and plywood plain exports[7]. Whereas when used to get an estimate of the bid price in winning a procurement project contract, a MAPE (Mean Absolute Percentage Error) result of around 3% and a coefficient of determination R2 = 0.88167. the results of this study are better compared to previous studies that use traditional methods only produce about 25% MAPE, even with the latest research previously only get about 19% MAPE[8].

To predict how far the influence of individual factors and the urban environment on diabetes and asthma in a community can be determined using decision tree analysis. And what are the interrelationships between individual factors and environmental factors in predicting the influence of diabetes and asthma[9].In another study, the problem of removing body parts, especially the legs due to diabetes can be prevented by predicting early treatment using a neural network[10]. In a tertiary institution in Indonesia, there is a study which analyzes the possibility of resignation of prospective new students using the Decision Tree method called C4.5 algorithm[11]. Not only used in tertiary institutions, the C4.5 algorithm is also used in Predicting the Graduation Rate of Kintamani SMPN1 Students[12].

Based on the SUO registration data that has been used for 5 years, these data can be used as time series data to predict future participation, in previous studies, time series data were used to predict energy use in a community. using machine learning[13]. Besides machine learning is also used in making predictions in the world of health, namely to predict the quality and production of cow's milk if the cow has a history of mastitis (inflammation of the breast tissue)[14].

In fact, every data taken for analysis is data that must be understood first, not all data can be used as a basis for decision making, to be able to interpret the data, training is required Analysis of these data, then from the data selected or selected, which data has a major influence on the final result[15].

3. METHODOLOGY

With the development of technology and the increasing dependence of organizations and society on the use of the internet, resulting in the growth and variety of data that is increasingly large. The growth and variation of this large data becomes a challenge in conducting business data analysis, to help carry out the Big Data process, we need a framework that can assess the performance of several data mining tools, such as RapidMiner[16],[17],[18]. Several studies have conducted benchmarking on data mining tools for large data such as RapidMiner and KNIME.[17] Even in some studies conducted a review of some data mining tools namely WEKA, Rapidminer, Tanagra, Orange and Knime[19].

This study uses quite large data, where the data will be selected for related variables and discarding unrelated variables, the use of RapidMiner is very helpful in the process of analyzing data using several data mining methods, in the world of Education , some studies have used a combination of

RapidMiner with several data mining methods[16]. In determining stock price predictions, RapidMiner can be used and has good results[20].

In Rapid Miner, the optimize selection method is used to get the best combination of free variables and has a correlation with the dependent variable. Several Optimize selection models are used:

Forward Selection: in this model, each independent variable will be entered gradually to get the variable that has the closest correlation with the dependent variable. This process will stop until there are no more potential independent variables.

Backward Elimination: In contrast to the forward selection model, all dependent variables will be entered first, then they will be issued one by one gradually by testing the parameters.

Optimize Selection (Evolutionary): Each attribute subset is considered as an individual, the evolution algorithm will select the individual attributes to be processed independently and mutated or by crossing with other attributes. For feature selection, mutations can be activated or deactivated. The Optimize Selection Evolutionary strategy is implemented with a genetic algorithm.

In this paper, a simulation is made of the predictions of online test takers using several methods:

Datasets that have been separated between training data and testing data will be compared using the Regression data mining model (Linear Regression, Decision Tree, Support Vector Machine and Artificial Neural networks) combined with feature selection methods (Forward Selection, Backward Elimination and Evolutionary) and without feature selection. Then comparing the results of the regression models by considering the RMSE and R2 values. It is expected that the best regression models that can be applied to predictions of online exam participants will be seen.

3.1 Data Preparation and Pre-Processing

At this stage, the researcher will collect student registration data that has conducted SUO registration for the last 5 years, namely the even semester of 2015 until the odd semester of 2019. In this process, researchers will take all the data of SUO applicant students from various Study Programs and from various Provincial Cities in Indonesia, which has an Open University campus, there are 38 cities in the province that will be recorded, the complete list are shown in Table 1.

Table 1: List of Cities

No.	City	No.	City	No.	City
1	Aceh	14	Bandung	27	Batam
2	Medan	15	Semarang	28	Tarakan
3	Padang	16	Yogyakarta	29	Makassar
4	Jambi	17	Surakarta	30	Majene
5	Bengkulu	18	Surabaya	31	Kendari
6	Pekanbaru	19	Malang	32	Manado
7	Palembang	20	Jember	33	Palu
8	PangkalPinang	21	Denpasar	34	Gorontalo
9	Palangkaraya	22	Mataram	35	Ambon
10	Lampung	23	Kupang	36	Ternate
11	Serang	24	Pontianak	37	Jayapura
12	Jakarta	25	Samarinda	38	Sorong
13	Bogor	26	Banjarmasin		

The total data taken was 349 rows, from this data the researcher determined the distribution of 60% training data by 210 rows, 20% validation data by 70 rows and 20% testing data by 69 rows. The method that will be used when testing is linear regression, Decision Tree, Support Vector Regression (SVR) and Artificial Neural Network (ANN) and optimization is done using feature selection (Backward, forward, exhaustive) then the performance metric data adopted are RMSE and R^2 . Experiments carried out on RapidMiner 9.6. Variable data as shown in the Table 2.

Table 2: List of Attributes

No.	Attributes
1	Semester
2	City
3	Number of SUO Registrants
4	Total Number of Students
5	Number of marital status students
6	Number of male students
7	Number of High School Graduates
8	Number of Suo Courses
9	Total number of courses
10	City Cluster

3.2 Implementation of Data Mining

To predict the number of online exam participants, the data is grouped by city type so that the predicted value can reflect the number of participants based on the student city. City data is grouped to see the effect of cities in Indonesia on the interests of participants from the city. Grouping data using the K-Means method and the results of grouping [21] are evaluated using the Silhouette Coefficient, where a value close to 1 is the optimal grouping[22],[23].From the results of testing with Silhoutte Coefficient, the data obtained by grouping as many as 6 clusters is more optimal when compared to grouping 3,4,5 clusters, even in the 7th cluster the value of Silhoutte Coefficient goes down again. Comparative data as shown in Table 3.

Table 3: Value of Silhoutte Coefficient

Cluster	SilhoutteCoeficient
Cluster 3	0.760428044
Cluster 4	0.855487096
Cluster 5	0.895528752
Cluster 6	**0.897247391**
Cluster 7	0.883239085

4. RESULTS

By obtaining city clusters, in the data mining process, attribute cities are removed and replaced with city clusters, then each regression model process either using feature selection or without feature selection, an optimize parameter module is added, and to evaluate the performance of the model or algorithm, it is used cross validation with 10 fold. The performance value of the regression data mining model that has been processed with feature selection and without feature selection as shown in the Table 4 to 11.

4.1 Result of Regression Model without Feature Selection

Table 4: Result of Regression Model without Feature Selection (Training)

No.	Attributes	RMSE	R^2
1	Decision Tree	154.698 +/- 54.964 (micro average: 163.249 +/- 0.000)	0.716 +/- 0.137 (micro average: 0.690)
2	Linear Regression	119.269 +/- 46.750 (micro average: 127.248 +/- 0.000)	0.825 +/- 0.085 (micro average: 0.799)
3	SVM	259.549 +/- 86.916 (micro average: 272.332 +/- 0.000)	0.541 +/- 0.087 (micro average: 0.141)
4	ANN	118.538 +/- 37.656 (micro average: 123.804 +/- 0.000)	0.827 +/- 0.080 (micro average: 0.811)

Table 5: Result of Regression Model without Feature Selection (Testing)

No.	Attributes	RMSE	R^2
1	Decision Tree	169.047 +/- 0.000	0.749
2	Linear Regression	207.387 +/- 0.000	0.643
3	SVM	309.305 +/- 0.000	0.498
4	ANN	214.688 +/- 0.000	0.611

Table 4 shows the results of the regression model training process without using feature selection, the ANN model has the best results with values root_mean_squared_error: 118.538 +/- 37.656 (micro average: 123.804 +/- 0.000) And squared_correlation: 0.827 +/- 0.080 (micro average: 0.811)with optimized parameters are Neural Net.momentum = 0.8, Neural Net.training_cycles = 41. However, in the testing process, the decision tree model has the best value, shown in Table 5, with the following values:root_mean_squared_error: 169.047 +/- 0.000 and squared_correlation: 0.749.

4.2 Result of Regression Model with Forward Selection

Table 6: Result of Regression Model withForward Selection (Training)

No.	Attributes	RMSE	R^2
1	Decision Tree	136.089 +/- 49.810 (micro average: 144.059 +/- 0.000)	0.769 +/- 0.112 (micro average: 0.743)
2	Linear Regression	121.198 +/- 41.652 (micro average: 127.477 +/- 0.000)	0.819 +/- 0.085 (micro average: 0.799)
3	SVM	237.772 +/- 70.154 (micro average: 246.911 +/- 0.000)	0.512 +/- 0.145 (micro average: 0.341)
4	ANN	117.127 +/- 34.965 (micro average: 121.733 +/- 0.000)	0.831 +/- 0.086 (micro average: 0.816)

Table 7: Result of Regression Model with Forward Selection (Testing)

No.	Attributes	RMSE	R^2
1	Decision Tree	237.197 +/- 0.000	0.514
2	Linear Regression	233.062 +/- 0.000	0.546
3	SVM	288.069 +/- 0.000	0.514
4	ANN	227.609 +/- 0.000	0.590

Table 6 shows the results of the regression model training and testing process using forward selection, the ANN model has the best results with valuesroot_mean_squared_error: 117.127 +/- 34.965 (micro average: 121.733 +/- 0.000) and squared_correlation: 0.831 +/- 0.086 (micro average: 0.816)with optimized parameters areNeural Net.training_cycles = 90 , Neural Net.momentum = 0.7. And the testing value as shown in Table 7 are root_mean_squared_error: 227.609 +/- 0.000 and squared_correlation: 0.590.

4.3 Result of Regression Model with Backward Elimination.

Table 8: Result of Regression Model with Backward Elimination (Training)

No.	Attributes	RMSE	R^2
1	Decision Tree	148.071 +/- 69.163 (micro average: 161.957 +/- 0.000)	0.755 +/- 0.151 (micro average: 0.693)
2	Linear Regression	119.269 +/- 46.750 (micro average: 127.248 +/- 0.000)	0.825 +/- 0.085 (micro average: 0.799)
3	SVM	257.921 +/- 85.589 (micro average: 270.400 +/- 0.000)	0.510 +/- 0.115 (micro average: 0.343)
4	ANN	115.880 +/- 35.535 (micro average: 120.684 +/- 0.000)	0.831 +/- 0.089 (micro average: 0.820)

Table 9: Result of Regression Model with Backward Elimination (Testing)

No.	Attributes	RMSE	R^2
1	Decision Tree	184.463 +/- 0.000	0.698
2	Linear Regression	207.387 +/- 0.000	0.643
3	SVM	298.509 +/- 0.000	0.481
4	ANN	214.274 +/- 0.000	0.671

Table 8 shows the results of the regression model training process using backward elimination, the ANN model has the best results with valuesroot_mean_squared_error: 115.880 +/- 35.535 (micro average: 120.684 +/- 0.000) and squared_correlation: 0.831 +/- 0.089 (micro average: 0.820)with optimized parameters are Neural Net.training_cycles = 100 and Neural Net.momentum = 0.7. However, in the testing process, the decision tree model has the best value, as shown in Table 9, with the following values root_mean_squared_error: 184.463 +/- 0.000 and squared_correlation: 0.698.

4.4 Result of Regression Model with Optimize Selection (Evolutionary).

Table 10: Result of Regression Model withOptimize Selection Evolutionary(Training)

No.	Attributes	RMSE	R^2
1	Decision Tree	140.557 +/- 47.477 (micro average: 147.597 +/- 0.000)	0.761 +/- 0.116 (micro average: 0.745)
2	Linear Regression	122.206 +/- 46.765 (micro average: 130.010 +/- 0.000)	0.822 +/- 0.067 (micro average: 0.791)
3	SVM	244.443 +/- 72.083 (micro average: 253.828 +/- 0.000)	0.507 +/- 0.126 (micro average: 0.284)
4	ANN	115.023 +/- 38.183 (micro average: 120.592 +/- 0.000)	0.841 +/- 0.082 (micro average: 0.824)

Table 11: Result of Regression Model with Optimize Selection Evolutionary (Testing)

No.	Attributes	RMSE	R^2
1	Decision Tree	181.677 +/- 0.000	0.708
2	Linear Regression	208.221 +/- 0.000	0.640
3	SVM	301.650 +/- 0.000	0.386
4	ANN	208.152 +/- 0.000	0.646

Table 10 shows the results of the regression model training process using optimize selection (evolutionary), the ANN model has the best results with values root_mean_squared_error: 115.023 +/- 38.183 (micro average: 120.592 +/- 0.000) and squared_correlation: 0.841 +/- 0.082 (micro average: 0.824)with optimized parameters are Neural Net.training_cycles = 100 and Neural Net.momentum = 0.8. However, in the testing process, the decision tree model has the best value, as shown in Table 11, with the following values root_mean_squared_error: 208.152 +/- 0.000 and squared_correlation: 0.646.

5. CONCLUSION

attribute	weight
City Cluster = cluster_0	1
City Cluster = cluster_1	1
City Cluster = cluster_2	1
City Cluster = cluster_3	1
City Cluster = cluster_4	1
City Cluster = cluster_5	1
Number of High School Graduates	1
Number of Suo Courses	1
Number of male students	1
Number of marital status students	0
Semester = GANJIL	0
Semester = GENAP	0
Total Number of Students	0
Total number of courses	0

Figure 1: List of Relevant Attributes

Based on the results of performance data mining regression, it can be seen that when the data is trained, the ANN method using optimize selection (evolutionary) has a good value RMSE = 115.023 +/- 38.183 (micro average: 120.592 +/- 0.000) and R^2 =0.841 +/- 0.082 (micro average: 0.824), but after testing, the decision tree method without feature selection has a better value ,RMSE=169.047 +/- 0.000 and R^2= 0.749.The results using optimize selection (evolutionary), there are several attributes that are more dominant compared to other attributes, this dominant attribute will affect the prediction results of online exam participants in Universitas Terbuka. The relevant attributes are shown in Figure 1.

REFERENCES

[1] T. Pardede and S. Listyarini, "**Sistem Ujian Online Sebagai Upaya Peningkatan pelaksanaan Ujian Dalam Pendidikan Terbuka Jarak Jauh**," *J. Pendidik. Terbuka Dan Jarak Jauh*, no. 2017/8/29, pp. 19–30, 2010.

[2] A. Dhankhar, K. Solanki, and A. Rathee, "**Predicting Student's Performance by using Classification Methods**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 1532–1536, 2019.

[3] O. Dahiya, K. Solanki, and S. Dalal, "**Comparative analysis of regression test case prioritization techniques**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 1521–1531, 2019, doi: 10.30534/ijatcse/2019/74842019.

[4] M. Misdiyanto, "**Prediksi Pesediaan Stock Pulsa Menggunakan Fuzzy Linear Regresion**," *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 3, no. 1, p. 29, 2019, doi: 10.30645/j-sakti.v3i1.93.

[5] R. Kamal, Iman Mustofa Hendro P, Tachbir Ilyas, "**Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya**," *Semnasteknomedia Online*, vol. II, no. 1, pp. 49–54, 2017, [Online]. Available: http://ojs.amikom.ac.id/index.php/semnasteknomedia/article/view/1712.

[6] I. M. M. Ghani and S. Ahmad, "**Stepwise multiple regression method to forecast fish landing**," *Procedia - Soc. Behav. Sci.*, vol. 8, no. 5, pp. 549–554, 2010, doi: 10.1016/j.sbspro.2010.12.076.

[7] S. Omar Gan and S. Ahmad, "**Multiple Linear Regression to Forecast Balance of Trade**," *Malaysian J. Fundam. Appl. Sci.*, vol. 7, no. 2, pp. 1–7, 2014, doi: 10.11113/mjfas.v7n2.255.

[8] P. Aleksandar, P. Silvana, and Z. Valentina, "**Multiple Linear regression model for predicting bidding price**," no. October, 2015, [Online]. Available: https://www.researchgate.net/publication/282646773_Multiple_Linear_regression_model_for_predicting_bidding_price.

[9] H. A. Cuesta, D. L. Coffman, C. Branas, and H. M. Murphy, "**Using decision trees to understand the influence of individual- and neighborhood-level factors on urban diabetes and asthma**," *Heal. Place*, vol. 58, no. November 2018, p. 102119, 2019, doi: 10.1016/j.healthplace.2019.04.009.

[10] P. Santhi, P. Santhi, N. Deeban, N. Jeyapunitha, B. Muthukumaran, and R. Ravikumar, "**Prediction of Diabetes using Neural Networks**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 985–990, 2020.

[11] Andie, "**Penerapan Decision Tree Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru**," *Technologia*, vol. 7, no. 1, pp. 8–14, 2016.

[12] G. Indrawan, "**Penerapan Metode Decision Tree (Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1**," pp. 35–44, 2016.

[13] J. S. Chou and D. S. Tran, "**Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders**," *Energy*, vol. 165, pp. 709–726, 2018, doi: 10.1016/j.energy.2018.09.144.

[14] M. Ebrahimi, M. Mohammadi-Dehcheshmeh, E. Ebrahimie, and K. R. Petrovski, "**Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Trees outperform other models**," *Comput. Biol. Med.*, vol. 114, no. May, p. 103456, 2019, doi: 10.1016/j.compbiomed.2019.103456.

[15] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, "**Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection**,"

Informatics Med. Unlocked, vol. 17, no. August, p. 100255, 2019, doi: 10.1016/j.imu.2019.100255.

[16] J. Silva, L. Hernandez, T. Crissien, O. B. Pineda Lezama, and J. Romero, "**Big Data Aplication for Selecting Theses Topics**," *Procedia Comput. Sci.*, vol. 160, pp. 538–542, 2019, doi: 10.1016/j.procs.2019.11.051.

[17] C. Oliveira, T. Guimaraes, F. Portela, and M. Santos, "**Benchmarking Business Analytics Techniques in Big Data**," *Procedia Comput. Sci.*, vol. 160, pp. 690–695, 2019, doi: 10.1016/j.procs.2019.11.026.

[18] H. Bagheri and A. A. Shaltooki, "**Big data: Challenges, opportunities and cloud based solutions**," *Int. J. Electr. Comput. Eng.*, vol. 5, no. 2, pp. 340–343, 2015, doi: 10.11591/ijece.v5i2.pp340-343.

[19] A. Naik and L. Samant, "**Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime**," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

[20] T. N. A. et al. R. Chowdhury, M.R.C. Mahdy, "**Predicting the stock price of frontier markets using modified Black–Scholes Option pricing model and machine learning**," *J. Comput. Appl. Math.*, p. 112822, 2020, doi: 10.1016/j.cam.2020.112822.

[21] J. Cai, J. Luo, S. Wang, and S. Yang, "**Feature selection in machine learning: A new perspective**," *Neurocomputing*, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[22] shima kashef and H. Nezamabadi-pour, "**MLIFT: enhancing multi-label classifier with ensemble feature selection**," *J. AI Data Min.*, vol. 0, no. 0, pp. 355–365, 2018, doi: 10.22044/jadm.2018.5780.1688.

[23] K. Polat and S. Güneş, "**A new feature selection method on classification of medical datasets: Kernel F-score feature selection**," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10367–10373, 2009, doi: 10.1016/j.eswa.2009.01.041.