Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Analysis Bidding Price Analysis Using K-Means Cluster, Machine Learning, And Deep Learning in Public Institution Competitive Bidding: Targeting Service Projects in The Chungcheong Region

Young-Hun Kim[1], Gitae Kim[2]

[1] *Department Of Smart Production & Management Engineering, Hanbat National University, Daejeon, Republic of Korea*
[2] *professor, Department of Industrial & Management Engineering, Hanbat National University, Daejeon, Republic of Korea*

The probability of participating in a public institution bidding, which is a burden of expenditure, from winning the bid to signing a contract is quite low. In addition, because the conclusion of a contract has a significant impact on the company's management, bidders are under a lot of stress regarding bidding. Therefore, in this paper, a successful bid price analysis study was conducted to enable bidders to easily predict the bid price. Clusters were analyzed using the big data analysis algorithm K-means Cluster, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, and K-Nearest. The accuracy of the results was analyzed using Neighbors Regressor, Multi-layer Perceptron Regressor, etc. To this end, a price-to-value analysis was performed based on the last three years of service bidding results data in the Chungcheong region.

Keywords: Bidding price, Competitive bidding, Clustering, Machine learning, Deep learning

1. Introduction

With the introduction of the e-Government system in the 2000s, electronic bidding began through the Public Procurement Service. When public institutions make announcements through competitive bidding in accordance with contract laws, they must announce competitive bidding through the Public Procurement Service's National Marketplace. In order to participate in competitive bidding, two of the 15 multiple reserve prices consisting of +7 and -8 are selected, and when the bidding process is conducted, the expected price is determined by taking the arithmetic average of the four most frequently selected multiple reserve prices. Before the 2000s, information related to bidding was obtained through public institution bidding information sheets, but since the 2000s, information related to bidding has been collected from the Public Procurement Service or bidding-related sites. Services provided by bidding-related sites include information on the Public Procurement Service opening bid results, information on the number of companies participating in the bidding, information on the expected price, information on multiple reserve prices, and data on past bidding results. Based on the information provided, bidding participants predict the bidding price by selecting a section where the expected price is high, selecting a section where the number of bidders is low, and selecting a section where the expected price is high and where there are few bidders. Kang Min-seok (2014), who studied the impact of the multiple reserve price creation range and multiple reserve price section setting method on the scheduled price based on competitive bidding standards, discussed the multiple reserve price creation range, multiple reserve price creation section, and multiple reserve price lottery method, studied[1]. In this study, it was found that in determining the expected price for a successful bid, it was most reasonable to select the multiple reserve price lottery method as the top two and the bottom two as the unequal division method. Kim Cheol (2017) studied the factors affecting the public IT project success rate prediction model and used linear regression, support vector machine regression, and random forest in an R program to create a success rate prediction model using public IT success characteristic data Developed [2]. In this study, variables that affect the bid success rate were identified and a plan for utilizing them for public IT bidding was presented. Daeseong Hwang (2020) studied bidding price prediction using deep learning in electronic bidding using bidding result data for electrical construction using Machine Learning and Deep Learning methods [3]. In this study, training and test data were split 7:3, and the prediction accuracy in the Deep Learning algorithm was high. Previous papers are papers on the creation range of multiple reserve prices, multiple reserve price creation sections, and the method of setting multiple reserve prices to the top two and bottom two, and bidding using bidding information data from the information and communication industry with a machine learning algorithm to Price prediction accuracy was analyzed. And using the bidding result data for electrical work, the accuracy of the bidding results was predicted using Machine Learnig and Deep Learning methods. This study used bid result data for the service industry to analyze similar characteristics by dividing them into the same clusters using K-mearns Cluster, and analyzed the accuracy of successful bid price prediction for the service industry using Machine Learnig and Deep Learning. It can be said that it is different from existing previous papers. Projects ordered by public institutions have good cash flow, so there are many businesses participating in competitive bidding, and the probability of participating in the bidding and winning the bid is quite low. Additionally, the performance of the bidding has a significant impact on the company's management. As a result, the company's management puts psychological pressure on the bidding performance on the bidding manager, and the burden of bidding develops into stress on the bidding manager. For this study, 32,813 bid result data for service industries ordered in the Chungcheong region were collected from January 1, 2019 to October 15, 2023. This data was preprocessed and K-means clustering was performed on 18,441 data. Similar characteristics of the data were analyzed using an algorithm and divided into 9 clusters. And with 17,925 pieces of data, the accuracy of bid price prediction was analyzed using Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, K-Nearst Neighbors Regressor, and Multi-layer Perceptron Regressor. As a result, Random Forest Regressor was the most accurate, with an accuracy of over 99%. As a result of this study, the company's management status will be improved and the bidding manager's work stress regarding bidding will be reduced.

2. Literature Review

2.1 Public bidding

In order to ensure fairness, bids that impose a burden on financial expenditure are subject to competitive bidding in accordance with the National Contract Act and the Local Contract Act. At this time, from bid notice to bidding and contract, you must go through the competitive bidding process in accordance with the National Contract Act or Local Contract Act. 15 multiple preliminary prices must be prepared and posted at least 5 days before the opening date in accordance with the procedures set forth in the government contract regulations, “Standards for Preparing Prices” and “Standards for Local Government Bidding and Contract Execution,” and businesses participating in the bidding must submit them before the opening date. You must participate in the bidding after meeting the qualifications according to the relevant laws and regulations[4][5]. To participate in the bidding, register on the Public Procurement Service's National Marketplace, access the virtualized site of the National Marketplace, and select two out of 15 multiple reserve prices. The ordering organization registers as a user on the Nara Marketplace and then proceeds to bid. At this time, the expected amount is calculated by taking the arithmetic average of the four most frequent multiple reserve prices. And after the opening of bids is completed, priority in qualification review will be given to the business that participated in the bidding closest to the scheduled price. The contract partner will be determined after screening the first-ranked company in accordance with the Ministry of the Interior and Safety’s contract regulations, “Standards for determining successful bidders when bidding for local governments”[6][7]. When calculating 4 out of 15 multiple reserve prices, the expected price is 1365, and the formula is as in (1).

$$nCr = \frac{nPr}{r!} -------------------------(1)$$

The lower limit of the expected price in competitive bidding is limited. In order to minimize problems caused by poor design due to receiving low-priced orders, the lower limit of the expected price is applied at the time of bidding, and the formula is as follows (2) and (3).

$$Grade = 80 - 20 \left| \left(\frac{88}{100} - \frac{Bid\ Price}{Expected\ Price} \right) \times 100 \right| ---------- (2)$$

$$Grade = 90 - 20 \left| \left(\frac{88}{100} - \frac{Bid\ Price}{Expected\ Price} \right) \times 100 \right| ---------- (3)$$

Businesses participating in competitive bidding must prepare the limit amount according to formula (2) in the case of a service bid with a basic amount of less than 100 million won among 1,365 scheduled prices, and (3) in the case of a service bid with a basic amount of less than 200 million won. Prepare a bid with the limited amount according to the number formula and submit the bid through a cadastral information processing device.

2.2 K-means Clustering

Unsupervised learning is used to explain the distribution of data, extract key features of data, or model the distribution of data. Representative methods include principal component analysis (PCA), linear discriminant analysis (LDA), and clustering. The second is a method of creating a new representation of data, and representative methods include artificial neural network (ANN), One-Hot Encoding, and Text Summarization [8][9]. K-means clustering, a type of cluster analysis, classifies the acquired data into k clusters. Each cluster is defined based on the similarity of the data, and the data within the clusters are similar to each other, and the data between the clusters are not similar to each other [10]. Groups are formed based on data characteristics, clustered based on the central point of the group, and data with similar characteristics are grouped to help recognize patterns [11].

2.3 Machine Learning Model and Deep Learnning Model

Decision Tree Regressor, a type of machine learning, is an algorithm that can predict the final dependent variable by composing a complex decision-making process into a combination of several simple decisions. It is a non-parametric model used for both regression and classification [12]. Random Forest was proposed by Breiman in 2001, and is a model that generates bootstrap sampling based on a decision tree and trains the same algorithm multiple times by extracting multiple subsets from the entire training dataset [13]. Support Vector Regressor is a type of classification algorithm that is a generalized model of Support Vector Machine. It is a model that predicts data by finding the optimal hyperplane that includes as much data as possible within the distance between support vectors [14],[15]. K-Nearest Neighbors Regressor is a model that operates by checking k Lables located at the closest distance from the value Fast and easy to use [16],[17]. Multi-layer Perceptron Regressor is a type of Deep Learning. It is a feedforward neural network with one or more hidden layers between the input layer and the output layer. The hidden layer is an acceptable continuous function and is a model that is advantageous for finding approximations and calculating weights [18],[19][20][21][21][22][23].

3. Proposed Work

When 32,813 data from three years of service industry bidding results in the Chungcheong region were preprocessed and 18,441 data were analyzed using K-means clustering, they were clustered into 9 groups as shown in Table 1.

cluster	1	2	3	4	5	6	7	8	9	Sum
Count	968	3,032	3,435	3,409	2,341	1,707	1,228	1,152	1,169	18,441

Table 1. K-means clustering data grouping

Table 2 shows the distribution range, average, and number of data of the expected price by amount, and the ratio of dividing the lowest price for the range by group was recorded in the expected price range. It was found that the scheduled price was distributed within a range of ±3% according to the government's contract regulations. Since the range from 10 million won to 20 million won is generally conducted through voluntary bidding, contracts are concluded in the range of 90% of the basic amount. Therefore, the expected price range was found to be relatively high. The average range of groups 2 to 9 is 0.8783 to 0.8768, and it can be seen that the expected price is slightly biased toward the negative sign.

Division	amount(won)	range	average	Count	Expected price range	lower limit price
1 group	10,000,000~20,000,000	0.87378~0.90356	0.88841	968	99.2931~102.6773	88%
2 group	20,000,000~30,000,000	0.86998~0.88688	0.87830	3032	98.8613~100.7818	88%
3 group	30,000,000~40,000,000	0.86984~0.88595	0.87798	3435	98.8454~100.6761	88%
4 group	40,000,000~50,000,000	0.86992~0.88596	0.87793	3409	98.8545~100.6773	88%
5 group	50,000,000~60,000,000	0.86819~0.88548	0.87730	2341	98.6579~100.6227	88%
6 group	60,000,000~70,000,000	0.86710~0.88504	0.87689	1707	98.5340~100.5727	88%
7 group	70,000,000~80,000,000	0.86563~0.88460	0.87621	1228	98.3670~100.5227	88%
8 group	80,000,000~90,000,000	0.86651~0.88502	0.87664	1152	98.4670~100.5705	88%
9 group	90,000,000~100,000,000	0.86705~0.88497	0.87689	1199	98.5284~100.5648	88%

Table 2. Expected price range, average, and lowest bid price for each group

Fig 2. is data visualizing the analysis results of the expected price by group. In Group 1, contracts were often carried out through private contracts or private bidding, and the number of companies that could participate in bidding was limited, resulting in a high expected price. It can be seen that groups 2 to 9 are widely distributed within a limited range. This indicates that the expected price shows a normal distribution. In order to apply an algorithm for predicting the bid price, it is possible to analyze the trend of the expected price distribution according to the probability distribution, and the section of the expected price that was widely distributed in the past can be analyzed recently. If the distribution is low at the time, it can be predicted that the probability of the scheduled price appearing in the range of that group will increase.

Fig 2. Visualization of expected price analysis results by group

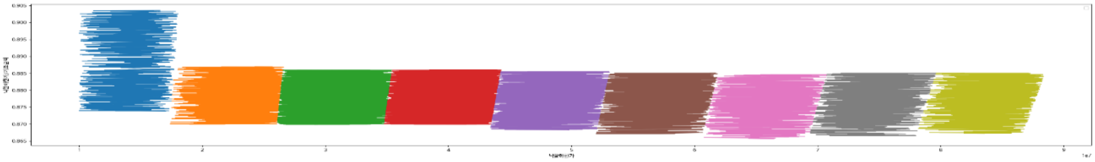

Table. 3 shows an example of preprocessed data of competitive bidding results.

Data	Basic amount	Expected price	EP/BA (100%)	EP/BA (0%)	1st priority Bid amount	1st priority Assessment rate	1st place basic preparation	Number of companies
5163	32,066,000	31,9015,900	99.5319	-0.4680	28,099,530	99.5798	87.6302	12
21251	74,015,000	73,398,150	99.1665	-0.8334	64,637,299	99.2386	87.3299	144
3965	51,664,000	52,212,850	101.0623	1.0623	46,548,000	102.3835	90.0975	3
21234	41,800,000	41,411,850	99.0714	-0.9285	36,452,940	99.0999	87.2079	46
19470	50,000,000	49,0971,250	98.1425	-1.8575	43,677,800	99.2677	87.3556	17

Tabe 4. includes 9 variables (columns) of 17,925 pre-processed bid result data, and the real number data (float64) of each variable is divided into estimated price, base price, scheduled price, expected price/basic price (100%), Analysis of trends in bidding price changes in the Chungcheong region using variables such as expected price/basic price (0%), 1st priority

bidding amount, 1st priority evaluation rate (100%), 1st priority basis comparison, and number of companies, estimated price and basic price , This is information that analyzes the comparative analysis of the expected price, the relationship between the first bid amount and the basic price, and the relationship between the number of participating companies and the bid price. To prevent overfitting of the model and improve generalization performance, the data split ratio was set at 80% of the total data as training data and 20% as test data. Table 5 shows the analysis results for each model, and the Random Forest model showed the highest accuracy and generalization performance. The Decision Tree model showed high accuracy but had a possibility of overfitting, the SVR model had a high possibility of underfitting, and the KNN model was sensitive to data size. MLP required a lot of data and learning time to increase data accuracy. It was analyzed that the Random Forest Model, which generally shows high accuracy and generalization performance, is most suitable for the learning purpose of predicting the characteristics of bidding result data and bidding price.

```
<class ‘pandas.frame.dataFrame’>
Int64Index: 17925 entries, 5163 to 15795
Data columns (total 9  columns):
#     Column               Non-Null Count   Dtype
---   -----------           -------------------  ---------
0  Estimated price            17925 non-null   float64
1  Basic amount               17925 non-null   float64
2  Expected price             17925 non-null   float64
3  EP/BA(100%)                17925 non-null   float64
4  EP/BA(0%)                  17925 non-null   float64
5  1st priority Bid amount       17925 non-null   float64
6  1st priority Assessemnt rate   17925 non-null   float64
7  1st place basic preparation    17925 non-null   float64
8  Number of companies        17925 non-null   float64
Dtypes: float64(9)
Momory usage: 1.4MB

Train_size : test_size = 8 : 2
```

Table 4. Data Information

Model	MSE	R2 Score	Study Time(ms)	Explanation
Decision Tree	95949447797.46814	0.999745	309.757	High accuracy (R2 Score), but possible overfitting
Random Forest	67120411441.58901	0.999822	259,076	Higher accuracy and generalization performance than DT
SVR	389033901300897.44	-0.032547	-	Sensitive to data distribution and possible underfitting
KNN	230768234718.29102	0.999388	-	Sensitive to data size and high computational cost
MLP	396074769078.2721	0.998949	480,383	Requires a lot of data and learning time for high accuracy

Table 5. Data example

When performing Parameter Grid Search with Random Forest Model, the combination of max_depth=8, min_samples_leaf=8, min_samples_Split=8, and n_estimators=100 was selected as the optimal hyperparameter. The prediction accuracy was very high at 0.9997 based on R2 Score when optimal hyperparameters were used. The RMSE of the training data was

251,927.41, and the RMSE of the test data was 305,923.92. There is a possibility of overfitting of the RMSE of the test data, but it was not considered a major problem because the prediction accuracy of the model was very high. < Table 6. Reference >. Fig 3 and Fig 4 are the learning curves of training and test data RMSE.

```
                    # Parameter definition for grid search
Param_grid = {
'n_estimators': [10, 50, 100, 200],
'max_depth': [None, 5, 8, 10, 20],
'min_samples_split': [2, 5, 8, 10],
'min_sam;le_leaf': [1, 2, 4, 8, 16]
}
Opimal parameters:
{'max_depth': 8 'min_samples_left': 8 'min_samples-split' : 8 'n_estimators': 100}
Optimal prediction accuracy: 0.9997

Training data RMSE : 251927.41295753248
Test data RMSE : 305923.9244689635
```

Table 6. Random Forest Grid Search

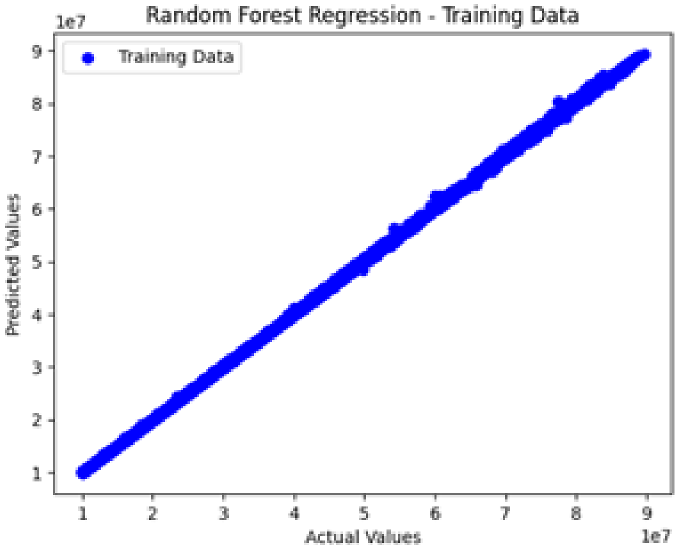

Fig 3. Trannig Data learning curve

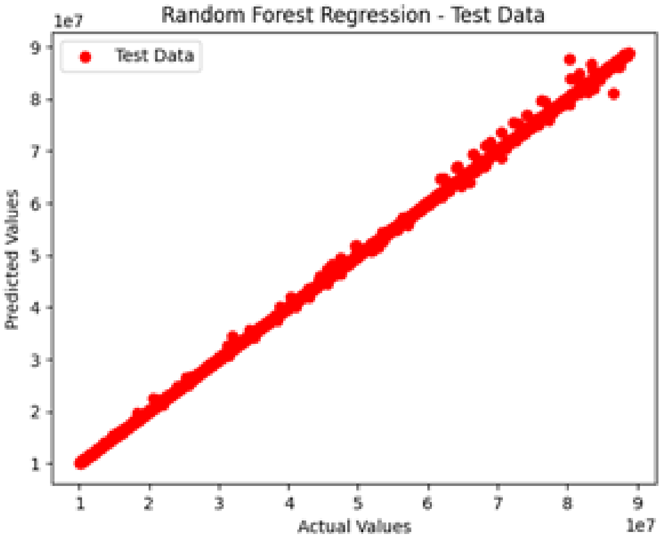

Fig 4. Test Data learning curve

4. Conclusion

This study used bidding result data from the Chungcheong region, analyzed the similarity of the data into 9 groups using K-means Cluster, analyzed it into 9 clusters, and visualized it. Through this cluster, it was found that the minimum floor rate of the bid price varies depending on the basic amount. In addition, in order to predict the accuracy of bidding results and bidding price, the accuracy of bidding results and analysis is measured using Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, K-Nearst Neighbors Regressor which are types of machine learning, and the accuracy of bidding results and analysis if measured using Multi-layer Perceptron Regressor which is a type of Deep Learning was analysis. As a result, Random Forest (RF) had a high prediction accuracy of 99.97%, and the test data RMSE was 305,923.92 and the training data RMSE was 251,927.41. The test data RMSE is higher than the training data RMSE, so there is a slight possibility of overfitting, but the prediction accuracy is very high, so there is a large possibility of overfitting is It was analyzed as not a problem. We preprocessed 32,183 bid result data from the Chungcheong region from January 1, 2019 to October 15, 2023, and used 17,925 bid result data as actual data to create predicted values trained using the expected price (actual value) and training data. When comparing the bid prices, it was confirmed that the values were almost close to the successful bid price. Although the actual and predicted values did not completely match, it was found that almost similar graphs appeared. Therefore, among the prices predicted in the experiment, there are many cases that fall between the successful bid price and the lower limit price, so it is analyzed that it is possible to predict the bid price with a high probability of winning when predicting the bid price using the Random Forest algorithm. The results of this study can be expected to increase the probability of winning a bid when predicting the bidding price using the Random Forest algorithm, and are expected to have a positive impact on alleviating the psychological stress of bidders and the company's business performance. Considering that 17,925 pieces of data were used as actual data when 32,183 pieces of data were preprocessed, it appears that a lot of unnecessary data was included in the data collection process. This can be seen as insufficient classification of unnecessary data during the data collection process, and it can affect more precise analysis, so it was necessary to collect more precise data during the data collection process. In addition, when analyzing Deep Learning with an algorithm, it was affected by computer performance, so the limitations of the research cost burden and analysis time were felt. The next study is expected to increase the accuracy of prediction by adding an algorithm using Deep Learnig, and conduct experiments to obtain the optimal hidden layer or node by changing the parameter values in more diverse ways, which will allow analysis to the winning bid price. The next study uses nationwide service bidding data to increase the accuracy of prediction by adding an algorithm using Deep Learnig, and experiments to obtain the optimal hidden layer or node by changing the parameter values in more diverse ways will bring the result closer to bid price.

References

1. M.S.Kang, "The effect of multiple reserve price generation range and section setting method on estimated price in bidding and successful bidding system", Mater's degree thesis at Korea University Graduate School of Management and Information, pp. 11-15, 43, 2014.
2. C. Kim, "Development of bidding-ratio prediction model for public information technoloty business projects using data mining method," Master's degree thesis at Yonsei University Graduate School of Department of Inustrial Information Management, pp. 15-20, 2017.
3. D. H. Hwang, Y. C. Bae. (2020). Bidding price prediction using deep learning in electronic bidding. Journal of the Electronics and Telecommunications Society of Korea, 15(1), 147-152.
4. Ministry of Strategy and Finance established regulations No. 653, 2023.06.16, partially revised, (contract regulations) standards for preparing expected prices.
5. Ministry of the Interior and Safety Regulation No. 252, June 29, 2023, partially revised, local government bidding and contract execution standards.
6. Ministry of Strategy and Finance, Contract Regulations, 2022.06.01. Partial revision, full text of contract regulations.
7. Ministry of the Interior and Safety Regulation No. 253, June 29, 2023, partially revised, criteria for determining successful bidders when bidding for local governments.
8. Tom M,Mitchell, Introduction to Machine Learning, 1997
9. Michael I.Jordan, Unsupervised Learning:A Theoretical and Empirical Analysis, 2002
10. wikipedia, K-means clustering, https://en.wikipedia.org/wiki/K-means_clustering
11. A. Goia, C. May, and G. Fusai, "Functional clustering and linear regression for peak load forecation", Elsevier, International Journal of Forecation", Elsevier, International Journal of Forecastion, Vol, 26, No. 4, pp. 700-711, Oct, 2010.
12. Cha Ki-wook, Hong Won-hwa. (2023). Development of a prediction model for the amount of decommissioned waste generated using a decision tree-based algorithm. Journal of the Architectural Institute of Korea, 39(3), 179-187.
13. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32
14. Jeong Sang-geun, Bae Jin-ho, Kang Hye-gyeom, and Kim Seong-yeol. (Date). Development of SVR-based short-term power demand prediction algorithm using public data. Korean Institute of Electrical Engineers conference proceedings, venue.
15. Taeyoung Kim and Jinho Kim, "Solar power generation prediction using support vector regression and Python library", Proceedings of the Korean Institute of Electrical Engineers

conference, pp. 540-541, 2018

16. Seon-ho Kim, Nak-hoon Choi, Jong-seok Oh. (2023). Dissolution prediction study of anti-tuberculosis drug capsules using lNN and ensemble. Journal of the Korean Society of Industrial-Academic Technology, 24(1), 531-537, 10.5762/KAIS.2023.24.1.531.
17. Z. Zhang, "Introduction to machine learning: k-nearest neigbors", Annals of translational medicine, vol. 4, no. 11, pp. 218, Jun. 2016. DOI:https://doi.org/10.21037/atm.2016.03.37
18. Su-han Kim, Ho-rim Wang, Won-ju Lee, Byeong-hyuk Ahn, Yu-jeong Kim, Joo-ho Lee, and Hyun-seong Shin. (2023). Molecular dynamics data-based hyperelastic constitutive equation modeling using MLP, GR, and RBF artificial neural networks. Journal of the Korean Society of Mechanical Engineers, Volume A, 47(1), 49-57, 10.3795/KSME-A.2023.47.1.049.
19. Adnaa, J. Daud, N. G. N., Ishak, M. T., Rizman, Z. I. and Rahman, M. I. A., 2018, "Tansig Activation Function (of MLP network) for Cardiac Abnormality Detection," AIP Conference Proceedings, Vol. 1930, No. 1, Article 02006.
20. Se-Joon Park,Yung-Cheol Byun, Electric Mobility Demand Prediction by Region using K-means Clustering", Journal of KIIT. Vol. 19, No. 11. Pp. 125-132, Nov. 30. 2021. http://dx.doi.org/10.14801/jkiit.2021.19.11.125
21. Maltare, N. N., Sharma, D., Patel, S. (2023). An Exploration and Prediction of Rainfall and Groundwater Level for the District of Banaskantha, Gujrat, India. International Journal of Environmental Sciences, 9 (1), 1-17 https://www.theaspd.com/resources/v9-1-1-Nilesh%20N.%20Maltare.pdf
22. Min, P.K., Mito, K. and Kim, T.H. (2024). The Evolving Landscape of Artificial Intelligence Applications in Animal Health. Indian Journal of Animal Research. https://doi.org/10.18805/IJAR.BF-1742
23. Kim, T. H. and AlZubi, A.A. (2024). AI Enhanced Precision Irrigation in Legume Farming: Optimizing Water Use Efficiency. Legume Research. https://doi.org/10.18805/LRF-791