MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 8 | No. 1 | Halaman 14 - 26
DOI : https://doi.org/10.26760/mindjournal.v8i1.14-26 Juni 2023

House Prices Prediction: Multiple Linear Regression vs Ridge vs Polynomial

JASMAN PARDEDE, RAYYAN

Department of Informatics, Institut Teknologi Nasional Bandung
Email : jasman@itenas.ac.id

Received 19 Desember 2022 / *Revised* 17 Februari 2023 / *Accepted* 13 Maret 2023

ABSTRACT

The phenomenon of falling or rising house prices has attracted the interest of researchers as well as many other interested parties. The house not only be used as a place to live, it is also used as an investment instrument. Errors in determining the price of the house can result in losses. However, with data from developers, machine learning models can be applied for price predictive analysis. Several methods are used such as multiple linear regression, ridge, and polynomial. Model performance was measured using evaluation matrices such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and also Mean Absolute Error (MAE). Multiple linear regression models yielding values of 0.0952, 0.3086, 0.2452, ridge yielding values of 0.0952, 0.3086, 0.2453, and polynomial yielding values of 0.0874, 0.2955, 0.2344. These results prove that the polynomial regression model with a value of degree = 2, coupled with a regularization technique using ridge regression with a value of Alpha = 100 can produces the best performance judging from the value of the error matrix it produces.

***Keywords**: Multiple Linear Regression, Ridge Regression, Polynomial Regression*

ABSTRAK

Fenomena turun atau naiknya harga rumah telah menarik minat dari peneliti juga banyak pihak lain yang berkepentingan. Rumah tidak hanya dijadikan sebagai tempat tinggal, rumah juga digunakan sebagai instrumen investasi. Kesalahan menentukan harga rumah dapat mengakibatkan kerugian. Namun, dengan adanya data – data dari pengembang, pembuatan model machine learning dapat diaplikasikan guna keperluan analisis prediktif harga. Beberapa metode yang digunakan seperti regresi linear berganda, ridge, dan polinomial. Performa model diukur menggunakan matriks evaluasi seperti Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan juga Mean Absolute Error (MAE). Model regresi linear berganda menghasilkan nilai 0.0952, 0.3086, 0.2452, ridge menghasilkan nilai 0.0952, 0.3086, 0.2453, dan polinomial menghasilkan nilai 0.0874, 0.2955, 0.2344. Hasil tersebut membuktikan bahwa model regresi polinomial dengan nilai degree = 2, ditambah dengan teknik regularisasi regresi ridge dengan nilai Alpha = 100 dapat menghasilkan performansi terbaik dilhat dari nilai matriks error yang dihasilkannya.

***Kata kunci**: Regresi Linear Berganda, Regresi Ridge, Regresi Polinomial*

1. INTRODUCTION

House is one of the main human needs besides clothing and food. In the Hierarchy of Needs, Maslow stated that the house is one part of the Basic Needs which can be categorized into 2 parts, namely Physiological Needs or Safety Needs **(Mcleod, 2018)**. The need for houses is not only used as a place to live, they are also often used as long-term investment instruments by investors to get additional income from these types of investments, especially for property entrepreneurs who certainly produce investments that can be considered quite promising **(Utama, 2011)**.

If the seller of the house makes a mistake, the result will be that the house will be less accepted on the market, this can make the house sellers have a smaller possibility or even lose the opportunity to get the maximum profit from the sale. To minimize errors in setting the selling price of the house, the seller must be careful in determining the price, that's because the selling price of the house is mostly raising and almost never goes down either in the short or long term **(Hendra et al., 2017)**.

The technique of making predictions or forecasting is one of the important tasks of a data scientist for many activities within an organization. Generating forecasts or predictions of good quality is also not an easy matter for both engines and most analysts **(Taylor & Letham, 2018)**. In 2015, research conducted by **(Gallo, 2015)** explained that regression analysis can be used to make predictions about the future, for example, predicting total product sales for the next six months, or also predicting the profit earned from sales for one month.

The previous studies using different datasets, a research was carried out using the multiple linear regression method compared to the decision tree regression method. This research was conducted to predict the selling price of houses. It is proved from the resulting error matrix that multiple linear regression has a slightly smaller error than that produced by the decision tree regression method **(Thamarai & Malarvizhi, 2020)**.

In 2021, a research conducted by **(Zhou, 2021)** was predicting house prices using the same dataset in this study, the method used in this study is polynomial regression combined with the PSO (Particle swarm optimization) technique. The results of the model evaluation are R-square with a value of 0.88, and it can be said to be good enough to make predictions in the case of house prices. However, the resulting R-square value is only obtained from the results of the model development iterations. The dataset is not divided into training data and test data so the model have a very high possibility of overfitting.

Based on the results of these previous studies, this research tries to compare the performance of several regression techniques such as multiple linear regression, ridge regression, and polynomial regression. Comparison of the performance of the three regression methods will be assessed from the several error matrices they produce at the model evaluation stage, the error matrices used are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Finally, based on the model evaluation results of the three regression techniques, the regression model that has the best performance can be used and implemented for the development of a simple website-based application for predicting the selling price of a house.

2. METHODS

2.1 Main Flowchart

In the main flowchart, it can be seen that the house prices sales listing dataset in King County, USA will be used. House sales data in the dataset were collected and obtained from May 2014 to May 2015. The amount of data available is 21,613 data, and has 21 features in it. The dataset is taken from the official Kaggle website. During Data Pre-processing, the dataset will first be analyzed using graphical depictions and data checking using several data visualization techniques as well as Exploratory Data Analysis (EDA). Furthermore, based on the results of data visualization and analysis, the data that will be used for modeling needs to cleaned first.

The data that has gone through the pre-processing stages of the data will then be divided into 2 parts, where the first part of the data will be used as training data, and the data in the second part will be used as a testing data. The training data consists of 15,129 data while the test data contains 6,484 data. After dividing the data, the next step is to use training data first, model training is carried out using 3 regression techniques, namely multiple linear regression, ridge regression and also polynomial regression. Each regression model generated by each technique will be tested using testing data before then evaluating each model it produces to determine the value of Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) in each model, these processes can be seen in Figure 1.

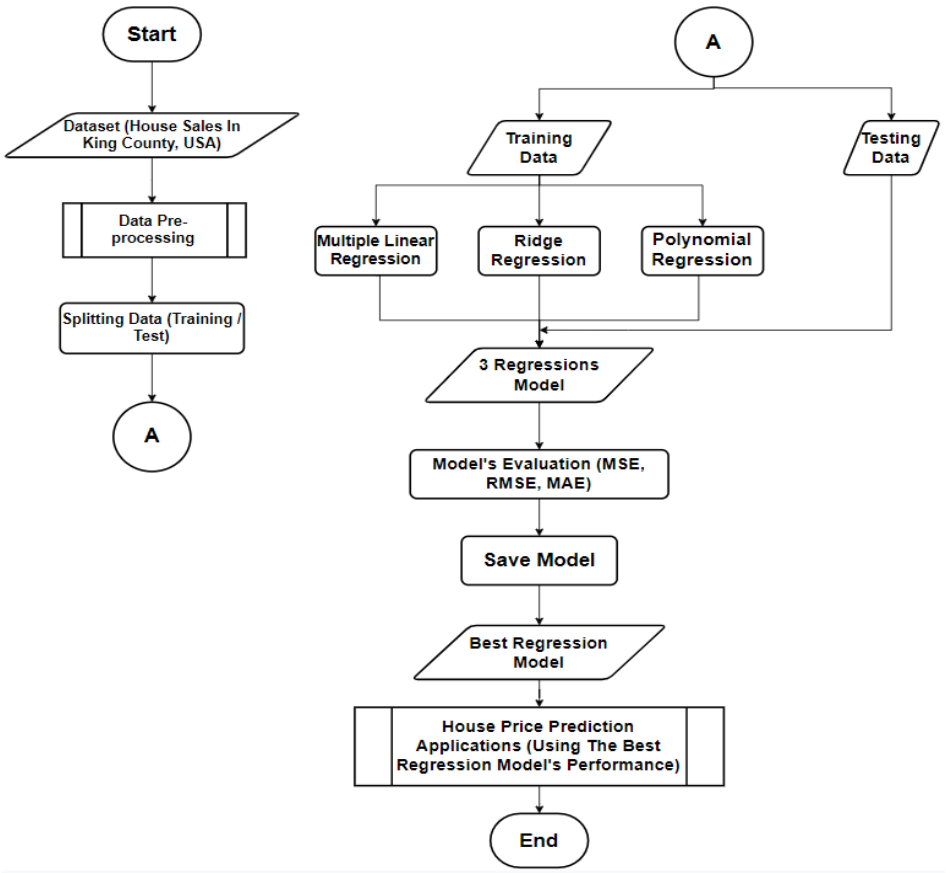

Figure 1. Main Flowchart

The evaluation results of each model will be compared with each other to find out which regression technique produces the model with the best performance based on the experimental results and analysis. The best model produced will then be stored for later use in making web-based applications in the case of predicting house prices.

2.2 Data Pre-processing

During the pre-processing stage, the dataset is first analyzed by depicting graphs and checking data using several data visualization techniques and also Exploratory Data Analysis (EDA). Based on the results of data visualization and analysis, the data that will be used for Machine Learning modeling needs to be cleaned first. The flowchart of Data Pre-processing can be seen in Figure 2.

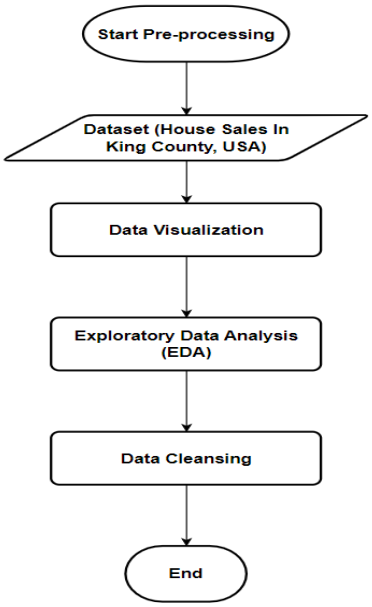

Figure 2. Data Pre-processing Flowchart

After carrying out all the pre-processing stages of the data, the data that is finally ready to be used for Machine Learning modeling using regression techniques.

2.3 House Prices Prediction Application

After completing the performance comparison of multiple linear regression models, ridge regression, and polynomial regression based on experiments and analysis, the best model produced will be saved for later use in making web-based applications to predict house prices in America and Indonesia. The flow of the application will look like the flowchart shown in Figure 3.

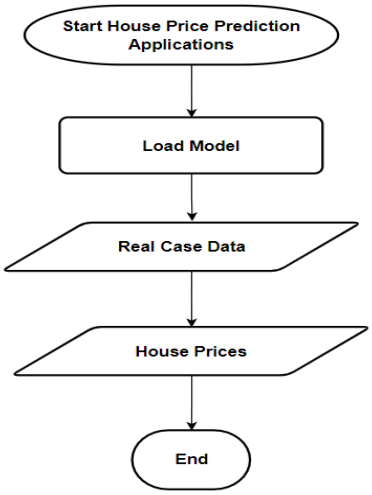

Figure 3. Application Flowchart

2.4 Multiple Linear Regression

Regression is divided into 2 categories, those are simple regression and multiple regression, the thing that distinguishes the two types of regression is that in simple regression there is only 1 independent variable which is used as a predictor to find out the results of the dependent variable. Whereas in multiple regression there are many independent variables that can be used as predictors equipped with their respective coefficients **(Tranmer et al., 2020)**. In research conducted by **(Sarstedt & Mooi, 2019)**, it was also explained that in regression analysis, the other most common method for making or estimating a regression model is called ordinary least squares (OLS). OLS adjusts the regression line to existing data with the aim of minimizing the distance between the predicted results and the actual data (Sum of squared distances). This is done in porpose of not to get negative distance values. By using the OLS technique, the equation formula for finding the intercept value as well as the coefficient of a simple or multiple linear regression model as shown in Equation (1).

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (1)$$

In equation 1, the value sought ($\hat{\beta}$) represents the relationship between each independent variable and the dependent variable. The value of X in the equation is the matrix of the independent variables, while the value of X^T is the transpose of the matrix X. Finally, the value of the variable Y is the dependent variable of the equation.

2.5 Ridge Regression

Ridge regression is a popular parameter estimation technique, ridge regression is commonly used to solve collinearity problems that often arise in multiple linear regression. Basically, ridge regression is very similar to ordinary linear regression, the difference between the two is that in ridge regression the coefficients in the regression model are estimated by minimizing a slightly different quantity, namely smaller **(James et al., 2013)**. In the calculation of the ridge regression, the OLS calculation formula used previously was slightly modified, namely by adding the multiplication between the lambda value (λ) and the X identity matrix, the OLS equation formula for the ridge estimator is shown in Equation (2).

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \quad (2)$$

The value ($\hat{\beta}$) in equation 2 represents the relationship between each independent variable and the dependent variable. The value of X in the equation is the matrix of the independent variables, while the value of X^T is the transpose of the matrix X. The value of the variable Y is the dependent variable of the equation. Lambda (λ) in this equation is the penalty value that will be given to the regression model.

2.6 Polynomial Regression

A Research conducted by **(Ostertagová, 2012)**, explained that in polynomial regression the dependent variable regression was also carried out on the strength of the independent variable. Polynomial regression is an extension and improvisation of standard linear regression, polynomial regression calculations can also use the OLS technique as used in Equation (1) before, but in polynomial regression the independent variables need to be raised to a power first according to the Degree value that are being used. The degree value in polynomial regression, usually used with a value of less than five, this is because when the degree becomes larger, the polynomial model will tend to be too fit with the model so that it can cause an overfitting behavior **(James et al., 2013)**.

3. RESULT AND ANALYSIS

3.1 Dataset

The dataset used in this study contains sales of houses in King County, USA. All house sales data in the dataset were collected and obtained from May 2014 to May 2015. The amount of data available is 21,613 data, and has 21 features in it. The dataset that will be used in this study can be retrieved from the official website of Kaggle House Sales in King County, USA (2014-2015). In accordance with the research's plan, the dataset is divided into 2 parts, namely training data of 70% (15,129 data) and test data of 30% (6,484 data). The distribution of this data is very important to do to prevent overfitting in the Machine Learning model.

3.2 Data Visualization and *Exploratory Data Analysis* (EDA)

During the visualization and Exploratory data analysis (EDA) stages, features in the dataset except for ID, Date, Zipcode, Latitude, and Longitude, are analyzed for the distribution of the data to the target variable, which is the price of a house. The distribution analysis of the data was divided into 2 types, those are the analysis of continuous numerical features, and the second one is the analysis of discrete numerical features. Numerical continuous features, are: bedroom, bathroom, size of living space, size of land area, flooring, size of living space above ground level, size of living space below ground level, year built, year renovated, averages of living space sizes from 15 closest neighbors, averages of land area sizes from 15 closest neighbors. Meanwhile, the features included in the Numerical Discrete Feature category are waterfront, view, condition, and grade. Before diving into the analysis of the two types of features above, the dataset is first checked for the NULL values contained in the features. The results show that there are no features that have a NULL value in it.

The next step is to check the skewness of the data on the dataset. The checking is done by displaying the data distribution value of each feature (Skewness), this needs to be done to find out whether the features in the dataset have normal data distribution or not. After checking, several variables/features have abnormal data distribution, this can be seen from the high skewness value (much greater or less than 0). These features are price, size of living space, size of land space, size of living space above ground level, averages of living space sizes 15 closest neighbors, and averages of land space 15 closest neighbors. Checking the distribution of data is also carried out from each of its features to the target variable, the price. Starting from continuous numerical features like the size of the living space and also the year the house was built have a big influence on the increase in the value of house prices. Other continuous numerical features apart from those two features show a fairly stable distribution of data, where an increase in value does not indicate a significant increase in the selling price of houses.

In features/variables that are numerically discrete, features that have a significant effect on increasing the selling price of a house are View, Waterfront, and Grade. The effect of those features on the increase in the selling price of a house shows an exponential increase in value. Other features such as the condition of the house also affect the increase in the selling price of the house, but the price increase tends to be quite stable with this feature, visualization using bar plots on discrete numerical features can be seen in Figure 4.

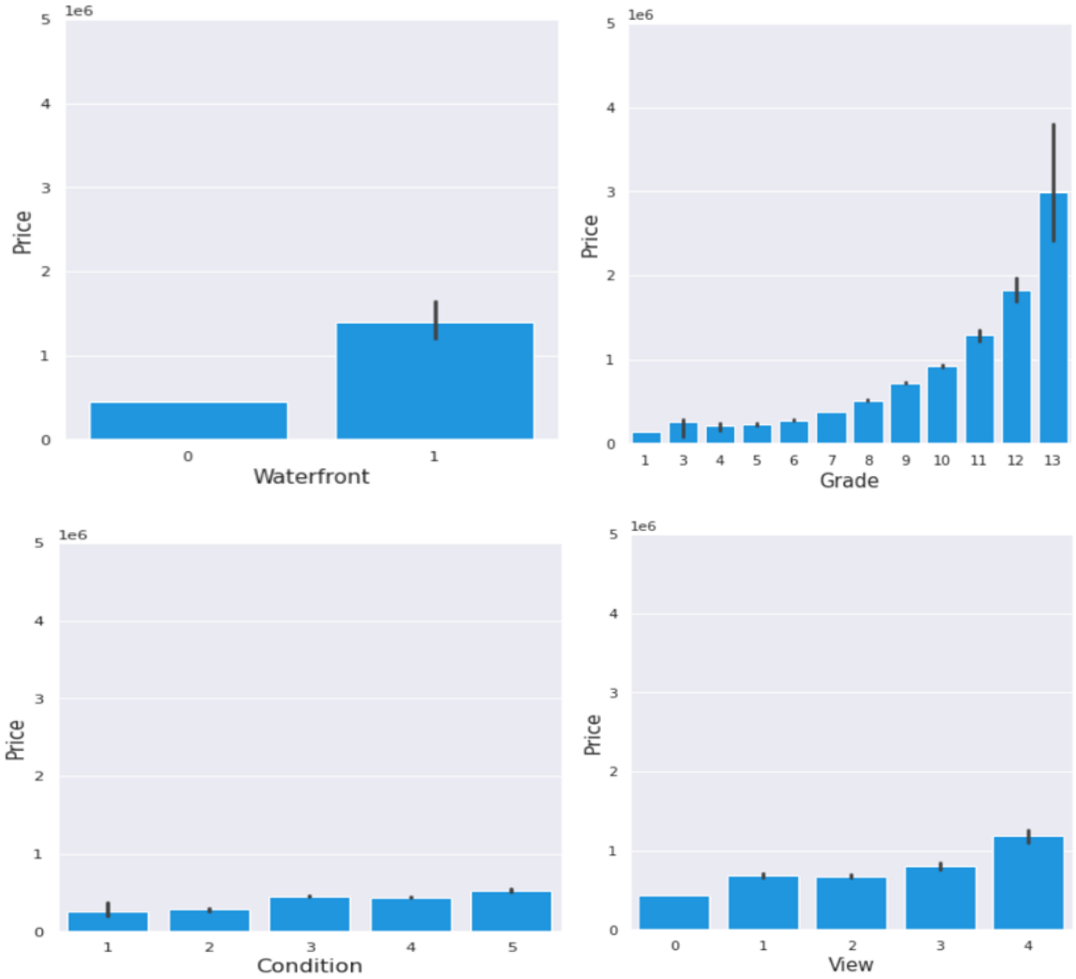

Figure 4. Discrete Numerical Features

3.3 Data Cleansing

Based on the results of data visualization and also Exploratory data analysis (EDA), the next step is to do data cleansing first before the data is ready to be used in the regression modeling. Since there is no data containing NULL values in the dataset, the first step to clean the data is done by dropping several features that will not be used for the modeling process, these features are ID, Date, Zipcode, Latitude, and Longitude features. Then, features that were previously found to have data that was not normally distributed can first be converted in value using a Log Transform. This transformation is done to make data that were previously not normally distributed have now turned into normally distributed data, this can be seen from the skewness value which changes to a range value of 0.

Furthermore, in data processing, the data before it is normally distributed and after the normal distribution is carried out will be separated, this is done to compare the model performance results generated from the two data.

3.4 Experiment Scenarios

Before carrying out regression modeling in this study, there were several experiments carried out on each method, the list of experimental scenarios can be seen in Table 1.

Table 1. Experiment Scenarios List

Methods	Experiments's Code	Experiments
Multiple Linear Regression	A1	Unstandardized data
	A2	Standardized data
Ridge Regression	B1	Alpha = 10
	B2	Alpha = 100
	B3	Alpha = 1000
Polynomial Regression	C1	Degree = 2
	C2	Degree = 3
	C3	Degree = 4
	C4	Degree = 2 + Alpha 10
	C5	Degree = 3 + Alpha 10
	C6	Degree = 4 + Alpha 10
	C7	Degree = 2 + Alpha 100
	C8	Degree = 3 + Alpha 100
	C9	Degree = 4 + Alpha 100
	C10	Degree = 2 + Alpha 1000
	C11	Degree = 3 + Alpha 1000
	C12	Degree = 4 + Alpha 1000

3.5 Models Training

The results of model training using training data on datasets whose data have not been normally distributed, produce model evaluation values as shown in Table 2.

Table 2. Models Training (Undistributed Data)

Experiments's Code	MSE	RMSE	MAE
A1	45221378376	212653.19	138205.878
A2	45221378376	212653.188	138205.878
B1	45221413510	212653.2706	138189.3322
B2	45224777601	212661.1803	138049.1341
B3	45475633648	213250.1668	137261.6514
C1	31543173742	177603.9801	120891.3656
C2	24864805902	157685.7822	110830.2669
C3	16692871617	129200.8963	92691.79759
C4	31517520948	177531.7463	121074.9663
C5	25005901506	158132.5441	111206.1858
C6	17349120139	131716.0588	95245.57618
C7	31526006414	177555.6432	121072.6547
C8	25234056932	158852.3117	111606.6789
C9	18248875729	135088.3997	97630.09319
C10	31848251289	178460.7836	121741.5383
C11	26473200584	162705.8714	113981.5366
C12	20258345984	142331.8165	102763.241

Meanwhile, the results of model training on the dataset, where the data has been normally distributed, produce model evaluation values as shown in Table 3.

Table 3. Models Training (Distributed Data)

Experiments's Code	MSE	RMSE	MAE
A1	0.093121021	0.305157371	0.242022657
A2	0.093121021	0.305157371	0.242022657
B1	0.093121155	0.30515759	0.242033493
B2	0.093132082	0.305175493	0.242137691
B3	0.093796006	0.306261337	0.243699071
C1	0.082892775	0.287911053	0.228292673
C2	0.073759975	0.271587877	0.213854209
C3	0.058613176	0.242101581	0.18444352
C4	0.082972479	0.288049438	0.228453391
C5	0.074126899	0.272262556	0.214615217
C6	0.061157997	0.247301429	0.191382601
C7	0.083048706	0.288181725	0.228705421
C8	0.074685577	0.27328662	0.215904063
C9	0.063680491	0.252349937	0.196403909
C10	0.084178217	0.290134826	0.231307578
C11	0.07799413	0.279274292	0.222322933
C12	0.068970424	0.262622207	0.206394858

Based on the results of Model Training, using both data that has not been normally distributed and which has, it can be seen that the yellow line in the experimental code B3 (ridge regression with Alpha (λ) = 1000) indicates the method with the worst model performance produced, this can be seen from the resulting error matrix which has the largest compared to other experiments. The best method produced at the model training stage is obtained from the experimental code C3, which is using the polynomial regression method with a Degree of 4.

3.6 Models Testing

The model that has been generated from the training results will then be tested using test data. The results of model testing using data that have not been normally distributed along with the values of the model evaluation results can be seen in Table 4.

Table 4. Models Testing (Undistributed Data)

Experiments's Code	MSE	RMSE	MAE
A1	50128346637	223893.6056	142881.166
A2	50128346637	223893.6056	142881.166
B1	50130073997	223897.4631	142864.6217
B2	50148530320	223938.6754	142725.5953
B3	50539303161	224809.4819	141984.7299
C1	42628209602	206466.0011	129107.6228
C2	5.1624E+11	718498.305	161737.6071
C3	2.00E+26	1.41515E+13	9.4285E+11
C4	42638113301	206489.9835	129414.933

Table 4 (Continued). Models Testing (Undistributed Data)

Experiments's Code	MSE	RMSE	MAE
C5	3.09021E+11	555896.7696	142780.0118
C6	1.09652E+14	10471507.21	417841.4796
C7	42236056413	205514.127	129313.139
C8	2.45225E+11	495202.0021	138213.5891
C9	3.98497E+13	6312660.428	295357.9226
C10	40693850733	201727.1691	129350.6421
C11	1.10259E+11	332052.5994	133694.4872
C12	4.16626E+12	2041142.936	186615.3262

Meanwhile, the results of model testing on the dataset, where the data has been normally distributed, produce model evaluation values as shown in Table 5.

Table 5. Model Testing (Distributed Data)

Experiments's Code	MSE	RMSE	MAE
A1	0.095282421	0.308678508	0.245287972
A2	0.095282421	0.308678508	0.245287972
B1	0.095283759	0.308680675	0.245301769
B2	0.095296763	0.308701738	0.245420803
B3	0.095863157	0.309617759	0.246969769
C1	0.087550756	0.29588977	0.234322362
C2	1.98256E+15	44525982.05	2884203.528
C3	1.11174E+13	3334270.518	232528.2493
C4	0.08743979	0.295702198	0.234385827
C5	0.098193678	0.313358704	0.237551861
C6	0.241710969	0.491641098	0.284455752
C7	0.087416	0.295661969	0.23446542
C8	0.090493296	0.300821037	0.233195848
C9	0.179454746	0.423620994	0.2601806
C10	0.088231986	0.297038694	0.236620547
C11	0.090691695	0.301150618	0.234831412
C12	0.142003439	0.376833437	0.248961567

Based on the results of model testing, it can be seen that on data that has not been normally distributed, the yellow line in experimental code C3 (Polynomial Regression with Degree = 4) produces a model with the worst performance when viewed from the resulting error matrix, therefore it can be said that the model is overfitted. Whereas for data that has been normally distributed, the worst model is produced from experiment C2, which uses the polynomial regression method with Degree = 3. The best model generated from data that has not been normally distributed is obtained from the experimental code C10 (Regression polynomial with Degree = 2 + Alpha (λ) = 1000). Whereas for data that has been normally distributed, the best model produced is obtained from the experimental code C7, which uses the polynomial regression method with a value of Degree = 2 and coupled with a regularization technique using the ridge regression method with an Alpha value (λ) = 100.

3.7 Models Comparison and Implementation

The evaluation value of the model generated from data that has not been normally distributed is far from the range of 0, this proves that the resulting model is invalid because it cannot fulfill one of the regression assumptions. On the other hand, the model generated from data that has been normally distributed produces a model evaluation value close to 0, this means that the resulting model is valid. Therefore, it can be concluded that the models that produce the most optimal performance will be taken from the models obtained in datasets that have been normally distributed. In models generated from experiments using normally distributed data, the multiple linear regression model produces MSE, RMSE, and MAE values of 0.0952, 0.3086, and 0.2452. In the ridge regression model, the best model evaluation value is derived from Alpha (λ) = 10, the model produces MSE, RMSE, and MAE values of 0.0952, 0.3086, 0.2453, whereas in the polynomial regression model, the best model evaluation value is derived from the Degree value = 2 + Alpha (λ) = 100 with MSE, RMSE, and MAE worth 0.0874, 0.2955, 0.2344. Figure 5 shows a graph of the error matrix values for each of the best models resulting from the three types of regression methods.

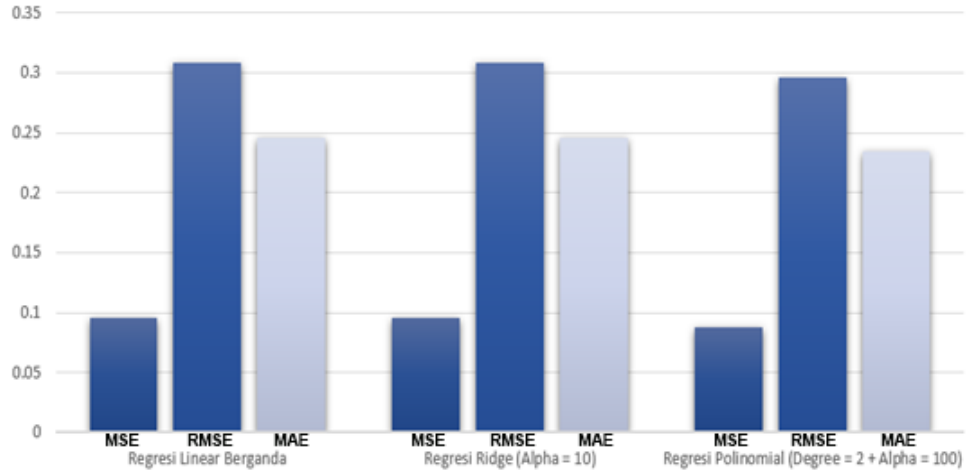

Figure 5. Error Matrix Values Comparison

The three models that have been generated are then stored on localhost using a tool from Python called Pickle. Each of these models will be used separately and implemented in making a house price prediction application, this is done to compare the prediction results obtained from each of the regression models. The application is website-based and created using the help of a Python programming language framework called Streamlit. After experimenting to predict house prices using this research's dataset, the original price of a house corresponds to what is stated in the dataset, which is $221,900. The results of the prediction of house prices generated by the multiple linear regression model, produce a house price value of $ 327,846. Whereas in the prediction results generated by the ridge regression model with an Alpha value = 10, the house price is estimated to be worth $ 326,290. The prediction generated by the polynomial regression model with a value of degree = 2 and Alpha = 100, produces a predicted price output that is worth $ 315,238.

Based on the experimental results, it can be said that the predictions made using a polynomial regression model with a value of degree = 2 and Alpha = 100, are proven to produce a predicted value that is closest to the original price of the house according to what is stated in the dataset. The difference between the original price of the house and the predicted results made using the model, is in the range of $ 93,000 where the value of the resulting predicted house prices is worth more than the original prices of houses in the dataset.

4. CONCLUSIONS

Based on the results of analysis and research experiments, the points that can be concluded are as follows:

1. The prediction of house prices using the dataset of house sales in King County, USA, can be done using the polynomial regression method with a value of degree = 2, and coupled with the application of regularization techniques using the ridge regression method with an Alpha value (λ) = 100. The resulting model performance is the best compared to other models in this study. This can be seen from the resulting MSE, RMSE, and MAE values of 0.087416, 0.295661969, and 0.23446542 respectively. The values of the error matrix are the smallest compared to other models resulted from this research.

2. Model training using the polynomial regression method with a value of degree = 4, both for datasets that have not been normally distributed and data that has, it is evident that the model produced using this method has an overfitting behavior. This was proven when testing the model, the error matrix values generated at the model evaluation stage were the largest compared to other models resulted from this research.

3. Based on the dataset used in this study, the features that most influence the increase in house prices (price) are building area (sqft_living), houses near water areas (waterfront), house design appearance (view), house rating values (grade), and finally, the year the house was built (yr_built). A slight increase in the value of these five features will also be followed by a significant increase in house prices.

REFERENCES

Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. In Renewable and Sustainable Energy Reviews (Vol. 74, pp. 902–924).

Febyanti, F. (2022). Pemodelan Faktor-Faktor yang Mempengaruhi Harga Rumah di Jabodetabek Menggunakan Metode Regresi Probit.

Hoerl, R. W. (2020). Ridge Regression: A Historical Context. Technometrics, 62(4), 420–425.

Mahesh, B. (2018). Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Six Stroke Engine View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review.

Mcleod, S. (2018). *Maslow's Hierarchy of Needs*.

Naser, M. Z., & Alavi, A. H. (2020). *Insights into Performance Fitness and Error Metrics for Machine Learning*.

Ranstam, J., & Cook, J. A. (2018). LASSO regression. British Journal of Surgery, 105(10), 1348.

Ray, S. (2019). A Quick Review of Machine Learning Algorithms

Reza Fahlepi, M., Widjaja, A., & Surya Sumantri No, J. (2019). Penerapan Metode Multiple Linear Regression Untuk Prediksi Harga Sewa Kamar Kost.

Roihan, A., Abas Sunarya, P., & Rafika, A. S. (2020). IJCIT (Indonesian Journal on Computer and Information Technology) Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. In IJCIT (Indonesian Journal on Computer and Information Technology) (Vol. 5, Issue 1).

Sarstedt, M., & Mooi, E. (2019). *Regression Analysis* (pp. 209–256).

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *American Statistician*, *72* (1), 37–45.

Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, *12* (2), 15–20.

Tranmer, M., Murphy, J., Elliot, M., & Pampaka, M. (2020). *Multiple Linear Regression (2nd Edition)*.

Wahyuni, E. D., Arifiyanti, A. A., & Kustyani, M. (2019). Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining. 263–269.

Zhou, C. (2021). House price prediction using polynomial regression with Particle Swarm Optimization. *IOP Conference Series: Earth and Environmental Science*, *1802* (3).