International Journal of Education, Information Technology and Others (IJEIT), August 2024, 7 (3), 191-199
DOI: https://doi.org/10.5281/zenodo.13688260
p-ISSN: 2654-2528 e-ISSN: 2623-2324
Accredited by Directorate General of Strengthening for Research and Development
Available online at https://jurnal.peneliti.net/index.php/IJEIT

Prediction of Boarding House Rental Prices Using Multiple Linear Regression Method

Nadif Rayhan Julio Mohede[1*], Basuki Rahmat[2], Kartini[3]

[1,2,3]Department of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur

Received: June 5, 2024
Revised: June 10, 2024
Accepted: June 13, 2024

Abstract

Given the significant influx of students to Surabaya, there is a high demand for affordable boarding house rental near universities. One of the issues that arises is figuring out the rent fee set by the landlord.. This study will focus on solving the problem using machine learning with Multiple Linear Regression (MLR) method. This study also focuses on developing a predictive model for temporary housing rental prices around the National Development University "Veteran" Java. Key variables include rental price, room type, room size, availability of air conditioning, WiFi, private bathrooms, kitchen access, 24-hour access, and distance to the university. The dataset was split into training and testing sets (80:20 ratio) for model development and evaluation. The MLR model achieved an R^2 value of 0.76, an RMSE of 211555.8, and a MAPE of 0.18, indicating high predictive accuracy.

Keywords: Machine Learning, Multiple Linear Regression, Temporary Housing Rental.

(*) Corresponding Author: nrayhan26@gmail.com

How to Cite: Mohede, N. R., Rahmat, B., & Kartini, K. (2024). Prediction of Boarding House Rental Prices Using Multiple Linear Regression Method. *International Journal of Education, Information Technology, and Others*, *7*(3), 191-199. https://doi.org/10.5281/zenodo.13688260

INTRODUCTION

Education is very important for the progress of a country, especially in facing global competition and filtering foreign cultural influences to suit Indonesian culture (Amilia et al., 2019). Education plays an important role in forming a society that is able to face modern challenges, including social and cultural changes in urban areas. Dense urban life, such as in Surabaya, creates a need for affordable housing, especially for students and workers.

Surabaya, the largest city in East Java and the second largest in Indonesia after Jakarta, is a major destination for students seeking quality education (BPS, 2022). Many students from outside the city come to Surabaya every year (Adibhadiansyah et al., 2016). The need for affordable housing in the city has made boarding houses a popular solution, especially around university campuses. However, determining a boarding house rental price that is fair and in accordance with the facilities provided is a challenge for boarding house owners.

The East Java "Veteran" National Development University, with its student population continuing to grow, is experiencing an increase in demand for boarding houses. Boarding house owners need a solution to determine competitive and fair rental prices. One proposed solution is to use data mining. Data mining is defined as a set of techniques used automatically to explore and uncover complex relationships in large data sets (Siregar et al., 2017).

The Multiple Linear Regression method can be used as a solution because it is included in data mining and is a data analysis tool that is often used. Multiple Linear Regression is a linear regression model that involves more than one independent variable or predictor (Febriyanto et al., 2020). This method is suitable for use in research that has more than one independent variable and has a high level of accuracy (Utomo et al., 2021).

This research aims to develop a prediction model for boarding house rental prices around the National Development University "Veteran" East Java using the Multiple Linear Regression algorithm. This model will consider variables such as room size, boarding house type, facilities (AC, internet, private bathroom), 24-hour access, kitchen access, and distance to campus. By implementing this algorithm, it is hoped that boarding house owners can determine rental prices that meet the needs of prospective tenants and maintain competitiveness in the market.

RESEARCH METHOD

This research aims to obtain a prediction model for boarding house rental prices within the scope of UPN "Veteran" East Java. The stages of this research were carried out in several stages as shown in Figure 1.

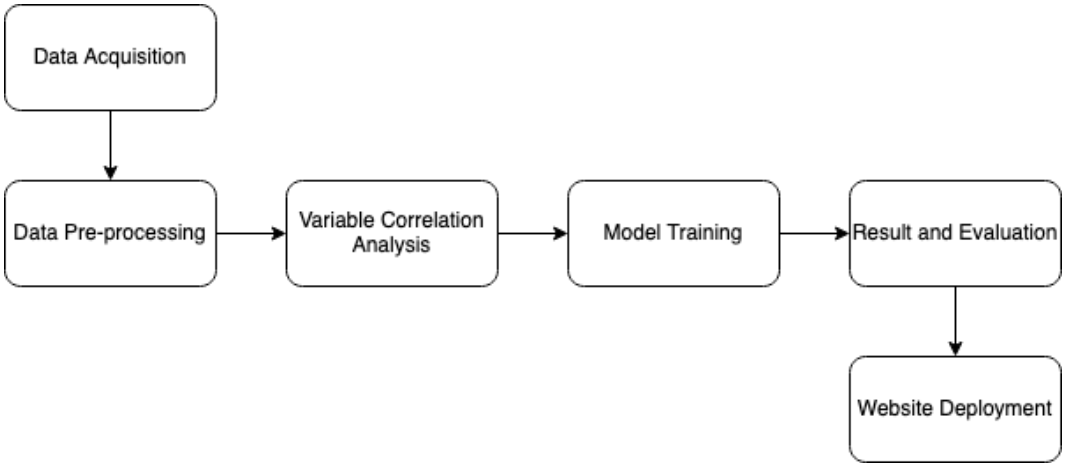

Figure 1. Stages of the research

Data Acquisition

In collecting data, a literature study was carried out first by searching and collecting various references through scientific journals related to predictions of production results.Then the data that will be used as a dataset in this research is secondary data from scraping results on the mamikos.com website which contains boarding house data in the city of Surabaya, with collection time in February 2024. Boarding house data obtained from scraping results of 200 boarding houses in area around UPN Veteran "East Java" with the variables used in this research, namely rental price, boarding house type, room size, AC, WiFi, private bathroom, kitchen access, 24 hour access, distance to UPN "Veteran" East Java. An example of the data model collected is as in the table below.

Jarak Menuju UPN (X_1)	Luas Kamar (X_2)	Tipe Kost (X_3)	AC (X_4)	WiFi (X_5)	Kamar Mandi Dalam (X_6)	Akses 24 Jam (X_7)	Akses Dapur (X_8)	Harga Sewa (Y)
Dekat	Sedang	Campur	Tidak	Tidak	Tidak	Iya	Iya	650000
Sedang	Sedang	Putra	Iya	Iya	Iya	Iya	Tidak	1500000
Dekat	Besar	Campur	Iya	Iya	Iya	Tidak	Iya	2800000
Dekat	Kecil	Putri	Iya	Iya	Iya	Tidak	Iya	1550000
Dekat	Kecil	Putra	Tidak	Iya	Tidak	Tidak	Iya	900000

Figure 2. Example of data model

The distance from the boarding house location to UPN "Veteran" East Java is divided into two categories, namely near and far. Assuming that if the distance to UPN "Veteran" East Java is below or equal to 1 kilometers then the distance is close, if the distance to UPN "Veteran" East Java between 1 kilometers to 2 kilometers than the distance is medium, and if the distance to UPN "Veteran" East Java is above 2 kilometers then the distance is considered far. Then the room area is divided into 3 categories, namely small, medium and large. Assuming that if the room area is below or equal to 9m^2 then the room is small, if the room size is more than 9m^2 or less than 12m^2 then the room size is medium size, and if the room size is more than 12m^2 then the room size is large.

Data Pre-processing

After collecting data, the next step is to carry out the data pre-processing stage with the procedure of changing categorical data into numerical data. The data value will be changed with the following assumptions: if the data is "Tidak" (No), "Kecil" (Small), or "Jauh" (Far), it will be converted to 0; if the data is "Iya" (Yes) or "Sedang" (Medium), it will be converted to 1; and if the data is "Besar" (Large) or "Dekat" (Near), it will be converted to 2.

Variable Correlation Analysis

Once data pre-processing is complete, variable correlation analysis will be performed to check the level of correlation between the independent variables (predictors) and the dependent variable. The goal of this analysis is to identify how strong the correlation is between each predictor and the dependent variable in the regression model, which helps understand the relative contribution of each variable to the model. In this research, the correlation matrix is used to analyze the relationship between variables. A correlation matrix is a table that shows the correlation coefficient between two or more variables, with values ranging between -1 and 1. Negative values indicate a negative relationship, while positive values indicate a positive relationship, and the closer the value is to 1 or -1, the stronger the relationship between these variables. A value of 0 indicates there is no linear relationship between these variables.

Model Training

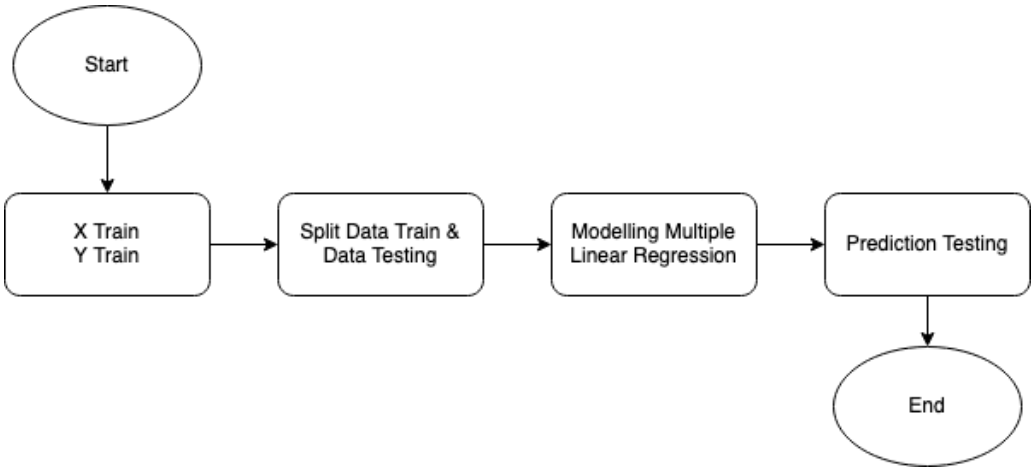

Figure 3. Flowchart of the model training

At this stage, model training is carried out using Multiple Linear Regression after data pre-processing and variable correlation analysis have been carried out. The dataset is then divided into variable x as the input variable and variable y as the output variable. Variable x data contains the distance to UPN "Veteran" East Java, room size, AC, WiFi, private bathroom, kitchen access and 24 hour access. For variable data, y is data on the rental price of the boarding house. After that, the data is separated into training data and testing data, with a ratio of 80% for training data and 20% for testing data. Then apply Multiple Linear Regression to the training dataset and after that use the model to predict rental price values on the testing dataset.

Results and Evaluation

The next stage is testing predictions with a dataset that has passed data pre-processing, followed by evaluating the trained model. Model evaluation was carried out using the R^2 (R-Squared), RMSE (Root Mean Squared Error), and MAPE (Mean Absolute Percentage Error) methods. R^2 measures how well the data fits the regression model, with values ranging from 0 to 1, where values closer to 1 indicate the model is good at explaining variations in the data. RMSE measures how well a model predicts the true value by calculating the root mean square of the difference between the predicted value and the true value, with smaller values indicating less prediction error and better model performance. MAPE measures relative error in the form of a percentage of the true value, providing an idea of the extent to which the model is experiencing error in predicting the data. The use of these three evaluation methods provides a holistic assessment of the accuracy of model predictions. The test results will produce output in the form of R^2 values, RMSE values, and MAPE values.

Website Implementation

The final process is installing the model on the website, where this process occurs after training and testing of the model is complete. Before that, the model is first saved in a file in .pkl format using the pickle library. After the model is saved in the file, the model can be installed on the website using the available flask framework to connect the model file to the website. The flask framework will run on files using the Python programming language and for website display using the HTML and CSS programming languages.

RESEARCH RESULTS AND DISCUSSION

The following section discusses the results related to prediction of boarding house rental prices using multiple linear regression method.

Data Pre-processing

	Nama_Kost	Tipe_Kost	Harga_Sewa	Luas_Kamar	AC	Kamar_Mandi_Dalam	WiFi	Akses_24_Jam	Akses_Dapur	Jarak_Menuju_UPN
0	Kost Medokan Sawah 01	Campur	650000	Sedang	Tidak	Tidak	Tidak	Iya	Iya	Dekat
1	Kost Omah Ibu Tina	Putra	1500000	Sedang	Iya	Iya	Iya	Iya	Tidak	Sedang
2	Kost Ivorya Tipe Avanthi	Campur	2800000	Besar	Iya	Iya	Iya	Tidak	Iya	Dekat
3	Kost Vira 2	Putri	1500000	Sedang	Iya	Iya	Iya	Iya	Tidak	Dekat
4	Kost Rumah KosQ	Putri	1550000	Kecil	Iya	Iya	Iya	Tidak	Iya	Dekat

Figure 4. Dataset before transformation

[101]

	Tipe_Kost	Harga_Sewa	Luas_Kamar	AC	Kamar_Mandi_Dalam	WiFi	Akses_24_Jam	Akses_Dapur	Jarak_Menuju_UPN
0	2	650000	1	0	0	0	1	1	2
1	0	1500000	1	1	1	1	1	0	1
2	2	2800000	2	1	1	1	0	1	2
3	1	1500000	1	1	1	1	1	0	2
4	1	1550000	0	1	1	1	0	1	2

Figure 5. Dataset after transformation

Based on the information in figure 3 and figure 4, the pre-processing stage is carried out with the procedure of changing categorical values into numerical values and deleting columns that are not needed in the next process, namely the "boarding_name" column. The aim of removing the “boarding_name” column is to help reduce data complexity and increase the efficiency of the modeling process.

Variable Correlation Analysis

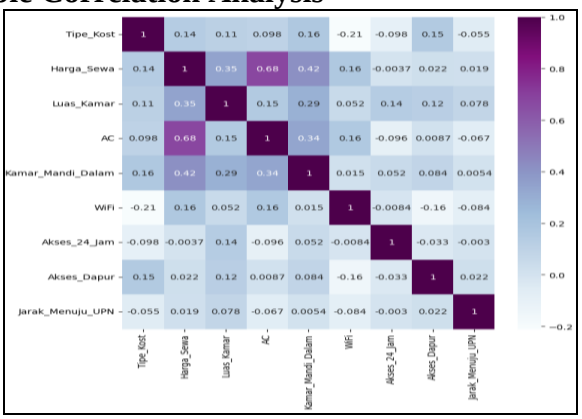

Figure 6. Matrix correlation between variables

Figure 5 is a schematic of the relationship between variables which is visualized in the form of a correlation matrix. Judging from the figure above, the following information can be obtained:

1. There is a strong positive correlation between Rental Price and AC (0.68). This shows that the presence of air conditioning in boarding houses has a big influence on increasing rental prices.
2. The positive correlation between Rental Price and Room Size is 0.35. This shows that the larger the room area, the higher the rental price.
3. There is a positive correlation between rental price and ensuite bathroom (0.42). This shows that having an en suite bathroom also increases the rental price.
4. The correlation between rental price and boarding house type is 0.14, which indicates a weak positive relationship between boarding house type and rental price.
5. The correlation between AC and room area is 0.15. This suggests that larger rooms are more likely to have air conditioning.
6. There is a negative correlation between WiFi and Boarding House Type (-0.21), which indicates that certain types of boarding houses may provide less WiFi facilities.
7. The correlation between Rental Price and WiFi is 0.16, indicating a very weak relationship between the presence of WiFi and rental price.
8. The correlation between rental price and distance to UPN is 0.019, indicating that distance to UPN has a very weak relationship with rental price.
9. Some variables such as 24 Hour Access, Kitchen Access, and Distance to UPN have a very weak correlation with Rental Prices, indicating that they may not have much influence on boarding house rental prices.
10. A fairly high correlation is also seen between several independent variables, such as between AC and private bathroom (0.34), which shows that boarding houses that have AC tend to also have private bathrooms.

Results and Evaluation

Before getting accuracy results, model training is carried out first. The process begins by separating the features (predictors) and targets (labels). The features are stored in variable Next, the model is initialized and trained using the training data (X_train and Y_train).

Next, the model evaluation process is carried out using model evaluation techniquesR^2, RMSE, and MAPE. R^2 measures the proportion of total variation of the dependent variable that can be explained by the independent variables in the model. The R^2 value ranges from 0 to 1, with information that the closer the value is to 1, the better the model is in explaining all data variables.

RMSE is used to measure how well the model predicts the true value by calculating the root mean square of the difference between the predicted value and the true value. The RMSE value ranges from 0 to infinity, with the information that if the value is smaller, closer to 0, the model has fewer errors in predicting and the model performance is also better. Meanwhile, MAPE is used to measure

relative error in the form of a percentage of the actual value. This gives an idea of the extent to which the model experiences errors in forecasting the data.

```
[44] Hasil Evaluasi Model
0s

Nilai RMSE (Root Mean Squared Error): 197182.47
Nilai R2 (R-squared): 0.79
Nilai MAPE (Mean Absolute Percentage Error): 0.16
```

Figure 7. Evaluation results from the model

From the results of the model evaluation above, the R^2 (R-Squared) value was 0.79, RMSE (Root Mean Squared Error) was 197182.47, and MAPE (Mean Absolute Percentage Error) was 0.16. By getting these values from the three evaluation methods, it can be said that the prediction model works quite well. After the model evaluation has been carried out and the results have been obtained, the process of saving the prediction model is then carried out.

Website Implementation

Figure 8. User interface for the website

Figure 9. User interface for the website

Above is the website display result of the boarding house rental price prediction system which contains 8 facilities that can be selected in the form of an input form with dropdown format. After the user selects each facility, the user can then press the 'Price Prediction' button to find out the rental price according to the facility selected by the user.

CONCLUSION

Implementation of the Multiple Linear Regression (MLR) method to predict boarding house rental prices based on factors such as room size, boarding house type, facilities (AC, internet, private bathroom, 24 hour access, kitchen access), and distance to UPN "Veteran" Jawa East has been carried out through several stages, starting with data pre-processing and variable correlation analysis. The dataset is divided into training data and testing data (80:20), and the MLR model is trained with the training data and tested with the testing data. Evaluation of the model using R^2, RMSE, and MAPE shows an R^2 value of 0.79, RMSE 197182.47, and MAPE 0.16, which indicates that the model is able to explain most of the variability in boarding house rental prices with a small prediction error. This model is then integrated into a website using the Flask framework, allowing users to predict boarding house rental prices based on relevant input variables.

BIBLIOGRAPHY

Adibhadiansyah, M., & Rochmawati, N. (2016). PENGEMBANGAN SISTEM INFORMASI KOS BERBASIS ANDROID. Jurnal Manajemen Informatika, vol. 5, pp. 68–73.

Amilia, S., & Iriyani. (2020). Pengaruh lokasi, Harga Dan Fasilitas Terhadap Keputusan Sewa Kamar Kost Mahasiswa Fakultas Ekonomi Universitas samudra. Jurnal Manajemen dan Keuangan, vol. 8, no. 3, pp. 267–280. doi:10.33059/jmk.v8i3.2328

Deni, M., & Latifah, R. (2022). PREDIKSI PENGISIAN BBM HSD DENGAN METODE MULTIPLE LINEAR REGRESSION. JUST IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer, 11(3).

Fachid, S., & Triayudi, A. (2022). Perbandingan Algoritma Regresi Linier Dan Regresi random forest Dalam Memprediksi Kasus Positif covid-19. JURNAL MEDIA INFORMATIKA BUDIDARMA, 6(1), 68. doi:10.30865/mib.v6i1.3492

Faruqhy, M. N., Andreswari, D., & Sari, J. P. (2021). Prediksi Prestasi Nilai Akademik Mahasiswa Berdasarkan Jalur Masuk Perguruan tinggi menggunakan metode multiple linear regression (Studi Kasus: Fakultas Teknik Universitas Bengkulu). Rekursif: Jurnal Informatika, vol. 9, no. 2, pp. 172–183. doi:10.33369/rekursif.v9i2.17108

Fitri, A., Syahra, Y., & Kustini, R. (2020). Penerapan Data mining Dalam Mengklusterisasi location best PB Tambahan Pada Regional IV PT indomarco prismatama cab.Medan Dengan menggunakan metode k-means. Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer), vol. 19, no. 2, p. 11. doi:10.53513/jis.v19i2.2330

Hayuningtyas, R. Y., & Sari, R. (2022). Implementasi data mining dengan ALGORITMA multiple linear regression untuk MEMPREDIKSI penyakit diabetes. Jurnal Teknik Komputer, vol. 8, no. 1, pp. 40–44. doi:10.31294/jtk.v8i1.11552

James, G., Witten, D., Hastie, T., and Tibshirani, R., An introduction to statistical learning. New York: springer, 2013.

Khder, M. (2021). Web scraping or web crawling: State of Art, Techniques, approaches and application. International Journal of Advances in Soft Computing and Its Applications, 13(3), 145–168. doi:10.15849/ijasca.211128.11

Luthfiarta, A., Febriyanto, A., Lestiawan, H., & Wicaksono, W. (2020). Analisa prakiraan Cuaca Dengan Parameter Suhu, Kelembaban, Tekanan Udara, Dan Kecepatan angin menggunakan Regresi Linear Berganda. JOINS (Journal of Information System), vol. 5, no. 1, pp. 10–17. doi:10.33633/joins.v5i1.2760

Maula, I., Hasanah, L. U., & Tholib, A. (2023). Analisis prediksi Harga Rumah di Jabodetabek MENGGUNAKAN multiple linear regression. Jurnal Informatika Kaputama (JIK), vol. 7, no. 2, pp. 216–224. doi:10.59697/jik.v7i2.135

M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, Applied linear statistical models. New Delhi: Mcgraw-Hill Education (India) Private Limited, 2013.

Novalyn, E. tri, Ginting, G., & Siburian, H. K. (2018). Pemanfaatan metode cart dalam memprediksi omset pakaian pria remaja studi kasus PT. Matahari departement store thamrin plaza medan. Pelita Informatika: Informasi dan Informatika, vol. 7, pp. 199–206.

Prasetyo, A., Salahuddin, S., & Amirullah, A. (2021). Prediksi produksi Kelapa Sawit menggunakan metode regresi Linier Berganda. Jurnal Infomedia, vol. 6, no. 2, p. 76. doi:10.30811/jim.v6i2.2343

Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1467–1474. doi:10.1016/j.dsx.2020.07.045

Rokhayati, Y., Utomo, N. S., & Sartikha. (2021). Prediksi Kelayakan operasional Mesin rivet menggunakan Regresi Linear berganda. Jurnal Sustainable: Jurnal Hasil Penelitian dan Industri Terapan, vol. 10, no. 1, pp. 10–15. doi:10.31629/sustainable.v10i1.2336

Triyanto, E., Sismoro, H., & Laksito, A. D. (2019). Implementasi Algoritma regresi linear berganda Untuk Memprediksi produksi padi di Kabupaten Bantul. Rabit: Jurnal Teknologi dan Sistem Informasi Univrab, vol. 4, no. 2, pp. 66–75. doi:10.36341/rabit.v4i2.666

Zer, R. H., Hayadi, B. H., & Damanik, A. R. (2022). Pendekatan machine learning Menggunakan Algoritma C4.5 Berbasis Pso Dalam Analisa Pemahaman Pemrograman website. Jurnal Informatika Dan Teknik Elektro Terapan, 10(3). doi:10.23960/jitet.v10i3.2700