Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

# House Price Prediction Using Exploratory Data Analysis and Machine Learning with Feature Selection

Fadhil M. Basysyar*, Gifthera Dwilestari

Information System, STMIK IKMI, 45131 Cirebon, Indonesia

* Correspondence: Fadhil M. Basysyar (fadhil.m.basysyar@gmail.com)

**Received:** 07-15-2022 **Revised:** 08-17-2022 **Accepted:** 09-13-2022

**Citation:** F. M. Basysyar and G. Dwilestari, "House price prediction using exploratory data analysis and machine learning with feature selection," *Acadlore Trans. Mach. Learn.*, vol. 1, no. 1, pp. 11-21, 2022. https://doi.org/10.56578/ataiml010103.

**Abstract:** In many real-world applications, it is more realistic to predict a price range than to forecast a single value. When the goal is to identify a range of prices, price prediction becomes a classification problem. The House Price Index is a typical instrument for estimating house price discrepancies. This repeat sale index analyzes the mean price variation in repeat sales or refinancing of the same assets. Since it depends on all transactions, the House Price Index is poor at projecting the price of a single house. To forecast house prices effectively, this study investigates the exploratory data analysis based on linear regression, ridge regression, Lasso regression, and Elastic Net regression, with the aid of machine learning with feature selection. The proposed prediction model for house prices was evaluated on a machine learning housing dataset, which covers 1,460 records and 81 features. By comparing the predicted and actual prices, it was learned that our model outputted an acceptable, expected values compared to the actual values. The error margin to actual values was very small. The comparison shows that our model is satisfactory in predicting house prices.

**Keywords:** House price index; Feature selection; Machine learning; Exploratory data analysis

## 1. Introduction

Real estate development is an important measure for a country to stimulate economic growth in the short term. As the economy improves, people tend to move from cities to rural areas, resulting in a boom of population. Housing demand rises in tandem with population growth. The growth of house prices is in lockstep with the market. In a specific region, the price of homes may spike suddenly with infrastructural development. For example, homeowners in a residential neighborhood prefer to increase the selling price of their houses, after issues like an impassable road and unstable electricity were resolved. The price increase of residential dwellings is frequently calculated by the House Price Index [1-4].

Despite its importance, the House Price Index has not been sufficiently explored by researchers in this century [5-7]. The overall home value is influenced by a lot of factors, including but not limited to physical states, concepts, and locations. Physical perception can detect the size of the property, the number and space of rooms, the availability of the yard, the area of land and structures, and the age of the property [8]. The price of a house is also affected by other physical attributes, such as its size, year of construction, number of bedrooms and bathrooms, and other interior amenities [9]. Concepts allude to the numerous marketing methods used by developers to persuade potential investors. The common concepts include the proximity of the property to hospitals, markets, educational institutions, airports, major roads, etc. The location of a property has a significant impact on its pricing, because the current land price depends largely by the surroundings.

For various stakeholders (e.g., tenants, homeowners, real estate specialists, lawmakers, and urban/regional planning agencies), it is critical to understand the patterns and determinants of house pricing [10]. A computer-based prediction system can assist people in determining whether and when to purchase a home [11-15]. Residential real estate, the major reservoir of middle-class equity, acts as a source of capital for new businesses. The rising property prices may enhance demand by raising the income of homeowners, but may also encourage debt-financed spending and erode financial resilience.

https://doi.org/10.56578/ataiml010103

There are in general two types of price forecasting strategies: the time series strategy to predict market patterns, and the strategy to determine the price of a commodity based on its features. The former strategy aims to clarify the relationship between current and historical rates, and the latter utilizes pricing and linear regression [16-18]. Following the second type of strategy, this paper carries out an exploratory data analysis based on linear regression, ridge regression, Lasso regression, and Elastic Net regression with feature selection.

## 2. Methodology

To estimate house prices based on the features of a relevant dataset, this study performs an exploratory data analysis based on linear regression, ridge regression, Lasso regression, and Elastic Net regression with feature selection [19-21]. The relevant data were collected and explored to analyze the dataset and identify the key sections in the dataset. Then, the data were preprocessed to make them suitable for model creation. These are the main processes of our methodology.

### 2.1 Data Reading and Exploration

The house data dataset from the Machine Learning Repository at kaggle.com were used to create our model. The selected dataset contains 1,460 records, which provide the aggregated data on 81 features for homes in various suburbs. Before developing a regression model, it is essential to carry out exploratory data analysis. This allows researchers to uncover underlying trends in the data, and assists with the selection of appropriate machine learning algorithms. As a result, data exploration was performed to comprehend the features present in the dataset as well as their functions.

The selected dataset contains the per-capita crime rate per town, the percentage of residential land allocated for lots, the ratio of non-retail commercial acres by town, and the Charles River dummy variable (1 if the tract bounds the river; 0 if otherwise), Nitric oxide concentration (parts per 10 million), the typical number of rooms in a house, the percentage of owner-occupied apartments built before 1980 (represented by age), the weighed distances to employment centers, the index of radial highway accessibility, the tax rate (the total value of the property), the pupil-teacher ratio (town-specific), the proportion of dark-skinned students (town-specific), the median of owner-occupied homes, and the percentage of people with a low socioeconomic status.

Since our model adopts supervised learning, the dataset was divided properly into a training set and a test set [22-24].

### 2.2 Initial Data Preparation

The data for model training and testing must be thoroughly scrutinized before modeling. Otherwise, the constructed model would be unable to learn the patterns very quickly. As shown in Figure 1, there was no permissible missing value in the dataset.

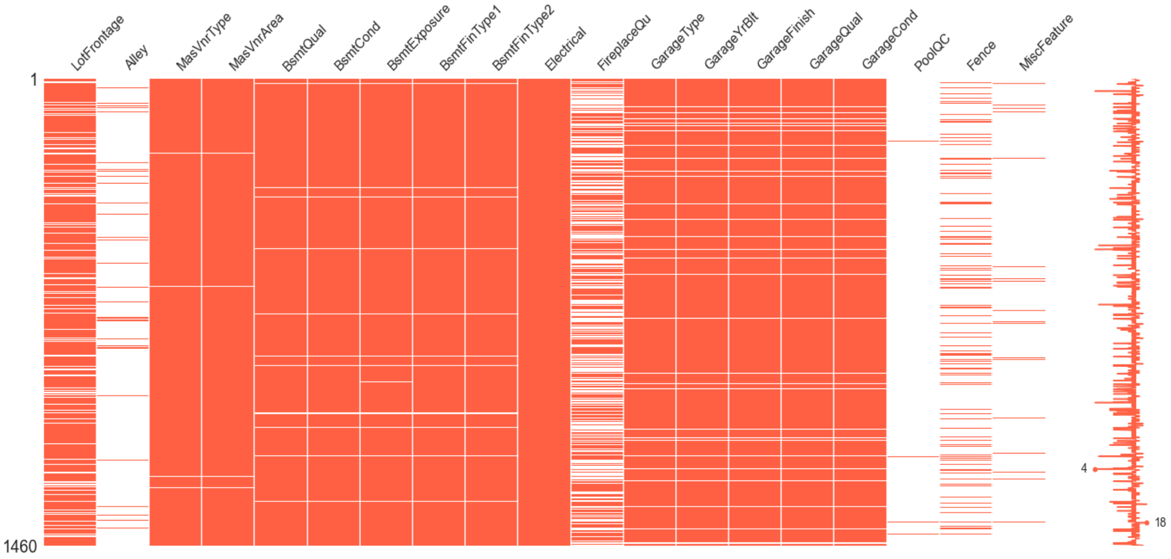


**Figure 1.** No missing value in the dataset

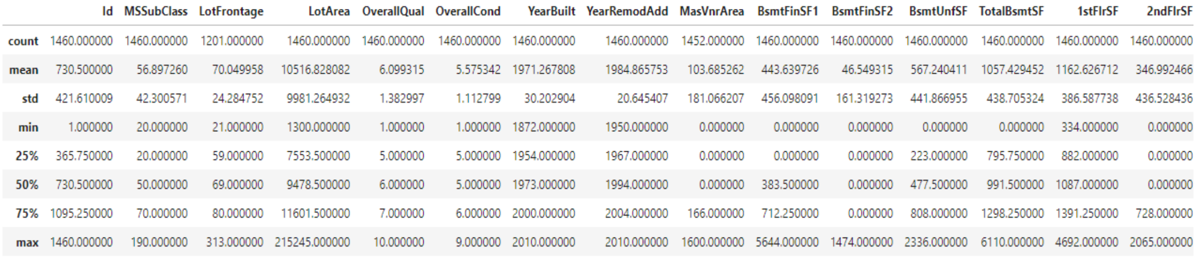
Next, the numerical values were normalized, and the classes were encoded one at a time. Following data exploration, the most suitable features were selected against the heatmap, and the feature data were processed preliminary. The training and test sets typically have various properties.

Scaling was performed to ensure that the components are of a reasonably similar size, as the values of individual features are likely on different scales. The scale difference may weaken the performance of our model. Here, the scaling was realized using the standard scaler function of the Phyton sklearn module.

The Standard Scaler assumes that the data are naturally distributed inside each process, and scales the data to cluster around 0 with a standard deviation of 1. The mean and standard deviation of a feature are determined, and the component is installed [25, 26]:

$$\frac{xi - means(x)}{stdev(x)} \tag{1}$$

Figure 2 describes the numerical columns and some specific percentiles statistically. Figure 3 visualizes the numerical variables as a box plot.



| | Id | MSSubClass | LotFrontage | LotArea | OverallQual | OverallCond | YearBuilt | YearRemodAdd | MasVnrArea | BsmtFinSF1 | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | 1stFlrSF | 2ndFlrSF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 1460.000000 | 1460.000000 | 1201.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1452.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 | 1460.000000 |
| **mean** | 730.500000 | 56.897260 | 70.049958 | 10516.828082 | 6.099315 | 5.575342 | 1971.267808 | 1984.865753 | 103.685262 | 443.639726 | 46.549315 | 567.240411 | 1057.429452 | 1162.626712 | 346.992466 |
| **std** | 421.610009 | 42.300571 | 24.284752 | 9981.264932 | 1.382997 | 1.112799 | 30.202904 | 20.645407 | 181.066207 | 456.098091 | 161.319273 | 441.866955 | 438.705324 | 386.587738 | 436.528436 |
| **min** | 1.000000 | 20.000000 | 21.000000 | 1300.000000 | 1.000000 | 1.000000 | 1872.000000 | 1950.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 334.000000 | 0.000000 |
| **25%** | 365.750000 | 20.000000 | 59.000000 | 7553.500000 | 5.000000 | 5.000000 | 1954.000000 | 1967.000000 | 0.000000 | 0.000000 | 0.000000 | 223.000000 | 795.750000 | 882.000000 | 0.000000 |
| **50%** | 730.500000 | 50.000000 | 69.000000 | 9478.500000 | 6.000000 | 5.000000 | 1973.000000 | 1994.000000 | 0.000000 | 383.500000 | 0.000000 | 477.500000 | 991.500000 | 1087.000000 | 0.000000 |
| **75%** | 1095.250000 | 70.000000 | 80.000000 | 11601.500000 | 7.000000 | 6.000000 | 2000.000000 | 2004.000000 | 166.000000 | 712.250000 | 0.000000 | 808.000000 | 1298.250000 | 1391.250000 | 728.000000 |
| **max** | 1460.000000 | 190.000000 | 313.000000 | 215245.000000 | 10.000000 | 9.000000 | 2010.000000 | 2010.000000 | 1600.000000 | 5644.000000 | 1474.000000 | 2336.000000 | 6110.000000 | 4692.000000 | 2065.000000 |

**Figure 2.** Statistical description of the dataset

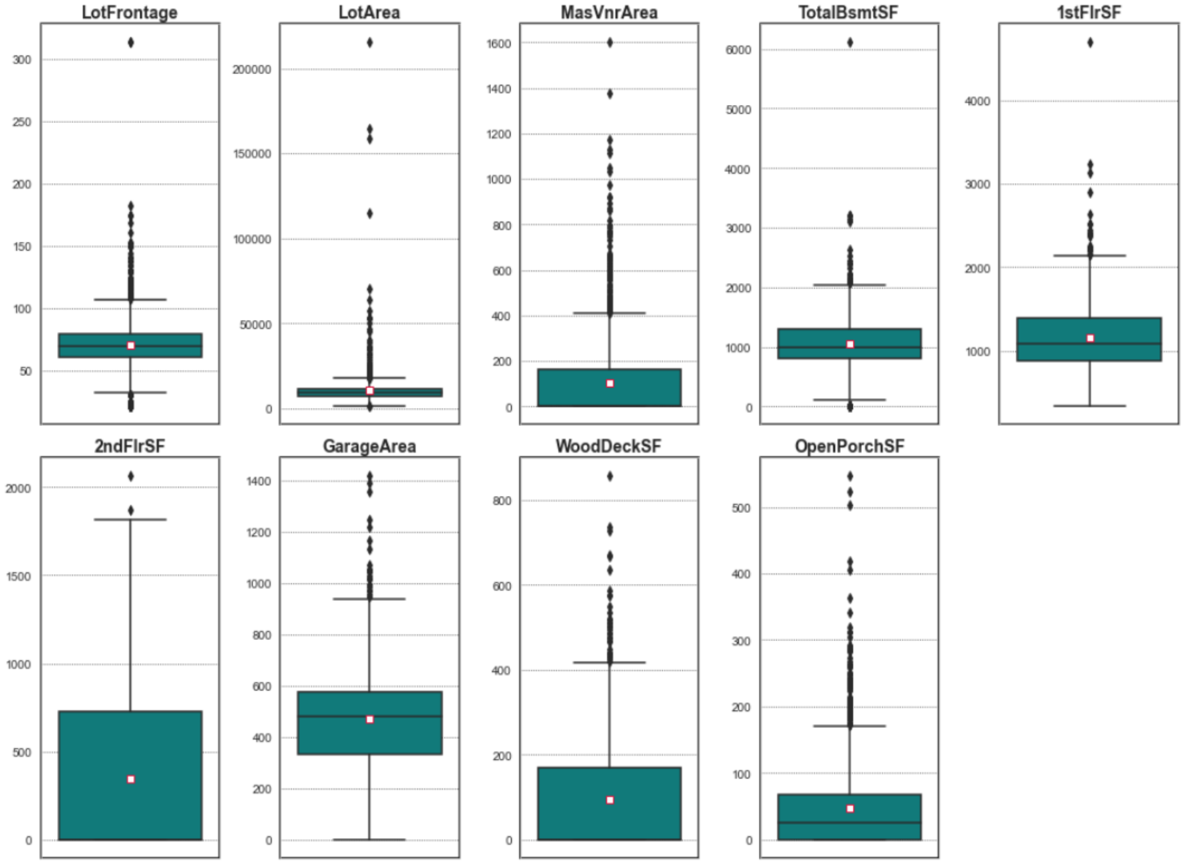


**Figure 3.** Box plot of numerical variables

## 3. Exploratory Data Analysis

The Results section may be divided into subsections. It should describe the results concisely and precisely, provide their interpretation, and draw possible conclusions from the results.

### 3.1 Visualizing Target

The target variable, SalePrice, has a slightly positive correlation with the target. Some of the features exert a substantial impact on the sale prices of their respective classes, while some others do not. The latter are regarded as irrelevant and be discarded, during the feature selection process (Figure 4).

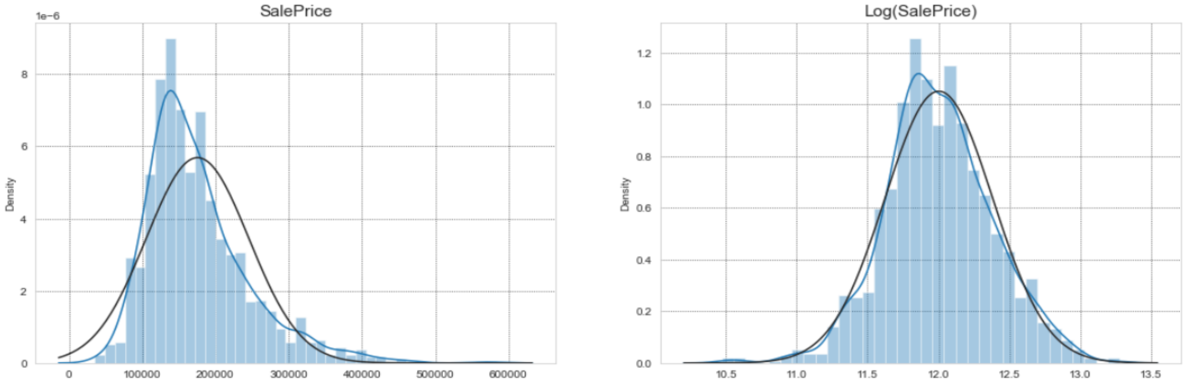


**Figure 4.** The target variable

There are also some binary and ordinal class variables. Many variables with SalePrice exhibit a strong rising trend. In some cases, the rise is nearly linear. Newer homes are typically pricier. Houses with newly built garages are more expensive than houses with older garages, which in turn are more expensive than houses without a garage. The age of the house from the year it was built or remodeled to the year it was sold appears to have a negative relationship with the price at which the property was sold. As illustrated in Figure 5, it is not necessary to add this as a new feature, because it will be handled by linear regression.

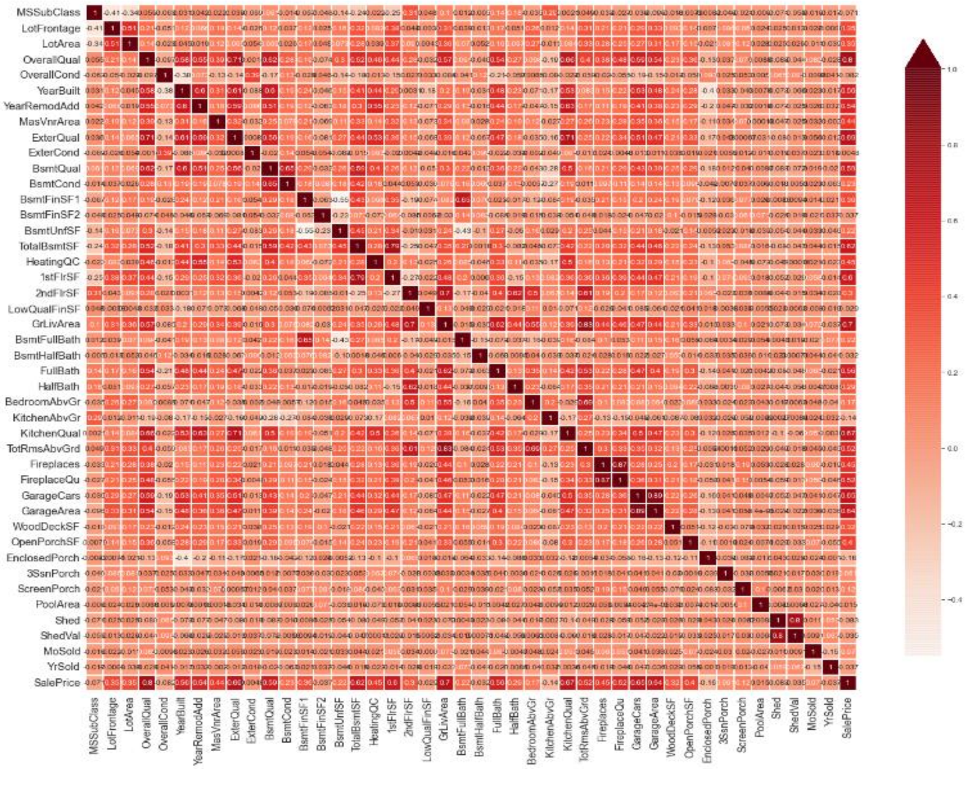
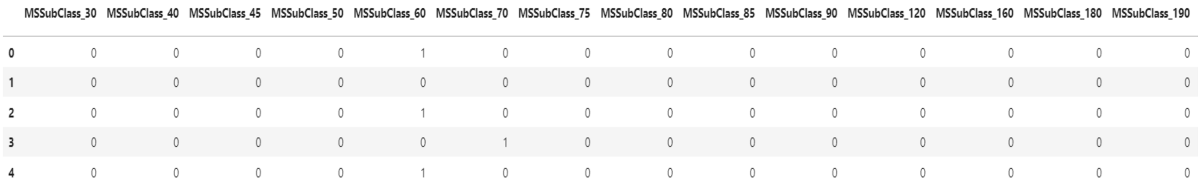


**Figure 5.** Correlations between variables

### 3.2 Data Preparation

During data preparation, the initial stage is to establish dummy variables for the remaining class variables, and the next stage is to perform a level drop. The columns of binary flat features like “Street” and “Utilities” are useless in modeling, because these features almost all belong to the same class. The two features could be removed later in the feature selection process. Finally, the original class variables are removed, and dummy variables are created for the remaining class variables (Figure 6). The latter variables can be concatenated with the leading df.

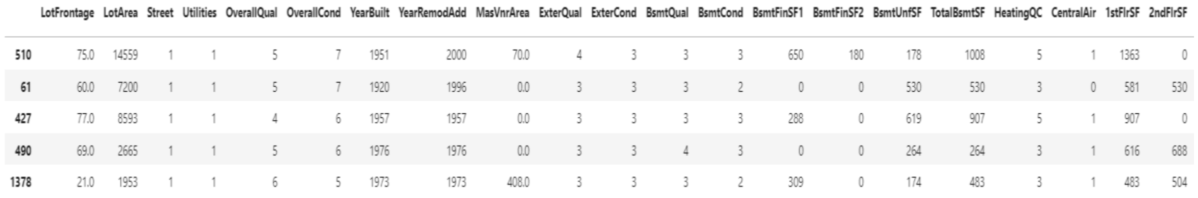


| | MSSubClass_30 | MSSubClass_40 | MSSubClass_45 | MSSubClass_50 | MSSubClass_60 | MSSubClass_70 | MSSubClass_75 | MSSubClass_80 | MSSubClass_85 | MSSubClass_90 | MSSubClass_120 | MSSubClass_160 | MSSubClass_180 | MSSubClass_190 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 6.** Fata preparation

### 3.3 Data Split

In machine learning, the concept of model training is merely an approximation of a set of parameters that can describe a particular dataset. The training procedure for the algorithm y = w + bx will give two estimates for the variables w and b. Using fewer training data will increase the estimation variance of each parameter. If fewer data tests were conducted, the estimation of the model outcomes will be more variable. Hence, the dataset was partitioned into two parts by the ratio of 70:30, namely, a training set and a test set. Figure 7 provides an example of the training data.



| | LotFrontage | LotArea | Street | Utilities | OverallQual | OverallCond | YearBuilt | YearRemodAdd | MasVnrArea | ExterQual | ExterCond | BsmtQual | BsmtCond | BsmtFinSF1 | BsmtFinSF2 | BsmtUnfSF | TotalBsmtSF | HeatingQC | CentralAir | 1stFlrSF | 2ndFlrSF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **510** | 75.0 | 14559 | 1 | 1 | 5 | 7 | 1951 | 2000 | 70.0 | 4 | 3 | 3 | 3 | 650 | 180 | 178 | 1008 | 5 | 1 | 1363 | 0 |
| **61** | 60.0 | 7200 | 1 | 1 | 5 | 7 | 1920 | 1996 | 0.0 | 3 | 3 | 3 | 2 | 0 | 0 | 530 | 530 | 3 | 0 | 581 | 530 |
| **427** | 77.0 | 8593 | 1 | 1 | 4 | 6 | 1957 | 1957 | 0.0 | 3 | 3 | 3 | 3 | 288 | 0 | 619 | 907 | 5 | 1 | 907 | 0 |
| **490** | 69.0 | 2665 | 1 | 1 | 5 | 6 | 1976 | 1976 | 0.0 | 3 | 3 | 4 | 3 | 0 | 0 | 264 | 264 | 3 | 1 | 616 | 688 |
| **1378** | 21.0 | 1953 | 1 | 1 | 6 | 5 | 1973 | 1973 | 408.0 | 3 | 3 | 3 | 2 | 309 | 0 | 174 | 483 | 3 | 1 | 483 | 504 |

**Figure 7.** Example of training data

### 3.4 Model Construction

The model is constructed in the following steps: Selecting features from the target; generating a function for developing a linear regression model with a Train R2 of 0.94 and a Test R2 of 0.86, a sign of overfitting. This problem could be remedied through careful selection and regularization of features. There appears to be some nonlinearity in the error terms. It goes against the assumption of linear regression that the error terms are independent of each other. As illustrated in Figure 8, the problem could be corrected by changing the regressor variables or the target.

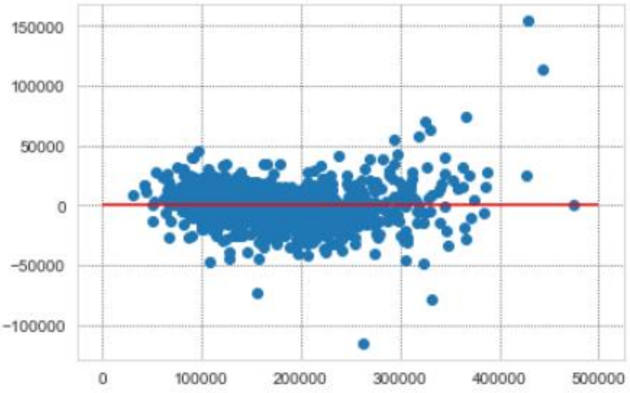


**Figure 8.** Nonlinearity error terms

In addition, the translation of the regression equation yields log(y) = X + a. Using a statistical model with an error of 0.5, the value was obtained as 0.95 for Train R2 and 0.86 for Test R2. As shown in Figure 9, the trend of error terms varied significantly between the two scenarios. In the event of a converted answer, the error words appear to be scattered randomly.

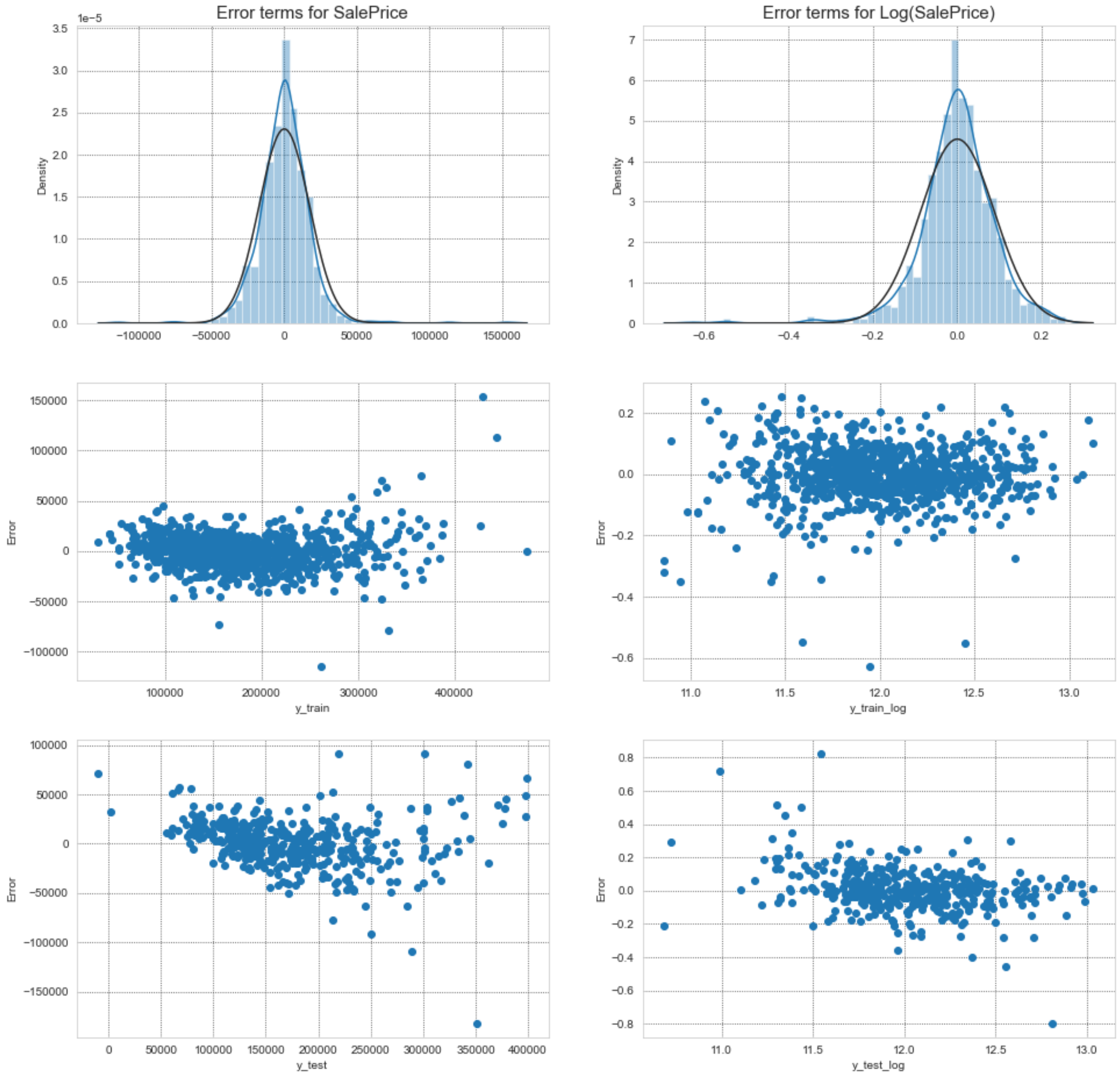


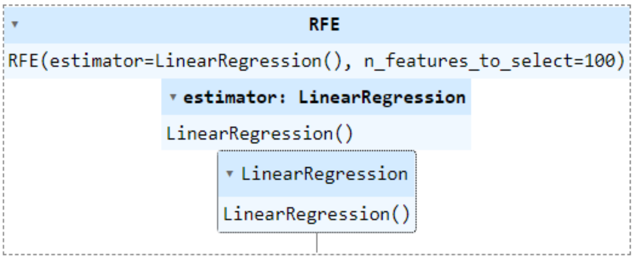
**Figure 9.** Comparison of error terms

## 4. Results and Discussion

Data exploration was performed before to better understand the dataset. The columns from the selected data provide a fascinating summary, through recursive feature elimination (RFE), linear regression, ridge regression, Lasso regression, and Elastic Net regression. This summary makes a lot of sense, for both variables are conditional and unconditional. These columns are assumed to be useless for regression tasks, such as trend prediction.

### 4.1 RFE Feature Elimination

As illustrated in Figure 10, the useless features were eliminated using RFE with selected elements. After the feature elimination, both the training set and the test set were updated.

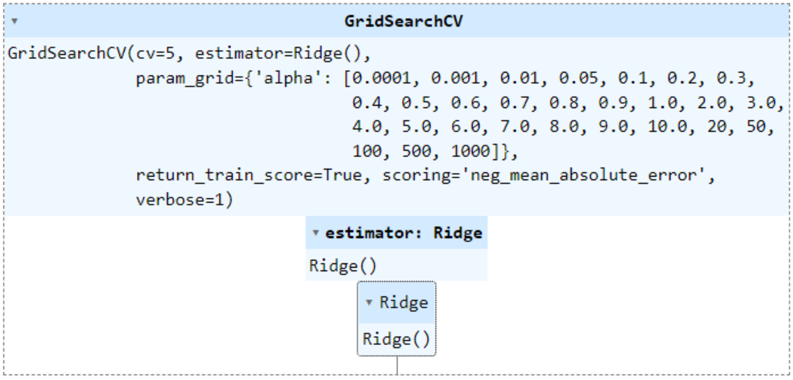


```
RFE(estimator=LinearRegression(), n_features_to_select=100)
```

**Figure 10.** RFE feature elimination

### 4.2 Ridge Regression

Following the RFE feature elimination, ridge regression was carried out to obtain a list of alphas. Five folds were fitted for each of the 28 candidates, for a total of 140 fits (Figure 11). Figure 12 depicts the mean training and test scores for various parameter settings considering negative mean absolute error and alpha.



```
GridSearchCV(cv=5, estimator=Ridge(),
             param_grid={'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3,
                                   0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
                                   4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50,
                                   100, 500, 1000]},
             return_train_score=True, scoring='neg_mean_absolute_error',
             verbose=1)
```

**Figure 11.** Ridge regression

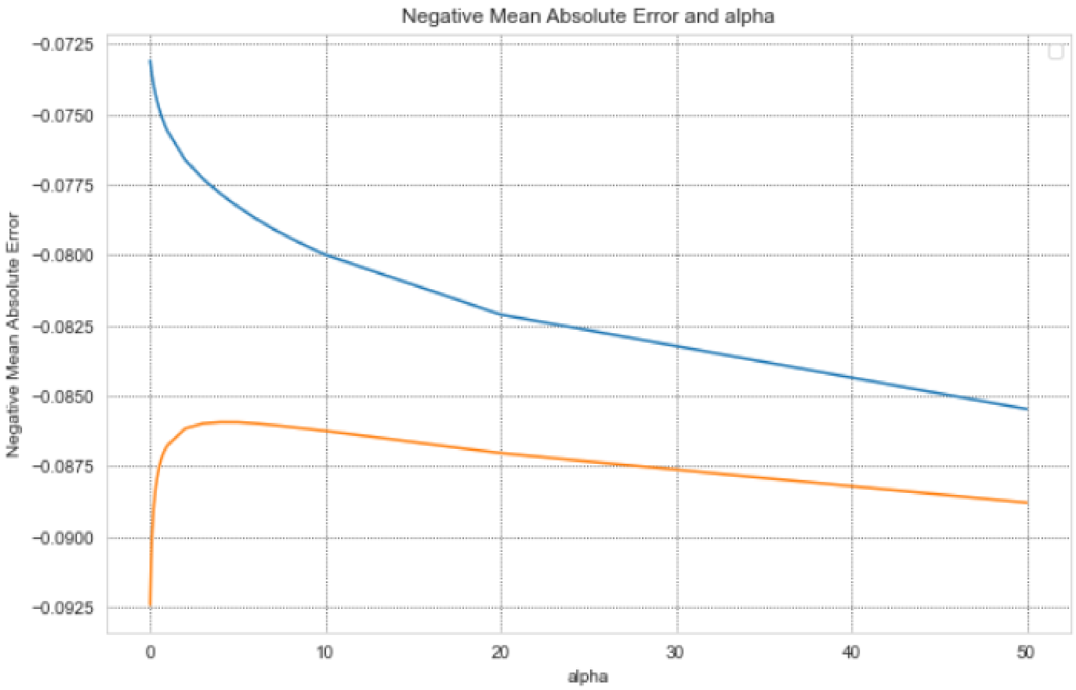
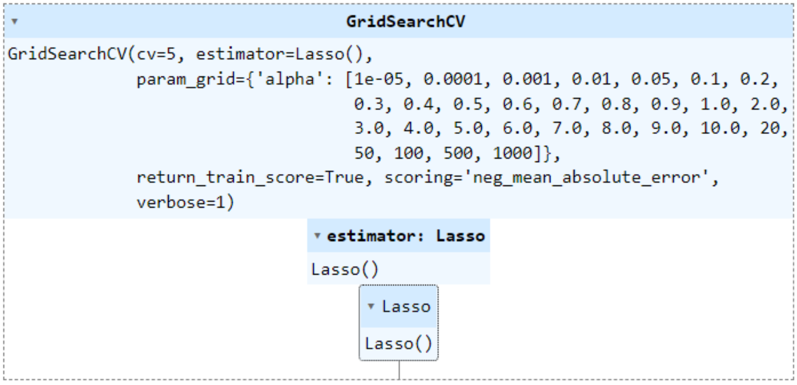


**Figure 12.** Negative mean absolute error and alpha

### 4.3 Lasso Regression

There are still a lot more variables to consider. The goal of this study is to develop a prediction model that can anticipate the price of a house in a certain setting based on a set of parameters. The model will also be used to assess the intensity of correlations between the response and the predictors, which is an important goal of house price prediction models. As a result, an intelligent method must be created for feature reduction, without sacrificing model performance.

Lasso generally performs well if only a few of the predictors are used to build a model, and if these predictors have a significant influence on the response variable. Thus, Lasso regression can work as a feature selection method to eliminate unimportant variables. As shown in Figure 13, Lasso regression consists of 5-fold fittings for each of 29 candidates, accounting for a total of 145 fits.



```
GridSearchCV
GridSearchCV(cv=5, estimator=Lasso(),
             param_grid={'alpha': [1e-05, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2,
                                   0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0,
                                   3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20,
                                   50, 100, 500, 1000]},
             return_train_score=True, scoring='neg_mean_absolute_error',
             verbose=1)
estimator: Lasso
Lasso()
Lasso
Lasso()
```

**Figure 13.** Lasso **r**egression

Figure 14 depicts the mean train and test scores for various parameters obtain through Lasso regression, as well as scores for various parameters in terms of negative mean absolute error and alpha.

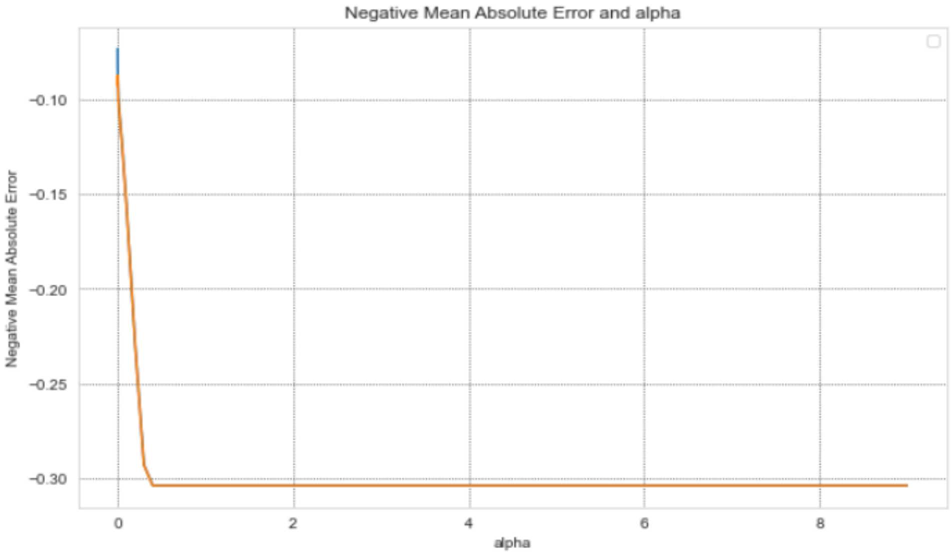
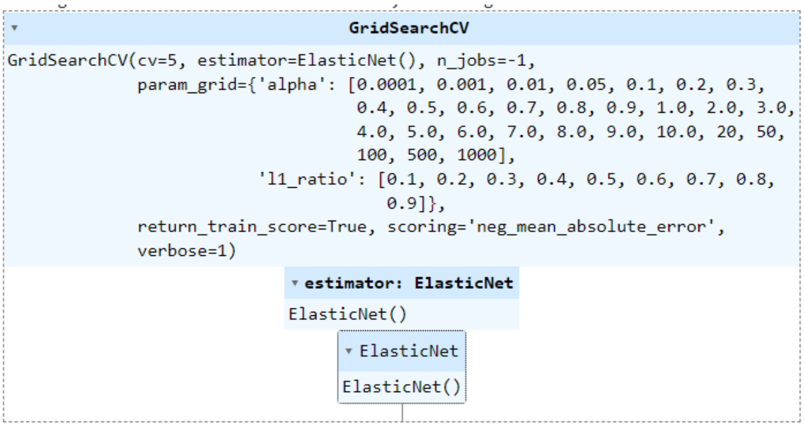


**Figure 14.** Negative mean absolute error and alpha

Lasso outperforms ridge regression when it comes to the prediction on unseen data. The regression losses of the two methods are approximately identical. Instead of using the better alpha, the salient features can be selected by slightly increasing the alpha. The above results show that Lasso functions as a feature selector capable of reducing unnecessary variables.

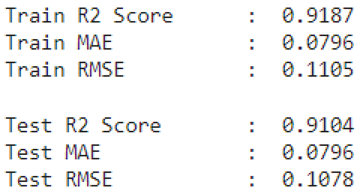
### 4.4 Elastic Net Regression

Elastic Net, as a regularized regression method, integrates the L1 and L2 penalties of Lasso and ridge regressions linearly. Figure 15 shows the Elastic Net fitting of five folds for each of the 252 candidates, accounting for a total of 1,260 fits. As shown in Figure 16, the Train R2 was 0.9187, the Train Mean Absolute Error was 0.0796, and the Train Mean Squared Error was 0.1105, whereas the Test R2 was 0.9104, the Test Mean Absolute Error was 0.0796, and the Test Mean Squared Error was 0.1078.



```
GridSearchCV
GridSearchCV(cv=5, estimator=ElasticNet(), n_jobs=-1,
             param_grid={'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3,
                                   0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
                                   4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50,
                                   100, 500, 1000],
                         'l1_ratio': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8,
                                      0.9]},
             return_train_score=True, scoring='neg_mean_absolute_error',
             verbose=1)
estimator: ElasticNet
ElasticNet()
ElasticNet
ElasticNet()
```

**Figure 15.** Elastic net regression



```
Train R2 Score     :  0.9187
Train MAE          :  0.0796
Train RMSE         :  0.1105

Test R2 Score      :  0.9104
Test MAE           :  0.0796
Test RMSE          :  0.1078
```

**Figure 16.** Elastic net models for best parameters

Figure 17 shows the prediction effect of our final model. It can be seen that the cost of a place is determined by the above grade living area, overall quality of home (i.e., quality of material finish), age of property, overall condition of the house, basement size, and zoning classification of the sale [27-29].

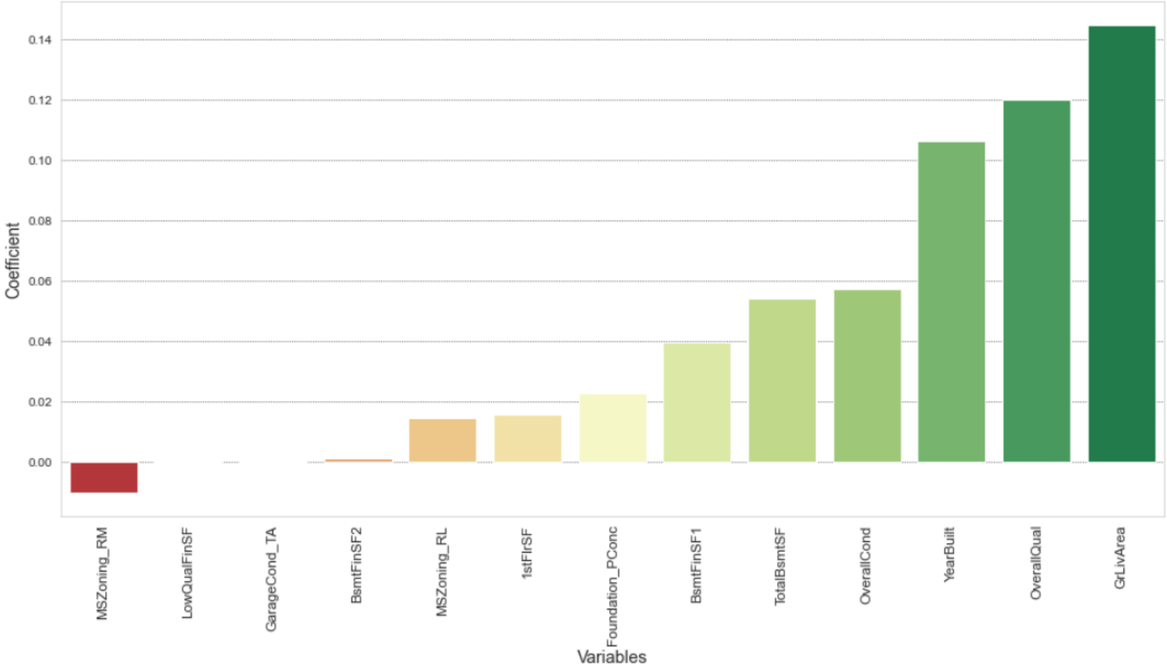


**Figure 17.** Prediction effect of our final model

## 5. Conclusions

Through the above experimental analysis, it is possible to measure and compare the performance of various models: For ridge regression, the Train R2 is 0.9195, Train Mean Absolute Error is 0.08, and Train Mean Squared Error is 0.11; the Test R2 is 0.9059, Test Mean Absolute Error is 0.08, and Test Mean Squared Error is 0.11. For Lasso regression, the Train R2 is 0.9122, the Mean Absolute Error is 0.0821, the Mean Squared Error is 0.1148; the Test R2 is 0.9113, the Mean Absolute Error is 0.0786, and the Mean Squared Error is 0.1073. For Elastic Net, the Train R2 is 0.9177, Train Mean Absolute Error is 0.0803, Train Mean Squared Error is 0.1112; the Test R2 is 0.9084, the Mean Squared Error is 0.0802, and the Mean Squared Error is 0.109. To sum up, Lasso outperformed the other models in terms of the performance on the test set. Furthermore, Lasso was used for intelligent feature selection with a modified alpha, which managed to reduce the variable set to the 13 most significant variables. According to our final model, the variables that affect the price of a house exhibit a nonlinear relationship between the regressors and the answer.

## Author Contributions

Fadhil Muhammad Basysyar conceived of the presented idea, developed the theory and performed the computations, verified the analytical methods. Gifthera Dwilestari encouraged Fadhil Muhammad Basysyar to investigate real-world applications of House Price Prediction and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

Fadhil Muhammad Basysyar and Gifthera Dwilestari carried out the experiment, wrote the manuscript with support and developed the theoretical formalism, performed the Data Analysis and Machine Learning with Feature Selection to the final version of the manuscript. Fadhil Muhammad Basysyar and Gifthera Dwilestari contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

## Data Availability

The data that support the findings of this study are available in kaggle at www.kaggle.com. These data were derived from the following resources available in the public domain: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.

## Acknowledgements

We would like to thank STMIK IKMI Cirebon, and International Information and Engineering Technology Association for the support and help with statistical analysis. We sincerely thank all the participants in our study.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

[1] Y. Aliefendioğlu, H. Tanrivermis, and M. A. Salami, "House price index (HPI) and Covid-19 pandemic shocks: Evidence from Turkey and Kazakhstan," *Int. J. Hous. Mark. Anal.*, vol. 15, no. 1, pp. 108-125, 2022. http://dx.doi.org/10.1108/IJHMA-10-2020-0126.

[2] A. Grimes, K. Sorensen, and C. Young, "Repeat sales house price indices: Comparative properties under alternative data generation processes," *New Zealand Econ Pap.*, vol. 55, no. 1, pp. 7-18, 2021. http://dx.doi.org/10.1080/00779954.2019.1612937.

[3] Z. Li, "Prediction of house price index based on machine learning methods," In 2nd International Conference on Computing and Data Science, (CDS 2021), Stanford, CA, USA, July 5, 2021, IEEE, 2021. pp. 472-476. https://doi.org/10.1109/CDS52072.2021.00087.

[4] "House Price Index," 2021, Distaque: Instituto Nacional de Estatística.

[5] D. Sayag, D. Ben-hur, and D. Pfeffermann, "Reducing revisions in hedonic house price indices by the use of nowcasts," *Int. J. Forecasting,* vol. 38, no. 1, pp. 253-266, 2022. http://dx.doi.org/10.1016/j.ijforecast.2021.04.008.

[6] Q. Ma, Z. Khan, F. Chen, M. Murshed, Y. Siqun, and D. Kirikkaleli, "Revisiting the nexus between house pricing and money demand: Power spectrum and wavelet coherence based approach," *Q. Rev. Econ. Financ.*, vol. 2021, 2021. http://dx.doi.org/10.1016/j.qref.2021.03.001.

[7] J. V. Duca, M. Hoesli, and J. Montezuma, "The resilience and realignment of house prices in the era of Covid-19," *J. Eur. Real. Estate Re.*, vol. 14, no. 3, pp. 421-431, 2021. http://dx.doi.org/10.1108/JERER-11-2020-0055.

[8] O. Bover and P. Velilla, “Hedonic house prices without characteristics: The case of new multiunit housing,” *SSRN Electron. J.*, vol. 2003, 2003. http://dx.doi.org/10.2139/ssrn.357280.

[9] M. L. Zagalaz-Sánchez, J. Cachón-Zagalaz, V. Arufe-Giráldez, A. Sanmiguel-Rodríguez, and G. González-Valero, “Influence of the characteristics of the house and place of residence in the daily educational activities of children during the period of COVID-19’ confinement,” *Heliyon*, vol. 7, no. 3, Article ID: e06392, 2021. http://dx.doi.org/10.1016/j.heliyon.2021.e06392.

[10] M. Khamdevi, “The characteristics linkage among Austronesian houses,” *AMERTA*, vol. 39, no. 2, pp. 147-162, 2021. http://dx.doi.org/10.24832/amt.v39i2.147-162.

[11] J. P. Gupta, A. Singh, and R. K. Kumar, “A computer-based disease prediction and medicine recommendation system using machine learning approach,” *Academia. Edu.*, vol. 12, no. 3, 2021. http://dx.doi.org/10.34218/IJARET.12.3.2021.062.

[12] V. Sathiyamoorthi, A. K. Ilavarasi, K. Murugeswari, S. Thouheed Ahmed, B. Aruna Devi, and M. Kalipindi, “A deep convolutional neural network based computer aided diagnosis system for the prediction of Alzheimer’s disease in MRI images,” *J. Int. Meas. Confederation*, vol. 171, Article ID: 108838, 2021. http://dx.doi.org/10.1016/j.measurement.2020.108838.

[13] L. Sun and D. Gao, “Security attitude prediction model of secret-related computer information system based on distributed parallel computing programming,” *Math. Probl. Eng.*, vol. 2022, pp. 1-13, 2022. http://dx.doi.org/10.1155/2022/3141568.

[14] A. Radovan, V. Šunde, D. Kučak, and Ž. Ban, “Solar irradiance forecast based on cloud movement prediction,” *Energies*, vol. 14, no. 13, Article ID: 3775, 2021. http://dx.doi.org/10.3390/en14133775.

[15] B. I. Sighencea, R. I. Stanciu, and C. D. Căleanu, “A review of deep learning-based methods for pedestrian trajectory prediction,” *Sensors*, vol. 21, no. 22, Article ID: 7543, 2021. http://dx.doi.org/10.3390/s21227543.

[16] X. Niu, J. Wang, and L. Zhang, “Carbon price forecasting system based on error correction and divide-conquer strategies,” *Appl. Soft Comput.*, vol. 118, Article ID: 107935, 2022. http://dx.doi.org/10.1016/j.asoc.2021.107935.

[17] W. Yang, J. Wang, T. Niu, and P. Du, “A hybrid forecasting system based on a dual decomposition strategy and multi-objective optimization for electricity price forecasting,” *Appl. Energ.*, vol. 235, no. 1, pp. 1205-1225, 2019. http://dx.doi.org/10.1016/j.apenergy.2018.11.034.

[18] W. Yang, S. Sun, Y. Hao, and S. Wang, “A novel machine learning-based electricity price forecasting model based on optimal model selection strategy,” *Energy*, vol. 238, 2022. http://dx.doi.org/10.1016/j.energy.2021.121989.

[19] A. T. Jebb, S. Parrigon, and S. E. Woo, “Exploratory data analysis as a foundation of inductive research,” *Hum. Resour. Manage. R.*, vol. 27, no. 2, pp. 265-276, 2017. http://dx.doi.org/10.1016/j.hrmr.2016.08.003.

[20] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, “Exploratory data analysis using python,” *Int. J. Innov. Technol. Exploring Eng.*, vol. 8, no. 12, pp. 4727-4735, 2019. http://dx.doi.org/10.35940/ijitee.L3591.1081219.

[21] M. Staniak and P. Biecek, “The landscape of R packages for automated exploratory data analysis,” *R J.*, vol. 11, no. 2, 2019. http://dx.doi.org/10.32614/rj-2019-033.

[22] X. Liu, F. J. Zhang, Z. Y. Hou, L. Mian, Z. Y. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE T. Knowl. Data En.*, vol. 2021, 2021. http://dx.doi.org/10.1109/TKDE.2021.3090866.

[23] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE T. Pattern Anal.*, vol. 43, pp. 4037-4058, 2021. http://dx.doi.org/10.1109/TPAMI.2020.2992393.

[24] S. R. Khonde and V. Ulagamuthalvi, “Hybrid architecture for distributed intrusion detection system using semi-supervised classifiers in ensemble approach,” *Adv. Model. Anal. B*, vol. 63, no. 1-4, pp. 10-19, 2020. http://dx.doi.org/10.18280/ama_b.631-403.

[25] Y. Luo, B. Y. Wang, Y. Zhang, and L. M. Zhao, “A novel fusion method of improved adaptive LTP and two-directional two-dimensional PCA for face feature extraction,” *Optoelectron. Lett.*, vol. 14, no. 2, pp. 143-147, 2018. http://dx.doi.org/10.1007/s11801-018-7226-7.

[26] F. Es-Sabery, K. Es-Sabery, J. Qadir, B. Sainz-De-Abajo, A. Hair, B. García-Zapirain, and I. D. L. Torre-Díez, “A map reduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier,” IEEE Access, vol. 9, Article ID: 58706, 2021. http://dx.doi.org/10.1109/ACCESS.2021.3073215.

[27] D. A. Fife and J. L. Rodgers, “Understanding the exploratory/confirmatory data analysis continuum: moving beyond the replication crisis,” *Am. Psychol.*, vol. 77, no. 3, pp. 453-466, 2021. http://dx.doi.org/10.1037/amp0000886.

[28] J. T. Behrens, “Principles and procedures of exploratory data analysis,” *Psychol. Methods*, vol. 2, no. 2, pp. 131-160, 1997. http://dx.doi.org/10.1037/1082-989X.2.2.131.

[29] E. Karageorgiou, “The Logic of Exploratory and Confirmatiry Data Analysis,” *Cogn. Crit.*, vol. 3, pp. 35-48, 2011.