



Forecasting Land Prices in Colombo District, Sri Lanka using Machine Learning and Spatial Network Analysis

W.G.K.D. Weerasinghe ^{a*}, Y.P.R.D. Yapa ^b

^a Department of Statistics and Computer Science, University of Peradeniya, Kandy 20000, Sri Lanka

^b Department of Statistics and Computer Science, University of Peradeniya, Kandy 20000, Sri Lanka

^a Email: dheelakafi@gmail.com

^b Email: roshany@gmail.com

Abstract

Land prices are essential indicators of economic and social development, but forecasting them presents challenges due to the multitude of factors that influence land value. This study explores the use of Machine Learning and geospatial analysis to forecast land prices in 50 key cities within the Colombo District of Sri Lanka. The dataset used comprises customer-submitted land price information provided by a real estate company 'Lanka Property Web' collected between 2018 and 2023. Geospatial data, including shapefiles, were obtained from the Survey Department of Sri Lanka and used to generate new variables, such as distances from land plots to significant infrastructure and locations, through spatial network analysis. A range of machine learning algorithms were evaluated for their performance in predicting land prices, including Linear Regression, Gradient Boosting, XGBoosting, Random Forest, and Artificial Neural Networks (ANN). Model performance was assessed based on predictive accuracy, with XGBoosting emerging as the most accurate, achieving an accuracy of 62.23%. In comparison, Linear Regression achieved 8.29%, Gradient Boosting 60.01%, Random Forest 58.54%, and ANN 48.01%. To further extend the forecast, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model was applied to predict land prices up to the second quarter of 2027. The final deliverable of this research is a user-friendly dashboard designed for individual land buyers and sellers. This dashboard provides dynamic, data-driven insights into land market trends, empowering users with the information needed to make informed decisions in a rapidly changing property market. This study demonstrates the potential of machine learning and geospatial

analysis in improving land price forecasting, offering valuable tools for stakeholders in the real estate sector.

Keywords: Colombo District; Land Price Forecasting; Machine Learning; Real Estate; Spatial Analysis

* Corresponding author. Email address: dheelakafi@gmail.com

1. Introduction

Sri Lanka has experienced rapid urbanization over the past few decades, with Colombo District at the center of economic growth and development. Land prices in this district reflect underlying socio-economic conditions and are vital indicators for investors, urban planners, and policymakers. Yet, forecasting land prices remains complex due to fluctuating market dynamics influenced by geographic, temporal, and socio-economic variables. Combining spatial network analysis with machine learning offers a promising avenue for robust land price forecasting, addressing challenges such as non-linearity, spatial heterogeneity, and temporal variation.

Traditional forecasting models often fail to incorporate spatial dependencies and complex interactions among variables, leading to suboptimal predictions. In Colombo District, the lack of a comprehensive forecasting framework impairs strategic urban planning and investment decision-making, potentially resulting in market inefficiencies, social inequities, and economic instability. Accurate land price forecasting can help identify emerging trends, detect potential bubbles, and promote equitable resource distribution. This research leverages cutting-edge Machine Learning techniques enriched with spatial variables to bridge this critical knowledge gap.

2. Materials And Methods

2.1. Data Collection

The primary dataset was sourced from Lanka Property Web (LPW), the leading online real estate platform in Sri Lanka. The dataset includes customer-submitted land sale advertisements within Colombo District, covering the period from January 2018 to September 2023. The raw dataset contained 58,352 entries with approximately 185 attributes per listing, capturing comprehensive property and listing details such as location (latitude and longitude), price, land size, textual descriptions, posting dates, and seller information. For analysis, a subset of around 50 main cities within Colombo District was selected based on data availability to ensure statistical reliability and representativeness.

2.2. Data Preprocessing

The original dataset contained over 58,000 entries, but due to high sparsity, over 120 columns were removed because they either contained more than 95% missing values or were irrelevant to land price prediction. In the next step, rows with missing values for essential features, such as the land price or geographical coordinates, were removed. Additionally, outliers in the land price data were detected using the Interquartile Range (IQR) method. These outliers were removed from the dataset to avoid distorting the model's performance due to extreme price variations. A new feature, price per perch, was computed for each land listing to standardize the price relative to the size of the land. This feature, the price divided by the land size (in perches), allowed for fair comparisons across listings of different sizes.

To standardize the city names across the dataset, a mapping dictionary was used. This ensured that each city name appeared consistently across all listings. Following these preprocessing steps, the dataset was reduced to around 27,000 entries that were suitable for further analysis.

2.3. Spatial Network Analysis and Feature Engineering

To capture the geographic context influencing land prices, spatial variables representing distances from each land plot to key amenities and infrastructure were engineered. Geospatial polygon and line shapefiles for roads, municipal council boundaries, hospitals, schools, bus stops, railway stations, and police stations were obtained from the Survey Department of Sri Lanka. Using ArcGIS 8 software and the Point & Polyline Tools v1.2 extension, least-cost path analysis was performed to calculate network distances between land parcels (latitude, longitude points) and nearest urban facilities [1]. Buffer zones were generated around amenities to determine proximity-based features and augment the dataset with categorical spatial indicators. The final spatial variables included: Distance to nearest hospital, Distance to nearest national school, Distance to nearest bus stop, Distance to nearest railway station, Distance to nearest main road [2].

2.4. Machine Learning Model Development

2.4.1. Cost Function

The cost function, that used in this project, is Mean Squared Error (MSE). The mean squared error serves as a quantitative measure assessing the proximity of the predicted response value for a given observation to the corresponding true response value [3]. This assessment is encapsulated by the formula:

$$(1)$$

Five Machine Learning algorithms were implemented to predict price per perch, trained and evaluated on the engineered feature set (excluding 'posted_date' for cross-sectional models).

2.4.2. Linear Regression

A baseline multivariate linear regression model was fit:

$$(2)$$

2.4.3. Random Forest Regressor

Random Forests build an ensemble of decision trees trained on bootstrapped subsets and select random feature subsets at each split to reduce variance. Tuned hyperparameters are as follows; Number of trees: 100, Maximum depth: None, Out-of-bag (OOB) error estimation for model validation.

2.4.4. Gradient Boosting Regressor

GBR sequentially fits decision trees on the residuals of previous trees, optimizing the loss function through gradient descent. [4]. Tuned hyperparameters are as follows; Number of estimators: 50, Learning rate: 0.2, Loss function: Mean Squared Error.

2.4.5. XGBoost Regressor

XGBoost is an efficient implementation of gradient boosting with regularization. Tuned hyperparameters are as follows; Number of estimators: 50, Learning rate: 0.1, Regularization parameters tuned via GridSearchCV.

2.4.6. Artificial Neural Network

The application of Neural Networks (NN) in modeling land price trends is described, with economic and social indicators being utilized [5]. A feedforward neural network was trained with; Input layer matching number of features, Two hidden layers with 64 and 32 neurons respectively, ReLU activation function, Output layer with linear activation, Optimizer: Adam, Maximum iterations: 10,000, Regularization parameter alpha: 0.001.

2.5. Seasonal Auto Regressive Integrated Moving Average (SARIMA)

To model temporal price trends, SARIMA models were fitted for each city's quarterly average price per perch. Selection of SARIMA utilized stationarity tests (Augmented Dickey-Fuller), autocorrelation and partial autocorrelation function plots, and Akaike Information Criterion (AIC) minimization [6]. The selected model captured both trend and seasonal effects, including anomalies such as COVID-19 induced shocks.

3. Results

3.1. Exploratory Data Analysis

Over the years from 2018 to 2023, the number of advertisements posted for each year was calculated, with a focus on grouping them by city to get an idea about available data. The few cities were removed those with fewer than 20 entries. Homagama emerged as the most active market with 4,664 listings, followed by Piliyandala with 3,373 entries. Other cities such as Padukka, Battaramulla, and Kaduwela also contributed significant data volumes. This spatial distribution reflects both population density and real estate market activity in Colombo District.

Table 1: Price metrics of each city

Main city	Min price	Max price	Average price	Std price	Median price	25% percentile	75% percentile
Angoda	200000	1160000	697277.7	223988.7	698484.5	538928.5	867500
Athurugiriya	250000	1300000	836564.9	248781.4	850000	650000	1025000

Main city	Min price	Max price	Average price	Std price	Median price	25% percentile	75% percentile
Avissawella	73529	240000	173325.5	36193.13	173611	138888	200000
Avissawella	85000	210000	156497.1	40058.96	149444	138888	192500
Battaramulla	800000	4000000	2459332	618228.7	2500000	2000000	2950000
Boralesgamuwa	750000	2800000	1692748	377578.7	1700000	1400000	1950000
Colombo 1	250000	685000	496213.6	139888.9	550000	380000	600000
Colombo 10	340000	704166	617810.8	49454.02	637804	600000	650000
Colombo 12	7500000	8500000	7628947	277546.7	7500000	7500000	7500000
Colombo 13	2600000	7083333	4773906	1150138	5000000	3800000	5966667
Colombo 14	3000000	9000000	6157751	1561705	6500000	4925000	7225000
Colombo 15	1650000	4716981	3457112	627951.4	3500000	3000000	4000000
Colombo 2	12000000	25000000	17190647	2612710	16500000	15020450	18750000
Colombo 3	9000000	28000000	18733459	3291340	19500000	16500000	20454545
Colombo 4	5434782	30000000	14705200	3104619	15000000	13000000	16000000
Colombo 5	3000000	13500000	6914172	1782343	6500000	5537500	8143750
Colombo 6	3200000	14285714	8309393	1882244	8450000	7000000	9619354
Colombo 7	9500000	25000000	16162624	2517027	16000000	14090909	18000000
Colombo 8	3300000	12000000	7634377	1746715	7900000	6500000	9000000
Colombo 9	2100000	6065217	4496390	972750.5	4619565	3970588	5163043
Dehiwala	1300000	5700000	3442399	680364.9	3500000	3000000	3950000
Ethul Kotte	900000	4500000	2989868	710503.4	3000000	2400000	3500000
Hanwella	47692	400000	301316.7	61350.52	325000	285000	335000
Hokandara	450000	1850000	1067634	318982.2	1000000	850000	1352174
Homagama	120689	883534	557194.7	106031.5	570000	510000	639508
Kaduwela	165000	565000	404782.5	67049.33	395000	365000	475000
Kahathuduwa	180000	800000	519515.6	108785.7	550000	430000	600000
Kesbewa	265000	1200000	812591	184580.5	900000	700000	900000
Kolonnawa	350000	1675000	943966.2	281838.5	950000	712500	1082576
Kosgama	75000	275000	182996.8	52289.48	200000	145000	225000
Kosgama	550000	700000	616666.7	76376.26	600000	575000	650000
Kottawa	375000	1875000	1085662	304793.6	1062500	825000	1350000
Maharagama	520000	3000000	1788811	406689.2	1925000	1500000	2000000
Malabe	460000	2500000	1321009	346393.8	1350000	1075000	1550000
Malabe	722222	833333	797222	52410.96	816666.5	780555.5	833333
Moratuwa	520000	2000000	1223163	293344	1200000	1000000	1468750
Mount Lavinia	1300000	4642857	2881325	558767	2900000	2500000	3300000
Nawala	2000000	6500000	4146727	788386.9	4000000	3500000	4750000
Nugegoda	1250000	5272727	2840019	726873.9	2800000	2400000	3500000
Padukka	80000	450000	268233.4	40671.67	260000	250000	290000
Pannipitiya	420000	2700000	1830013	313326.2	1961290	1850000	2000000

Main city	Min price	Max price	Average price	Std price	Median price	25% percentile	75% percentile
Piliyandala	240000	1450000	812220.3	232480.8	875000	650000	975000
Pita Kotte	1150000	4305555	2681087	636406.2	2600000	2200000	3200000
Polgasowita	180000	950000	419558.6	81634.74	430000	380000	430000
Rajagiriya	843000	7000000	3581343	1145920	3543293	2800000	4500000
Rathmalana	1000000	3500000	1960801	466134.9	1900000	1626016	2300000
Thalagala	500000	661538	602207.9	63389.46	650000	550000	650000
Thalahena	800000	2100000	1427653	297582	1433333	1200000	1687419
Thalangama	1700000	2700000	2163462	278503.3	2050000	2000000	2300000
Thalangama	1100000	1900000	1546429	289550.1	1575000	1375000	1750000
Thalawathugoda	500000	3000000	1823085	415209.1	1825000	1500000	2000000
Watareka	275000	550000	417032.1	35158.91	407692	400000	426576
Wattala	1500000	1500000	1500000	228365.2	1500000	1500000	1500000
Wellampitiya	400000	1100000	769000.9	173018.1	750000	625000	900000

Price metrics for each city were calculated by highlighting the significant variability in land values across the district. The minimum and maximum prices vary significantly across different areas. The average prices give an indication of the central tendency of the data. Colombo 3 has the highest average price (Rs. 18,733,459), while Avissawella has a lower average price (Rs. 173,325.5). Colombo 12 has a high standard deviation (Rs. 277,546.7), indicating a wider range of property prices, whereas Avissawella has a much lower standard deviation (Rs. 36,193.13). Colombo 2, Colombo 3, Colombo 7, and Colombo 8 stand out with high maximum prices, reflecting premium real estate in these areas. There were a downward trend in land prices in 2020 and 2021 likely due to the COVID-19 pandemic.

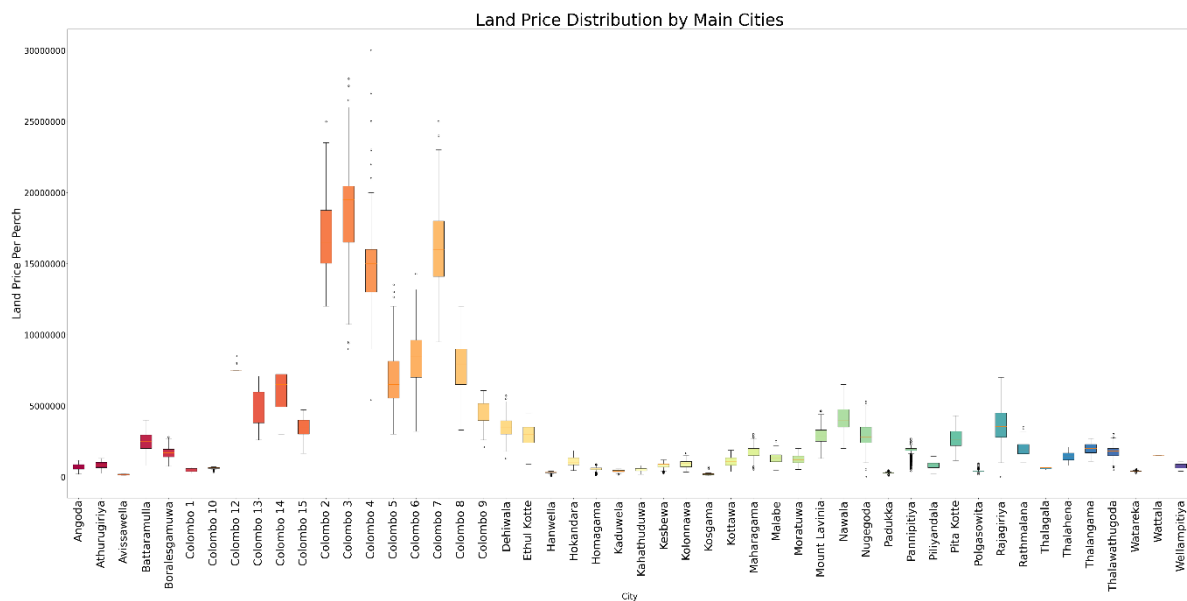


Figure 1: Land price distribution by city from 2018 to 2023

This boxplot visualization shows the distribution of land prices per perch across various main cities in Sri Lanka. Each box represents the interquartile range (IQR) of land prices for a specific city, with the line inside each box indicating the median price, and the whiskers and dots representing the spread and outliers, respectively. From the chart, it's clear that land prices are significantly higher and more variable in the Colombo metropolitan area, particularly in Colombo 3, 4, 5, 6, and 7, indicating high demand and possibly greater commercial or residential value. Colombo 1, 2, 8, and 10 also show relatively high prices, though with slightly lower variation.

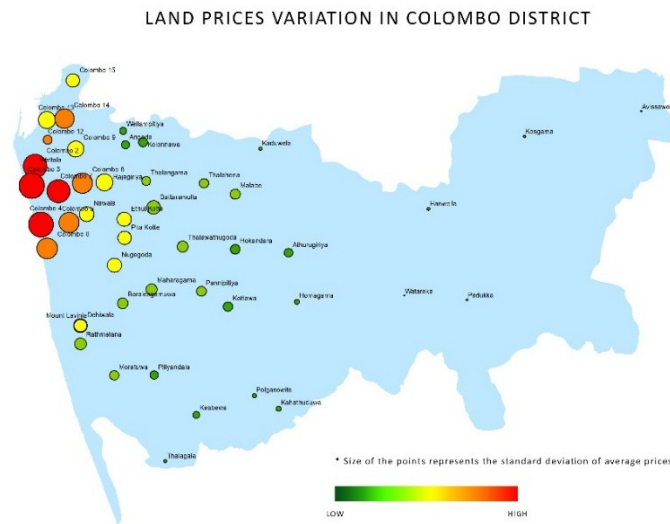


Figure 2: Deviation of prices in main cities

In this figure it shows the average prices in colors and standard deviation of price in size of the circles. It can be clearly seen that urban areas near Colombo city have higher prices and standard deviation compared to other places.

By further analysis, it was seen that the price of lands mainly varies depending on the distance from Colombo 1 to lands. Cities like Colombo 1-15 areas were highly priced. After that some areas like Nawala, Nugegoda, Rajagiriya, Ethul Kotte and Mount Lavinia had a middle price range. This can be shown in the map of Colombo district like below.

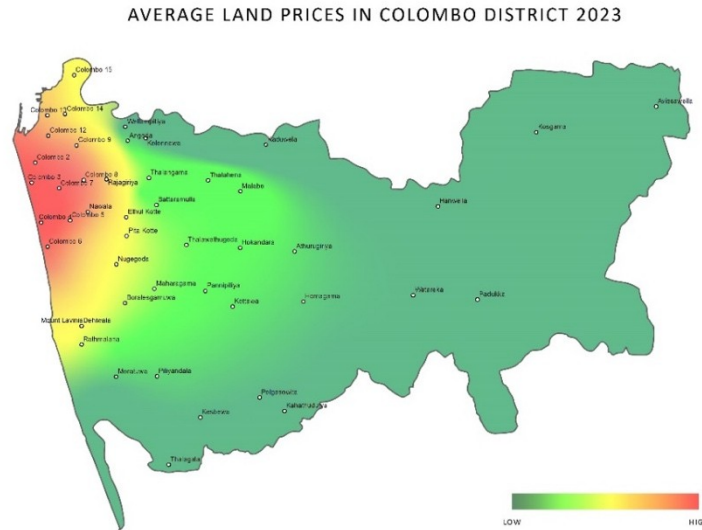


Figure 3: Land price variation with the distance from center in Colombo District

The spatial heatmap illustrates land price variation across Colombo District, displaying a clear gradient with decreasing prices as the distance from Colombo city center increases.

3.2. Temporal Price Trends

Time series analysis was conducted using monthly average land prices by city. Time series plots these monthly averages alongside smoothed trend lines for multiple cities, revealing a steady increase in land prices over the six-year period. A pronounced dip during 2020 and early 2021 corresponds to the onset of COVID-19, followed by recovery and accelerated price growth. Peripheral cities such as Homagama and Piliyandala show more moderate price growth relative to urban centers like Colombo 3 and Colombo 8.

3.3. Machine Learning Model Performance

To build machine learning models to predict the land prices, distance to key amenities such as distance to nearest hospital, distance to nearest school, distance to nearest railway station, distance to nearest bus stop and distance to nearest main road were selected as the independent variables using a correlation analysis. By taking these variables, RandomForestRegressor, GradientBoostingRegressor, LinearRegression, XGBRegressor and Artificial Neural network Models were fitted to the dataset.

Figure 4: Summary of Price prediction models

Model Name	Training Accuracy	Testing Accuracy
Linear Regression	15.62%	8.29%
Artificial Neural Network	76.23%	48.01%

Random Forest	90.26%	58.54%
Gradient boost	88.48%	60.01%
XG boost	91.38%	62.23%

Linear Regression performed poorly, achieving only 8.29% testing accuracy, reflecting the limitations of linear assumptions in this complex domain. The Artificial Neural Network achieved moderate success with 48.01% testing accuracy, indicating the benefits of capturing nonlinearities but also the challenges in training. Random Forest and Gradient Boosting models achieved higher accuracies, 58.54% and 60.01% respectively, demonstrating the effectiveness of ensemble methods. XGBoost outperformed all models with a testing accuracy of 62.23%, benefiting from advanced regularization and efficient training.

3.4. Time Series Forecasting

Time series forecasting model was build using a SARIMA model on a dataset. The 'posted_date' column is converted to a datetime format, and additional features such as 'year' and 'quarter' are extracted. The training data is selected based on a cutoff date ('2023-09-01'), wherein records with 'posted_date' earlier than this date are considered as the training set. The user is then prompted to input a main city as a keyboard input, which filters the dataset and creates a subset specific to the chosen city. The SARIMA model is then run and fitted to the training data. Model parameters, including the order (p, d, q) and seasonal order (P, D, Q, S), are specified for this purpose. The number of lag observations included in the model is 1 suggests an autoregressive order of 1. The number of times the raw observations are differenced to achieve stationarity. Value = 1 implies first-order differencing. Size of the moving average window order is specified as 1. Number of time steps in a seasonal period indicates a seasonal order of 4, suggesting that the seasonality occurs every 4 quarters. The model is employed to forecast land prices per perch for the period spanning '2024-01-01' to '2027-03-31'. Back testing on recent data indicated that the model achieved a Mean Absolute Percentage Error (MAPE) below 10%, confirming good forecasting accuracy.

4. Discussion

In the examination of land prices within the Colombo district of Sri Lanka, this study aimed to provide a comprehensive understanding of the economic and social landscape through the application of machine learning and geospatial analysis. Customer-submitted data from 'Lanka Property Web' spanning the years 2018 to 2023 was utilized, with a special pre-processing approach that involved outlier removal. The integration of spatial network analysis, shape files from the Survey Department of Sri Lanka, and machine learning techniques resulted in the creation of a user-friendly dashboard for individual land sellers and buyers.

Descriptive Statistics of bare land prices in different main cities within the Colombo district revealed significant variations in minimum and maximum prices. Colombo 3 emerged with the highest average price, while Avissawella exhibited a lower average. Colombo 12 demonstrated a high standard deviation, indicating a wider range of property prices, whereas Avissawella had a lower standard deviation, suggesting that urban areas near to Colombo city limits have higher average land prices and standard deviation of price.

Time series analysis spanning the years 2018 to 2023 revealed a general increase in land prices across most cities, with a noticeable decline in 2020-2021, potentially attributed to the COVID-19 pandemic. Spatial network analysis provided valuable insights, including distances to key amenities and infrastructure. This analysis contributed to the generation of variables essential for constructing machine learning models. In the model-building phase, five machine learning models were applied to predict land prices: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, XG Boosting Regressor, and Artificial Neural Network. The models were evaluated based on their training and testing accuracies. Among them, XG Boosting Regressor was selected as the best model for this data because of its training Accuracy of 91.38% and testing Accuracy of 62.23%.

A time series forecasting model using SARIMA was developed to predict land prices for the period 2024-2027. In conclusion, the user-friendly dashboard created in this analysis of the combination of machine learning, geospatial analysis, and time series forecasting has improved the understanding of land prices in the Colombo district, aiming investors and landowners in making informed decisions in a rapidly changing property landscape. The ultimate outcome, a user-friendly dashboard, stands as a valuable resource for individuals involved in land transactions, providing them with data-driven insights for well-informed decision-making.

As a continuation of this study, this model can be developed for the all the district in Sri Lanka. Since the original dataset contained only the data from Colombo district and there are very few data entries in other districts, this study had limited only to Colombo. But this can be developed to other districts with their significant variables. For example, altitude is not a significant factor for Colombo, but it will be significant for Kandy district. Apart from that, this study can drill down into another smaller level of location. City is the smallest unit in this study. As a future work, this can be developed up to street level.

5. References

- [1] M. Tresidder, "Using GIS to Measure Connectivity: An Exploration of Issues.," *Field Area Paper*, pp. 14-16, 2005.
- [2] Özer and Özlem, "Accessibility of Spatial Networks: Using ArcGIS network," 2017.
- [3] Hodson, O. Timothy, Over, M. Thomas, Foks and S. Sydney, "Mean Squared Error, Deconstructed," *Journal of Advances in Modeling Earth Systems*,, 2021.
- [4] D. Otchere, T. Ganat, J. Tackie-Otoo and M. Taki, "Application of gradient boosting regression model for the evaluation of feature," *Journal of Petroleum Science and Engineering*, p. 208, 2022.
- [5] P. Velumani, N. Nampoothiri and P. Kavithra, "Predicting the Land Value using Regression Techniques and Artificial Neural Network," *International Journal of Engineering and Advanced Technology*, p. 9, 2019.
- [6] Fattah, J. Ezzine, L. Aman, Z. Moussami, H. Lachhab and Abdeslam, "Forecasting of demand using ARIMA model," *International Journal of Engineering Business Management*, 2018.