WJERT, 2026, Vol. 12 Issue 9, 145-158. **Original Article** **ISSN 2454-695X**

World Journal of Engineering Research and Technology

www.wjert.org

Impact Factor: 8.067 **Coden USA: WJERA4**

DESIGN AND DEVELOPMENT OF AN END-TO-END PREDICTIVE MODELING SYSTEM TO TURN MESSY RAW DATA INTO SOLID PREDICTIONS

B. Lalitha Bhavani[1]*, N. Varshitha Sai[1], U. Kasivarun[1], M. Majji Babu[1], P. Kiran Sai[1], G. Krishnaveni[1]

[1]Department of Information Technology, Sir C. R. Reddy College of Engineering, Eluru-534007, Andhra Pradesh, India.

Article Received on 05/08/2026 **Article Revised on 25/08/2026** **Article Accepted on 03/09/2026**

***Corresponding Author**

B. Lalitha Bhavani

Department of Information Technology, Sir C. R. Reddy College of Engineering, Eluru-534007, Andhra Pradesh, India.

https://doi.org/10.5281/zenodo.22267252

How to cite this Article: B. Lalitha Bhavani[1*], N. Varshitha Sai[1], U. Kasivarun[1], M. Majji Babu[1], P. Kiran Sai[1], G. Krishnaveni[1] (2026). Design And Development Of An End-To-End Predictive Modeling System To Turn Messy Raw Data Into Solid Predictions. World Journal of Engineering Research and Technology, 12(9), 145–158.

ABSTRACT

This paper presents a Predictive Modeling System. In real world, the data is raw which means some of the data is missing, data is not structured etc. So the predictions we make are manual and not correct. Without the clean data we make wrong predictions that lead to wrong outcomes. Also it takes lot of time. So, to make predictions as accurate as possible and to reduce the time taking process we had created a Predictive Modeling System. To solve this issue, we had created a predictive modeling system which is used to do data preprocessing, feature engineering, model selection, validation. Data preprocessing means taking a file and preprocessing it using filling missing values, cleaning noisy data etc., Feature Engineering is used to create new features from the existing ones. Model selection is used to select models like we use random forest and gradient boosting. Validation is used to validate the model that the model is correctly defined or not or is there any issue.

KEYWORDS: Predictive Modeling · Machine Learning · Feature Engineering · Data Preprocessing · Random Forest · Linear Regression · Healthcare Analytics · Financial Prediction.

1 INTRODUCTION

In present world, predictive modeling is a featured technique used in many industries such as financial, healthcare, loan prediction, and manufacturing. This model is mainly used to learn from the existing data to make predictions. The model we build is more accurate that give most close predictions with the actual once. The importance of this project is working of the model that should work more efficiently as possible. It mainly focusses on development and outcome process in smoothly by regeneration of the predictive model.

In this project we used to standardize preprocessing for reduction of the manual preprocessing which means when we generate any dataset firstly it preprocesses. Next, it will select a nice algorithm that can be matched to our project. The validation process is done so that we know how well our model is performing. Finally, gives nice predictive analytics and used for decision making.

The scope is mainly used to define the technique that we have used like for our dataset we used supervised learning where it uses both classification and regression. We had also used some widely used programming languages such as Python with scikit-learn, TensorFlow etc.

1.1 Techniques for End-to-End Predictive Model

The below are the three approaches that we can use.

Data-Centric Approach: The Data-Centric approach is mainly used before model deployment to increase data quality, reliability, robustness and prediction accuracy.

Modeling-Centric Approach: The Model-Centric approach is mainly used after preprocessing stage to increase predictive performance through advanced model techniques and optimization. In this approach different ML models are trained, compared and tuned.

Pipeline Automation and MLOps Approach: This approach includes all stages of predictive modelling system with a structured and automated format. Here, it ensures scalability, reproducibility and faster decision making. This approach converts simple predictive model into an end-to-end production-ready model.

2 Literature Survey

Imran Fayyaz, G. G. Md. Nawaz Ali & Samantha Syeda Khairunnesa[1] (2025) used several multiple linear regression models for the cars’ price prediction and solved the overfitting problem by a broad Feature Engineering Pipeline (encoding, imputation, outliers removal, normalization, car age and pairwise interaction features). After feature engineering ensemble methods proved best in R^2 and MAE and RMSE.

Philomine Roseline[2] (2025) suggested EnStaR framework combining layered multi-regressors along with the Gradient boosting techniques resulted in optimized hyperparameters for reporting employees' performance scores. The ensemble stacking resulted better than single regressors with MAE, MSE and R^2 metrics.

Phuong Anh Nguyen, Bach Le, Luu Tuan Hoang, Xuan Le & Le Anh Ngoc[3] (2025) proposed a hybrid house price forecasting methodology, which incorporated ML-based text features with the contextual information from an LLM. In this approach, the inclusion of contextual LLM knowledge with structured real estate data enabled more accurate results as compared to standard ML models.

Anjan Kumar Reddy Ayyadapu[4] (2025) constructed a scalable machine learning framework for predictive analytics in big data context with focus on feature engineering and data pre processing in high-dimensional setting. Hybrid approaches improved the computational efficiency and prediction accuracy for big data context.

Clement O. Ogeh, G. C. Omede & F. O. Okorodudu[5] (2025) designed a web-enabled intelligent market prediction system with the implementation of regression modeling and enhancement of data handling methods. The proposed methodology enhances prediction accuracy and facilitates strategic business management.

Millicent Auma Omondi, Josu Nguinab, John Kamwele Mutinda, Amos Kipkorir Langat, Leonard Sanya, Ouraga Aime Cervert Ballou & Jeremy Nilandu Mabiala[6] (2025) implemented interpretable machine learning models to estimate premiums. Features comprised property risk and location. The methodology exhibited high performance in establishing the premiums and was interpretable.

MM Rahman, Babakalli Alkali, Amit Kumar Jain, Jess Gutirrez, C. Mcneil & Jake R Nelson[7] (2024) demonstrated the use of various machine learning models such as XG Boost, Decision Tree, Random Forest & Gradient Boosting for predictive maintenance of CCTV on electric trains. They used extracted features from sensor data to give better prediction system and safer transportation services.

Zalak A. Patel, Ayushi Porwal, Kajal Bhandare & Jongwook Woo[8] (2024) explored different regression algorithms in machine learning to estimate the net prices at USA colleges using family financial and school admittance characteristics. The statistical tests indicated that ensemble type strategies had the best results in term of predicting learning costs.

Arya Vinod, Anup K. Prasad, Sameeksha Mishra, Nirasindhu Desinayak, Sach Mukherjee, Anubhav Shukla, Nirasindhu Desinayak, B. C. Sarkar & Atul Kumar Varma[9] (2024) used a hybrid of machine learning models to actually predict a variety of FTIR to ash and coal content. The feature selection of wavelength feature was made to gain machine learning prediction, which was an excellent application in the field of material science.

Farzana Islam[10] (2024) innovated data-triggered machine and statistics learning models for CO_2 emission estimate & sustainable factory designing in direction to carbon neutrality. Feature learning & sustainability indicators enriched the predication practice for emission lowering & ecological planning.

Table 1: Summary of Research Studies on End-to-End Predictive Modeling System Using AI and Machine Learning Techniques.

Author(s)	Year	Focus Area	Techniques /Modules Used	Key Findings
Imran Fayyaz et al.	2025	Car price prediction	Linear Regression & Random Forest	Ensemble and neural networks could all achieve the highest R2 score and accuracy. 2
Philomine Roseline	2025	HR analytics & employee performance prediction	EnStar, ML Supervised & Regression algorithms	Using EnStaR, the performance forecast is better than any single algorithm.
Nguyen et al.	2025	Real estate house price prediction	Advanced ML & LLMs	Innovatively integrating LLMs and ML to predict houses prices.
Anjan Kumar Reddy Ayyadapu	2025	BDA hybrid ML model for large-scale prediction	Hybrid ML on big data	Hybrid ML model is able to execute big data predictive analytics.
Ogeh et al.	2025	Market forecasting	Linear Regression, Web–Based Deployment	Web-based regression and optimization make intelligent market forecasting decision.
Omondi et al.	2025	Insurance premium prediction	ML models	Interpretable ML to make transparent forecast for insurance premium.
Rahman et al.	2024	Transportation predictive maintenance	Random Forest Regression	CCTV-failure predicting could improve service reliability.

Patel et al.	2024	college net price prediction	Linear Regression & Random forest	Comparative evaluation of regression models for college net price prediction.
Vinod et al.	2024	Materials science phosphorus esti-mation	Multi-model ML & FTIR	Multi-model FTIR spectroscopy estimates phosphorus accurately in coal and ash.
Farzana Islam	2024	CO2 emission re-duction & carbon neutrality	Data-driven predic-tive models for emissions	Data driven CO2 prediction for industrial carbon neu-trality.

3 Problem Statement

In real life, the data for model input bought back are usually extremely chaotic, thus affects the model's performance and wasting our manual work dramatically. Moreover, all state-of-art approaches based on data mining methods cannot satisfy all kinds of high level operation of end to end predictive modeling system involving data cleaning, data preparing, feature extraction, model training and validation, and so on. However, there are still huge powerful possibility in the future to develop reproducible and scalable predictive modeling system for data to valuable knowledge conversion. The comparison also shows that ensemble learning with feature engineering is a critical step towards an end-to-end system for prediction.[2]

4 Proposed Methodology

4.1 Dataset Details and Context

In the methodology, there are two dataset implementations where one dataset is created that is Financial Dataset known as Loan Approval Scoring Dataset and the other dataset was taken from Kaggle that is Healthcare Dataset known as Medical Cost Personal Dataset. The second dataset is obtained from.[11] Fields include.

- age – patient's age in years
- sex – patient's gender
- bmi – body mass index
- children – number of children
- smoker – yes or no value
- region – represents one of eight directions
- charges (target column) – cost for their treatment

4.2 Financial Dataset

This dataset is generated and used for loan approval. Following are the steps involved.

1. Data Generation and Saving

Here, this dataset is generated which is a Loan approval Scoring dataset in csv file where it contains the fields as shown:

- age – years of the applicant
- income – applicant income
- credit score – past credit history of an applicant
- loan amount – requested by an applicant
- employment years – shows no of years employed
- approval score (target column) – calculated score showing the probability of the approval of the loan

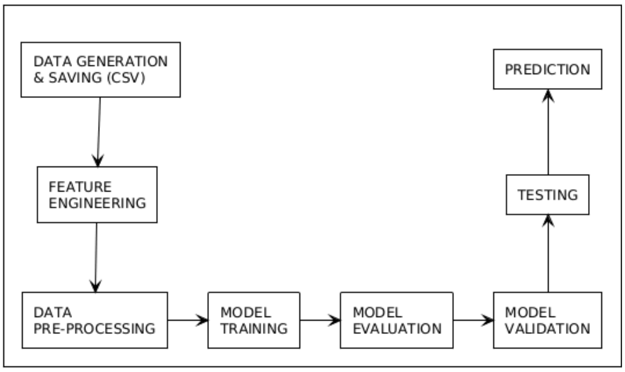

Fig. 1: Workflow diagram of the end-to-end predictive modeling system.

2. Feature Engineering

Key features are generated from the raw data by applying certain formulas. Feature engineering techniques are used extensively to do predictive modeling.[3] [14]

The features are.

- income per year employed: average income earned, per year, of employment.
- loan amount ratio: How much loan is received in comparison to income.
- credit age product: attribute that combines credit score and age; it expresses how credit-worthy one can be.

3. Data Pre-processing

Prior to training a model we need to pre-process it. Data pre-processing and feature scaling are important steps in predictive modeling pipelines as discussed in.[4] Here, we have preprocessed it using the technologies given below:

- **Pandas**: Used to manage dataset, drop columns and select features/target.
- **scikit-learn**: Used for feature scaling for normalizing data.
- **Joblib**: Used to save the trained model as scaler.pkl file.

4. Model Training

Trained our model with the following technologies.

- **scikit-learn**
- train test split() – splitting data into train and test
- LinearRegression() – Linear regression model
- RandomForestRegressor() – ensemble based regression model

Regression based approaches and ensemble based methods had been influenced extensively in predicting analytics.[1]

- **Joblib**: To save trained models as .pkl files (linear model.pkl, textttrandom forest model.pkl)

5. Model Evaluation

Model evaluation refers to the ability of the machine learning model to perform on unseen data (test).

- **scikit-learn**
- mean absolute error (MAE): The mean absolute error, measuring how close predictions are to the outcomes
- mean squared error (MSE): The mean squared error, measuring how close predictions are to the outcomes, effectively measuring the error magnitude and penalizing large errors more than small ones
- r2 score (R^2 Score): That tells you how well model fit to the data producing a value – for the better-fit model the closer it is to 1

Model performance assessment using MAE, RMSE and R^2 Score is a comparison in regression-based predictive systems.[12]

- **Numpy**: Computes RMSE from MSE

6. Model Validation

Model validation comparison can be made for linear regression and random forest in a bar graph with R^2 Score. Comparison of performance improves model selection and the confidence of the predictive system.[13]

7. Testing

Testing is done by using Postman with POST API method, a URL, Key as Content-Type, value as application/json and a body with input of all the columns. Here, for doing all these calculations there are two app.py to run the app and train.py to actually train the model to test it. After testing we get:

- confidence: an approximate value of confidence to approve the loan
- prediction: 0 or 1
- loan status: “APPROVED” or “REJECTED”
- response time ms: time taken to get the response

Web deployment and intelligent prediction system has been investigated in.[5] Machine learning model is used along the database driven web deployment framework through[15] in financial intelligent risk control system.

8. Prediction

The outcome of the prediction after testing is given below.

- 0 – refers loan rejection
- 1 – refers loan approval

4.3 Healthcare Dataset

This dataset contains medical cost information. Following are the steps involved.

1. Data Collection

Source: The data has been pre collected and uploaded on Kaggle; hence it does not require the realtime data from Hospital.

Background: Provided a healthcare dataset that was downloaded from Kaggle (Medical Cost Personal Dataset). Demographic and medical attributes have been included. Due to the domain knowledge random missing values were taken into account and is ready for predictive modeling using various preprocessing approaches.

2. Feature Engineering

Key features are generated from the raw data by applying certain formulas. Feature engineering techniques are used extensively to do predictive modeling.[13] [14]

The features are.

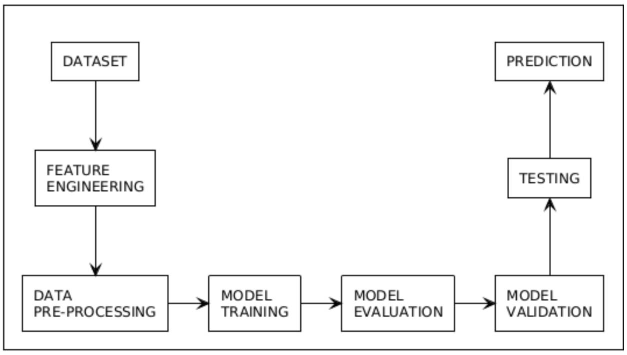

Fig. 2. Workflow diagram of the end-to-end predictive modeling system.

- bmi age product: The index of body mass index multiplied by age which naturally shows the influence of health overall risks
- children age ratio: Ratio indicates the dependents over age

Web-based deployment and intelligent prediction platform are discussed in papers.

3. Data Pre-processing

Our data need to be Pre-processed before training a model. Data pre-processing & Feature scaling are one of the most important stages of a predictive modeling pipeline as described in.[4] We did pre-processing on the dataset using the following technologies.

- **Pandas**: Used to manage dataset, drop columns and select features/target
- **scikit-learn**: Used for Feature Scaling for normalizing data
- **Joblib**: Used to Save the trained model as scaler.pkl file

4. Model Training

Trained our model with following technologies.

- **scikit-learn**
  - train test split() – splitting data into train and test
  - LinearRegression() – Linear regression model
  - RandomForestRegressor() – ensemble based regression model

Regression based approaches and ensemble based methods had been influenced extensively in predicting analytics.[1]

- **Joblib**: To save trained models as .pkl files (insurance linear.pkl, insurance rf.pkl)

5. Model Evaluation

Model evaluation refers to the ability of the machine learning model to perform to unseen data (test).

- **scikit-learn**
- mean absolute error (MAE): The mean absolute error, measuring how close predictions are to the outcomes
- mean squared error (MSE): The mean squared error, measuring how close predictions are to the outcomes, effectively measuring the error magnitude and penalizing large errors more than small ones
- r2 score (R^2 Score): That tells you how well model fit to the data producing a value – for the better-fit model the closer it is to 1

Model performance assessment using MAE, RMSE and R^2 Score is a comparison in regression-based predictive systems.[12]

- **Numpy**: Computes RMSE from MSE

6. Model Validation

Model validation comparison can be made for linear regression and random forest in a bar graph with R^2 Score. Comparison of performance improves model selection and the confidence of the predictive system.[13]

7. Testing

Testing is done by using Postman with POST API method, a URL, Key as Content-Type, value as application/json and a body with input of all the columns. Here, for doing all these calculations there are two app healthcare.py to run the app and train healthcare.py to actually train the model to test it. After testing we get.

- confidence: an approximate value of confidence to approve the loan
- prediction: 0 or 1
- Risk level: “HIGH” or “LOW”
- response time ms: time taken to get the response

Web-based deployment and intelligent prediction platform are discussed in.[5] Financial intelligent risk control system are now using more and more machine learning models embedded in a database managing deployment platform.[15]

8. Prediction

The prediction test result as below.

- 0 – indicates Low medical cost
- 1 – indicates High medical cost

4.4 Multiple Linear Regression

Multiple Linear Regression is a supervised statistically approach that aims to predict a real value continuous target attribute according to a set of other real valued independent attributes. It considers the dependencies among these attributes as being linear which lead to the following prediction function. The coefficients are estimated by minimizing the mean squared error between predicted and actual values. It seems to be very easy to use and comprehensible thus popular in cost prediction and financial forecasting.[8]

The prediction function is.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

The cost function minimized during training is.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

4.5 Random Forest

To adapt it to make a forecast Forest of decision trees is a method.[7] It explore strata of the data using different parameters to make each tree then it averages the results of all the trees to get the result. this makes the system more robust and prevents him from becoming too flexible and makes it useful for data that has a lot of structure.

The ensemble prediction is.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

5 RESULTS AND DISCUSSION

5.1 Performance Evaluation

The end-to-end predictive modeling system described in this paper was evaluated on.

- Synthetic financial dataset
- Live healthcare insurance dataset

The scores above were compared to benchmark studies[5] and.[6]

5.2 Financial Dataset Results

The results that the system under test produced using the synthetic financial dataset were.

- Linear Regression: $R^2 = 0.67$
- Random Forest: $R^2 = 0.62$ The results for study[5] were.
- Linear Regression: $R^2 = 0.89$
- Random Forest: $R^2 = 0.92$

The proposed system underperforms study[5] because the data used in this work is synthetic and the topic-specific optimization techniques used in[5] could not be applied.

5.3 Healthcare Dataset Results

The achievable R^2 for a real-world healthcare insurance dataset using the system defined in this paper is:

- Linear Regression: $R^2 = 0.7826$
- Random Forest: $R^2 = 0.87$ While that from study[6] is.
- Linear Regression: R^2 = NR (Not Reported)
- Random Forest: $R^2 = 0.8137$

As we can observe above, the proposed system achieves a higher R^2 score on the healthcare dataset when compared with the existing system.

Table 2: Comparative R^2 Performance of Existing and Proposed Regression Models on Financial and Healthcare Datasets.

Dataset	Model	Literature R^2	Proposed R^2
Financial	Linear Regression	0.89	0.67
Financial	Random Forest	0.92	0.62
Healthcare	Linear Regression	NR(Not Reported)	0.7826
Healthcare	Random Forest	0.8137	0.87
Financial	**Average (LR + RF)**	**0.905**	**0.645**
Healthcare	**Average (LR + RF)**	**0.8137***	**0.8263**

The table clearly shows there is lower quality in financial because it is synthetically generated but for the high quality of results that can be achieved on real-world data with the carefully structured approach of feature engineering, validation etc., of the end to end pipeline.

In the literature, the R^2 value for the Linear Regression with healthcare data set is missing (absent), noted by NR while the proposed system is able to give a full evaluation on both data sets. They show that the predictive ability of their approach are comparable or better.

6 CONCLUSION AND FUTURE SCOPE

The proposed methodology exhibits stable and promising predictive accuracy in financial and health datasets, with a better R^2 score (Random Forest) recorded in the health task, suggesting ensemble based methodology to be a promising technique. Further research can be led by advanced boosting scheme, automatic feature construction, best parameter optimization, utilizing large scale real practice data set etc.

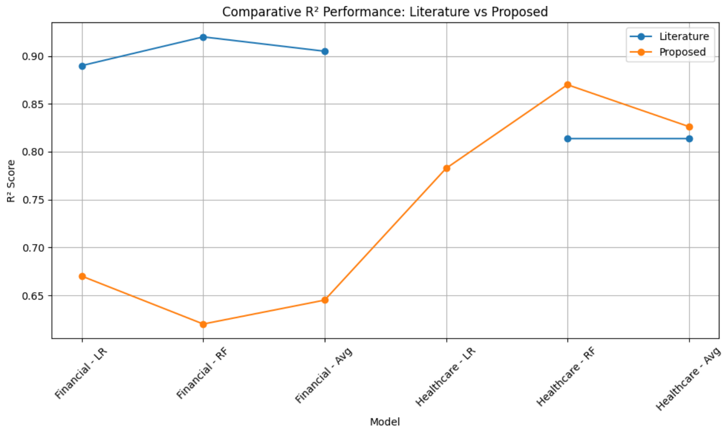

Fig. 3: Comparative line graph of literature and proposed R2 performance across financial and healthcare datasets.

REFERENCES

1. Fayyaz, I., Ali, G.G.M.N., Khairunnesa, S.S.: Advanced feature engineering and machine learning techniques for high accurate price prediction of heterogeneous pre-own cars. Vehicles, 2025; 7(3): 94. https://doi.org/10.3390/vehicles7030094
2. Roseline, P.: An ensemble framework of stacked multi regressors and gradient boosting algorithms with tuned hyperparameters EnStaR to predict employee performance score using HR analytics. Int. J. Appl. Math, 2025; 38(3s): 283–303. https://doi.org/10.12732/ijam.v38i3s.146
3. Nguyen, P.A., Le, B., Hoang, L.T., Le, X., Ngoc, L.A.: New method of house price forecasting through advanced machine learning algorithms and large linguistic models, 2025; 179–194. https://doi.org/10.4018/979-8-3373-2647-4.ch011
4. Ayyadapu, A.K.R.: New hybrid machine learning model to traditional predictive analytics in big data frameworks. ATICS, 2025; 2(1): 27–38. https://doi.org/10.64091/atics.2025.000103
5. Ogeh, C.O., Omede, G.C., Okorodudu, F.O.: Development of a web-based intelligent market forecasting system using regression modeling and data optimization. Int. J. Res. Sci. Innov, 2025; 12(6): 1015–1025. https://doi.org/10.51244/ijrsi.2025.12060082
6. Omondi, M.A., Nguinab´e, J., Mutinda, J.K., Langat, A.K., Sanya, L., Ballou, O.A.C.,

Mabiala, J.N.: Interpretable property insurance premium prediction using machine learning models. J. Artif. Intell. Appl, 2025; 2(3): 349–370. https://doi.org/10.54364/jaiai.2024.1121

7. Rahman, M., Alkali, B., Jain, A.K., Guti´errez, J., Mcneil, C., Nelson, J.R.: Predictive maintenance optimisation for CCTV systems in electric multiple unit trains using machine learning techniques. Civ. Comp. Proc, 2024; 7: 1–11. https://doi.org/10.4203/ccc.7.8.11.
8. Patel, Z.A., Porwal, A., Bhandare, K., Woo, J.: US college net price prediction comparing ML regression models, 2024. https://doi.org/10.48550/arxiv.2406.08071.
9. Vinod, A., Prasad, A.K., Mishra, S., Desinayak, N., Mukherjee, S., Shukla, A., Desinayak, N., Sarkar, B.C., Varma, A.K.: A novel multi-model estimation of phosphorus in coal and its ash using FTIR spectroscopy. Dent. Sci. Rep, 2024; 14(1). https://doi.org/10.1038/s41598-024-63672-x.
10. Islam, F.: Data-driven approaches for achieving carbon neutrality: Predictive models for reducing CO2 emissions and enhancing industrial sustainability, 2024. https://doi.org/10.33915/etd.12428.
11. Mirichoi0218: Medical cost personal dataset. Kaggle, https://www.kaggle.com/datasets/mirichoi0218/insurance.
12. Cherguy, O., Chmielowski, R., Hachem, E., Lahouij, I.: Deep learning prediction of dry friction in DLC coatings using literature-derived data. Tribol. Lett, 2025; 73(4). https://doi.org/10.1007/s11249-025-02056-2.
13. Yenikaya, M.A., Oktaysoy, O.: Applying machine learning approaches for improvement in energy efficiency: Comparative study of energy estimation models. EKEV Akademi Dergisi, 2025; 103: 196–210. https://doi.org/10.17753/sosekev.1636999
14. Iyiola, Z., Guzha, T., Nwosu, C.J., Devegowda, D., Ezeh, M.C., Obeng, S.D.A.: Study of effect of feature engineering methods in prediction of log properties. J. Nat. Gas Sci. Eng. (2025). https://doi.org/10.2118/228658-ms
15. Guo, H., Liu, Y., Hu, L., Zhang, X.: Machine learning enterprise financial intelligent risk control system based on new database. Int. J. Global Econ. Manag, 2024; 5(2): 283–289. https://doi.org/10.62051/ijgem.v5n2.30