Leveraging Data Mining, Machine Learning, and Web Scraping for Forecasting Rental Housing Prices in Tunisia

ALA BALTI[1,*], MOHAMED NAJEH LAKHOUA[2], MOUNIR SAYADI[1]
[1]University of Tunis, Research Laboratory SIME,
National High School of Engineering of Tunis, ENSIT,
TUNISIA

[2]University of Carthage, Research Laboratory Smart Electricity & ICT, SEICT,
National Engineering School of Carthage, ENICarthage,
TUNISIA

**Corresponding Author*

Abstract: - The Tunisian real estate market has experienced a notable 13.5% surge in prices since 2018, marking a substantial departure from the preceding five years, during which there was a 9% growth, as indicated by data from the National Institute of Statistics (INS). According to the 2020 Rental Barometer, the average monthly rent for unfurnished apartments stands at 1,360 Tunisian dinars. Our initiative, titled "Predicting Real Rental Prices," employs advanced machine learning techniques to provide accurate predictions for rental prices. Users of this platform can plan moves, organize properties into categories, and customize rental price insights based on their preferences. This project is based on machine learning and uses deep learning algorithms to predict rental prices, thereby meeting the needs of both lessors and tenants. The model ensures a thorough and accurate forecasting approach by accounting for a number of issues, such as the effect of furniture and building conditions on rental prices.

Key-Words: - Prediction, Deep Learning, Machine Learning, Data Analyse, Correlation, Linear Regression, Random-Forest.

Received: November 13, 2023. Revised: April 19, 2024. Accepted: June 12, 2024. Published: July 22, 2024.

1 Introduction

An increasingly significant component of the real estate market is the estimation of rental home prices, [1]. The increasing need for precise pricing data makes the incorporation of state-of-the-art technologies and techniques imperative. In order to forecast rental property prices, this paper presents a comprehensive method that combines web scraping, machine learning, prediction, linear regression, and data mining. Our goal is to provide precise and useful rental price predictions by using web scraping to gather data from multiple property listings, machine learning techniques for data analysis, linear regression for relationship modeling [2] and data mining for more in-depth understanding.

We can address the complexity of rental property markets by using a data-driven approach thanks to the integration of these techniques. The wealth of information from previous studies and industry best practices serves as the foundation for this investigation, [3].

The project's main goal is to help lessors and tenants make educated choices regarding the cost and type of accommodations they will need. Its primary emphasis is on utilizing deep learning to forecast real estate rental prices. This forecast is supported by well-trained models that make use of extensive databases that include a variety of properties in different locations with unique attributes. But sometimes, during execution, difficult factors like building condition and furniture condition are overlooked. For example, the age of the building may cause even the most opulent properties to lose value, illustrating just one of the many complexities in this dynamic market.

2 Web Scraping and Data classification

2.1 Web Scraping

One automated technique for obtaining data from websites is web scraping. This approach is typically used to gather a sizable amount of data for a number of uses, such as machine learning and data analysis. The

user must examine the page from which the data is to be extracted and recognize the desired features in order to accomplish this, The scraping technique is described in Figure 1.

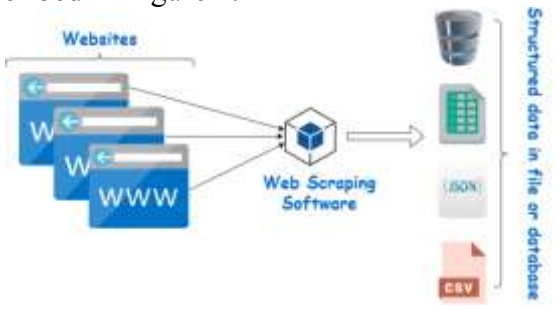

Fig. 1: Web Scraping

2.2 Data Classification

The classification of data in a Machine Learning project is an essential step. To work with this type of data, it must be integer or real data. To do this, we need to adopt a reliable and precise classification to characterize the location and type of dwelling. Let's take the example of an S+3 apartment in deluxe location with a surface area of 120m² and containing a parking space and a garden. This example will be expressed by a matrix containing the characteristics in Table 1 (Appendix).

After collecting our data and saving them in an Excel file, we set the features of our model with this model "Title, Location, Number of rooms, Surface area, Garden, Swimming pool, Parking space".

3 Feature Engineering

Feature engineering is a process used to detect noise in the database. Noise can be due to a measurement error or even false information.

In Machine Learning, it is strictly recommended to focus on the quality of the information, as this will reflect the reliability of the desired result. We're going to examine this type of uncertainty in order to increase the accuracy of our model, Figure 2 shows the data preparation flowchart.

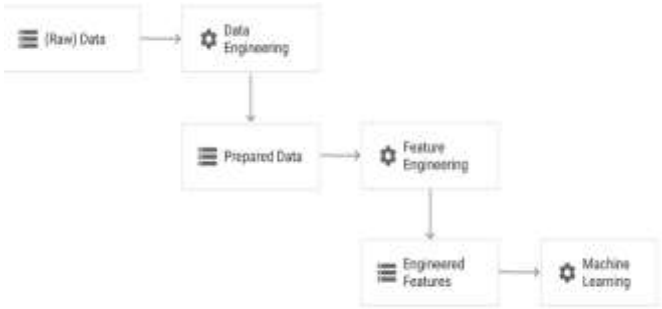

Fig. 2: Data preparation flowchart

After preparing our database, we noticed the existence of a few individuals who's Target (property rental price) is illogical.

Example: A 2-room apartment with a parking space is worth 0 dinars. We therefore need to eliminate this type of data from our database to avoid disrupting our model's learning process. We use the Dropna function: this is a predefined Python function for deleting individuals whose feature values are empty.

4 Data Analysis

Data analysis is the process of examining data using graphs and curves to understand the evolution and relationships between features.

Data analysis is therefore the set of processes that transform raw data into usable information. It involves techniques such as data mining, data visualization, statistical modeling and machine learning, [4].

There are several methods of data analysis, including:

- Descriptive analysis: this explores the characteristics of data using descriptive statistics such as mean, standard deviation etc, [5].
- Exploratory analysis: enables data to be visualized using techniques such as histograms, scatter plots etc.
- Correlation analysis: measures the relationship between two variables. (Example: number of pieces and price) to identify which ones have a strong influence.

Data visualization is fundamental to the creation of a housing price prediction model. It enables the detection of noise and outliers, [6].

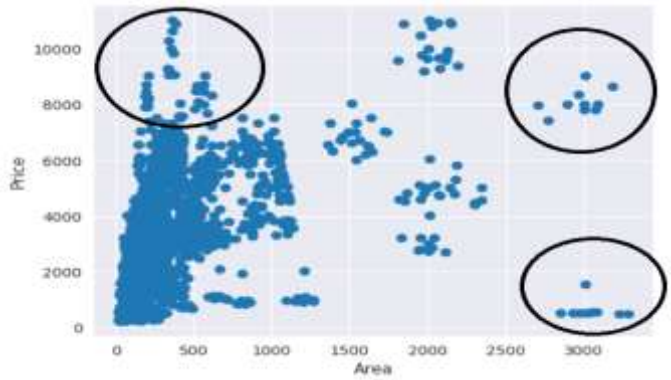

Fig. 3: Noise identification

Referring to Figure 3, we were able to visually identify a few outliers that will degrade the quality of the prediction.

Example:

A dwelling with an area of over 3000m² is worth

less than 500 dinars. Also, a dwelling with an area of less than 200m² is worth more than 10000 dinars. This is not the case in our database.

Faced with this situation, we need to remove these individuals from our database so as not to affect the prediction.

In the same way, we have improved the data for other features, in particular the number of rooms.

To better understand what we're talking about, a distribution analysis of the variables is essential. Hence, the features in our database will be described by the following histograms (Figure 8 in Appendix).

5 Descriptive Analysis

In this section, we'll look at calculating the mean and standard deviation to describe the price distribution of our database. The price distribution of the dataset after enhancement is shown in Figure 4 and Figure 5, respectively, in order to accomplish the discriptive analysis. On the other hand, Figure 6 displays the number of available homes according to price.

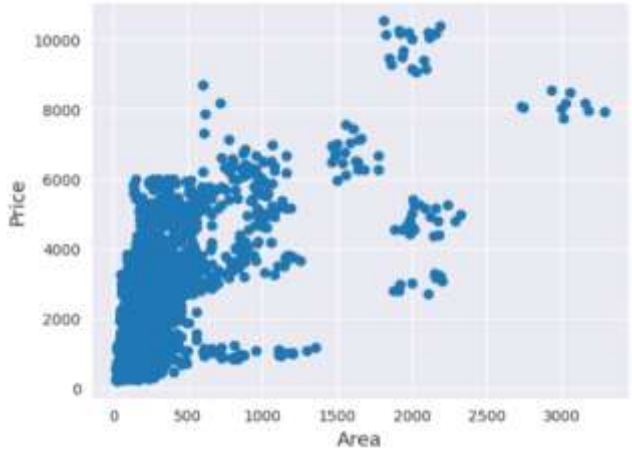

Fig. 4: Distribution after enhancement

As a result, when the surface area increases considerably, so does the price. The same goes for the number of rooms. Our analysis will therefore be based on these two variables, [7].

Referring to Figure 7, we can see that there are individuals whose number of rooms are greater than or equal to 8 with a price above 1500 dinars. These are outliners that we have not been able to detect visually. So we still need to work on the quality of the database, [8].

After rectification, we need to work on the part number histogram to understand the part number distribution of the properties in our database, and to find out the most frequent part number in them.

As a result, each of the target variable's explanatory variables must be set apart. Hence we need to group our base into 4 new variables:

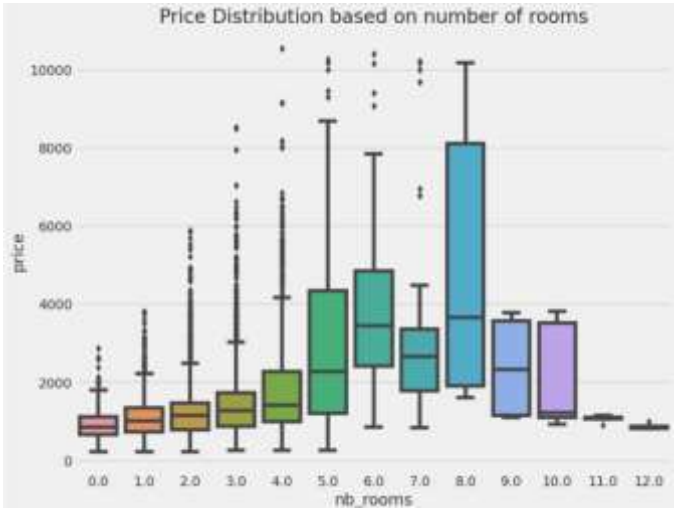

Fig. 5: Price distribution by number of rooms

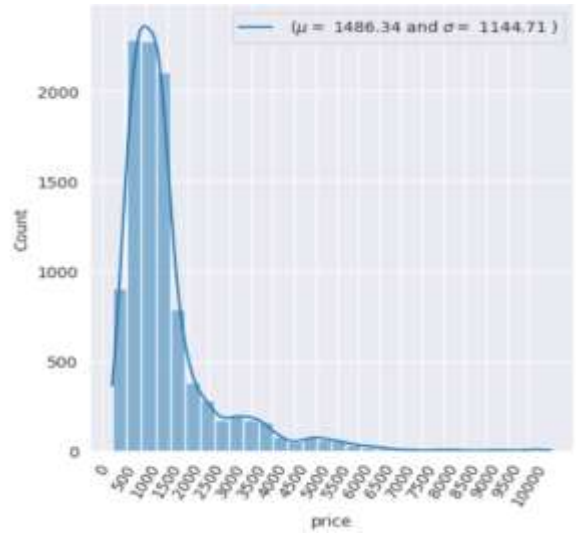

Fig. 6: Number of homes available according to price

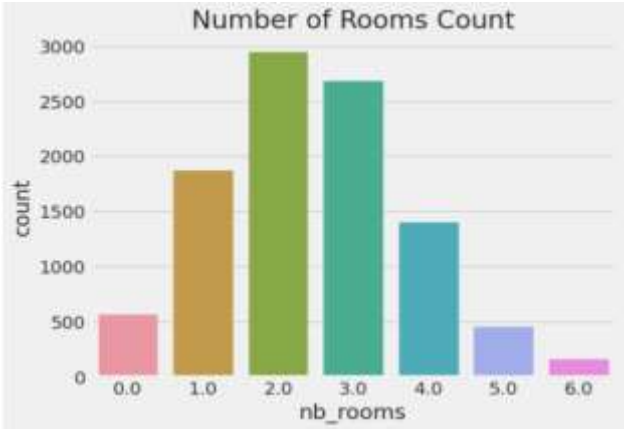

Fig. 7: Histogram number of rooms

From this histogram, we can draw that the most

frequent number of pieces in our database is:

- 1 room: with a number more than 1600.
- 2 rooms: with a number almost equal to 3000.
- 3 rooms: with a number over 2600.
- 4 rooms: with a number almost equal to 1400.

Similarly, surface area is the most important variable in our model, and the one that contributes most to price variation.

6 Prediction with Linear Regression Model

A linear regression model creates a mathematical function by exploiting the project's explanatory variables (number of rooms, surface area, etc.) to establish a relationship between them and our single target variable, price.

To do this, the model draws a straight line representing a mathematical relationship between the variables, which will be used as a reference to predict prices.

X_train , Y_train : for training the model on the explanatory variables.

- X_test , Y_test : to test 20% of the explanatory variables in our database.

X: the variable that must contain the explanatory variables

Y: the variable that should contain the target variables

To calculate MAE, you can use the following formula [9-10]:

$$MAE = \left(\frac{1}{n}\right) * \sum \left| Y_{\_i} - \hat{Y}_{\_i} \right| \quad (1)$$

Where:

MAE is the Mean Absolute Error.

n is the number of observations.

y_i represents the actual values.

ŷ_i represents the predicted values.

Σ indicates the sum over all observations.

In summary, MAE is a measure of the average distance between predictions and actual values, and it is often used to evaluate the performance of a model or forecasting method.

The equation for the coefficient of determination, often referred to as R-squared (R^2), is as follows [11], [12]:

$$R^2 = 1 - \left(\frac{SSR}{SST}\right) \quad (2)$$

Where:

R^2 is the coefficient of determination.

SSR is the sum of the squared residuals (the differences between the observed values and the values predicted by the regression model).

The total sum of squares, or SST for short, is a representation of the whole variation in the dependent variable. It is the total of the squared differences between the dependent variable's mean and observed values.

R-squared calculates the percentage of the dependent variable's overall variation that the independent variables in a regression model account for. Usually, it falls between 0 and 1, where 0 means that no variation in the model is explained and 1 means that all variation in the model is explained. A model that fits the data better is indicated by an R-squared value that is closer to 1, whereas a value that is closer to 0 denotes a poor fit, [13], [14].

From Figure 10 (Appendix), and by comparing R-squared and MAE of the two models, we can draw that both models have a good prediction with a more or less low error rate (MAE).

The Random Forest model is more accurate than the Linear Regression model, since: R-Squared (Linear Regression) < R-Squared (Random Forest) and MAE (Random Forest) < MAE (Linear Regression), [11], [12], [13], [14].

This suggests that the Random Forest model is better suited to our database, and can provide more accurate predictions. However, it's important to note that this doesn't necessarily mean that the Random Forest model is always the best option for all regression problems.

Figure 9 displays the Linear Regression Model, the Random Parameter Initial Model, and the Machine-To-Find Final Model.

The performance evaluation of two predictive models, Random Forest and Linear Regression, is shown in the Table 2. Mean Absolute Error (MAE) and R-Squared (R^2) are the evaluation metrics that are utilized. The Linear Regression model mean absolute error (MAE) is 414.909, meaning that there is an average deviation of 414.909 units between its predicted and actual values. With an R-Squared value of 0.655, the model accounts for 65.5% of the variability in the dependent variable. By contrast, the Random Forest model performs better, with an average deviation of 321.974 units, indicating more accurate predictions. It also fits the data better than the Linear Regression model, with an R-Squared value of 0.794, which means it accounts for 79.4% of the variability in the data.

Table 2. Performance evaluation of two models

	MAE	R-Squared
Linear Regression	414.909	0.655
Random Forest	321.974	0,794

7 Conclusion

The Real Rental Price Prediction project remains a necessity, as it facilitates the valuation of a property as well as relocation.

To get a good quality prediction, you need to go through all the steps mentioned in this paper, including data collection, filtering, the right choice of rankings for the explanatory variables and, above all, data analysis, since the latter has a strong effect on the quality of the estimate. Finally, the reliability of the models created should always be calculated, to facilitate the choice of the right model.

We've been able to analyze the data we've collected using histograms and distribution curves. This part is fundamental in the realization of such a project. In fact, it enables us to detect outliers and filter the database to obtain accurate, relevant data. Also, we have used to the correlation study, we were able to identify the most important explanatory variables that contribute most to the increase in property rental prices, such as surface area and number of rooms.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors utilized ChatGPT and GENEMI for information gathering and assistance in manuscript preparation. The authors reviewed and edited the content as necessary and take full responsibility for the final content of the publication.

References:

[1] Sun, Y., Wu, G., Wei, W., & Li, Y. (2016). Predicting Housing Prices with a Hybrid Model. *Procedia Computer Science*, 91, 866-872. doi:10.1016/j.procs.2016.07.356

[2] Kim, H. J., & Kang, J. (2018). Predicting rental prices of rental houses using a machine learning approach. *Expert Systems with Applications*, 95, 48-57. doi:10.1016/j.eswa.2017.11.022

[3] Zhang, X., Zheng, L., Ma, Z., & Wu, D. (2017). A prediction model of housing rental price in Airbnb. In *Proceedings of the 2017 International Conference on Management, Education and Social Science (ICMESS 2017)*. Atlantis Press.

[4] Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020). Crop prediction using machine learning. *In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE.

[5] Yoshida, T., Murakami, D., & Seya, H. (2022). Spatial prediction of apartment rent using regression-based and machine learning-based approaches with a large dataset. *The Journal of Real Estate Finance and Economics*, 1-28, https://doi.org/10.1007/s11146-022-09929-6.

[6] Neloy, A. A., Haque, H. S., & Islam, M. M. U. (2019). Ensemble learning based rental apartment price prediction model by categorical features factoring. *In Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, Association for Computing Machinery, ISBN: 978-1-4503-6600-7.

[7] Ma, Y., et al. (2018). Estimating warehouse rental price using machine learning techniques. *International Journal of Computers Communications & Control*, 13(2), 235-250.

[8] Wang, K., Zhao, H., & Li, J. (2023). Machine Learning-Based House Rent Prediction Using Stacking Integration Method. *American Journal of Management Science and Engineering*, 8(2), 50-55.

[9] Balti, A., Lakhoua, M. N., & Sayadi, M. (2024). Overview of Smart City Technologies: A Case Study of Designing a Multi-Service Smart Kiosk for Citizens. *International Journal of Computers*, 9, 12-21. IARAS Journals.

[10] Ala, B., & Najah, L. M. (2024). Tutorials and mobile learning in higher education: Enhancing and accessibility. *Advances in Mobile Learning Educational Research*, 4(1), 920-926.

[11] Balti, A., Sayadi, M., & Fnaiech, F. (2012, October). Invariant and reduced features for Fingerprint Characterization. *In IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society* (pp. 1530-1534). IEEE.

[12] Balti, A., Khelifa, M. M. B., Hassine, S. B., Ouazaa, H. A., Abid, S., Lakhoua, M. N., & Sayadi, M. (2022, May). Gait analysis and detection of human pose diseases. *In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)* (Vol. 1, pp. 1381-1386). IEEE.

[13] Balti, A., Sayadi, M., & Fnaiech, F. (2011, March). Segmentation and enhancement of

fingerprint images using K-means, fuzzy C-mean algorithm and statistical features. *In 2011 International Conference on Communications, Computing and Control Applications (CCCA)* (pp. 1-5). IEEE.

[14] Balti, A., Yassin, M., Lakhoua, M. N., & Sayadi, M. (2023, December). Predicting Laptop Prices in the Tunisian Market Using Data Mining and Machine Learning Methods. I*n 2023 IEEE Third International Conference on Signal, Control and Communication (SCC)* (pp. 1-6). IEEE.

APPENDIX

Table 1. Housing classification methods

Title			Location			Number of rooms	Garden	Swimming pool	Parking space
Villa	Apartment	House	Level 1	Level 2	Level 3	1 to 5	0 or 1	0 or 1	0 or 1
1	2	3	1	2	3				

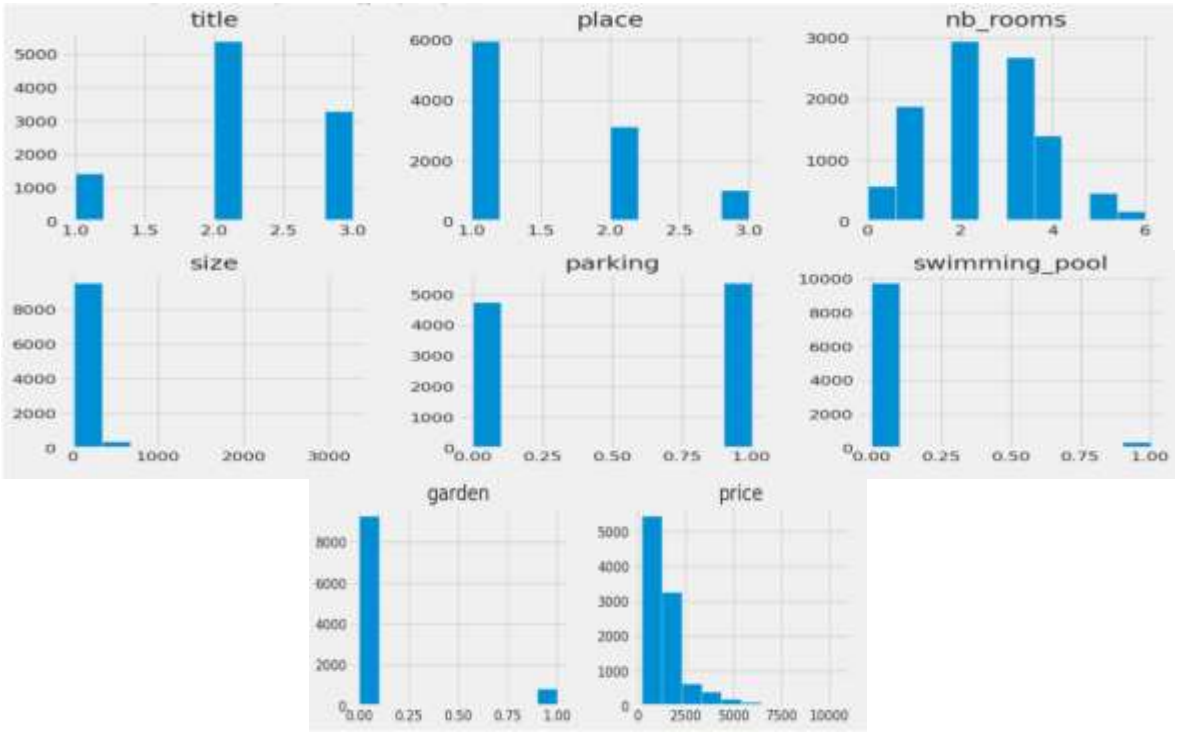

Fig. 8: Distribution of variables

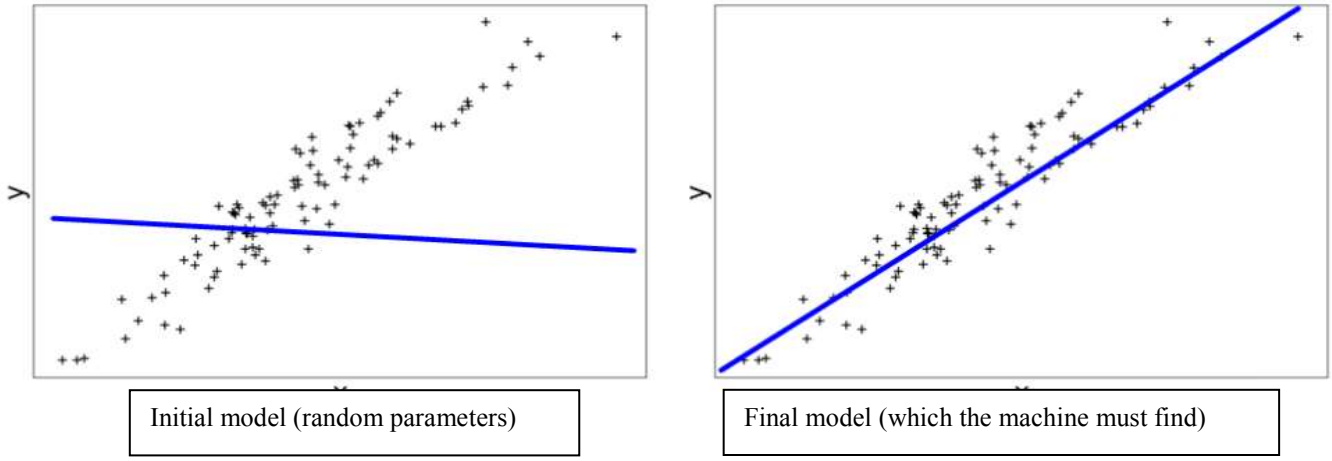

Fig. 9: Linear Regression Model

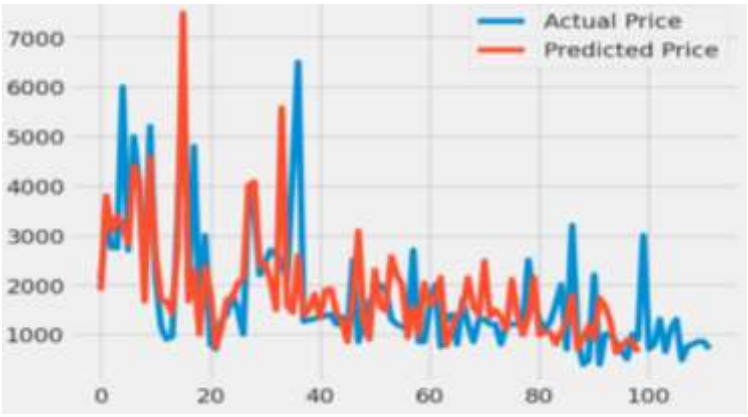

Fig. 10: Distribution of actual and predicted prices from the Linear Regression model

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
AB wrote the program, created the study design, and developed and revised the text. MNL directed the idea, carried out the research, and revised the finished work. MS examined the studies that were connected to the literature on concepts of forecasting and prediction.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The authors have no conflicts of interest to declare.