Support Vector Regression with Gaussian kernel for Housing Prices Prediction

Support Vector Regression with Gaussian kernel for Housing Prices Prediction

Dingyang Miao[1, a, *, †], Hongru Tang[2, b, *, †], Boshen Wang[2, c, *, †]

[1] School of Physics, Xiamen University, Xiamen, Fujian, 361005, China

[2] School of Economics, Xiamen University, Xiamen, Fujian, 361005, China

[*]Corresponding author email: [a]19720182203945@stu.xmu.edu.cn

[b]15220182202642@stu.xmu.edu.cn

[c]15220182202658@stu.xmu.edu.cn

[†]These authors contributed equally.

Abstract. The housing sector is one of the main sources of economic growth in both developing and developed countries. It is reported that nearly half of people buy or sell houses at an inappropriate price. Based on the public data set of Boston housing prices, this essay analyzed the factors affecting house prices and selected the five most important factors based on the decision tree with the ID3 algorithm. Then, this essay developed the support vector regression (SVR) with Gaussian kernel to predict housing prices. Experimental results showed that our method achieves superior accuracy and effectiveness compared with the SVR with linear kernel, KNN, and decision tree. To verify the applicability of our model, this research applied this model in Beijing housing price data, and it also achieved satisfactory fitting results.

1. Introduction

Housing price is a major social and economic problem related to the life quality of residents. For most people, buying a house is undoubtedly a huge investment. However, a report by Churchill, a British insurance company, pointed out that 47% of homebuyers regretted their purchase in the first year [1]. This phenomenon is the instability of housing prices, which are affected by many factors such as politics, region, and economy [2]. Therefore, people often lack understanding of the housing market situation but only rely on third-party house agents. The unequal cognition differences result in price chaos in the housing market. The seller sold his house at the wrong time, and the buyer bought the house at an unreasonable price.

This article developed the support vector regression (SVR) with Gaussian kernel to predict housing prices. The researchers analyzed the correlation between house prices and the factors, showing their significant relationship. Then the researchers selected the five most important factors based on the decision tree with the ID3 algorithm and divide the data set into a training set and a test set. To verify the validity of the methods, this essay compared SVR with Gaussian kernel with different regression methods[3]. The researchers visualized the above values by histogram and plotted the predicted values against the actual values. As a result, this research verified the accuracy and effectiveness of this model. Besides, the researchers utilized real-life data of Beijing housing prices to conduct experiment verification [4].

2. Methodologies

2.1. Feature Selection

To improve the efficiency of the learning process, the researchers designed the decision tree model based on the Iterative Dichotomies 3 (ID3) algorithm to make feature selection [5]. The features chosen need to have classification ability for training data. If the features participate in classification and the results of random classification are not different, the features would be regarded with no classification ability, so they will not have a special impact on the learning accuracy if they are omitted.

2.1.1. Entropy

The entropy measures the uncertainty of random variables. It could be calculated as follows:

$$H = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \quad (1)$$

We can find that entropy is the expected value of information. Therefore, the smaller the information entropy is, the higher the purity of information, which means the less information is, the less the classification field is, the less the categories are contained in it [6].

2.1.2. Information Gain

The information gain measures the degree of information uncertainty reduction under certain circumstances. The value of information entropy could calculate it before classification minus the value of information entropy after classification [7].

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|} \quad (2)$$

The greater the information gain is, the better the classification effect would be.

2.1.3. Selection Process

The ID3 algorithm takes the descending rate of information entropy as the criterion for selecting test attributes. That is, the attribute with the highest information gain that has not been used to partition is selected as the partition criterion in each node, and then the process continues until the generated decision tree can perfectly classify training samples [8].

2.2. Support Vector Regression (SVR)

The SVR is an application of a support vector machine (SVM) to regression problems. So before introducing SVR, the essay will briefly introduce what SVM is.

The SVM is a set of supervised learning methods used for classification. And the basic model is a classifier finding the biggest margin in the space [9]. The SVM is to classify two kinds of data points and predict which class the new data point belongs. In SVM, people are supposed to use a (p - 1) to separate p-dimensional data points. Similarly, people also need to find a hyperplane in SVR, as shown in figure 1.

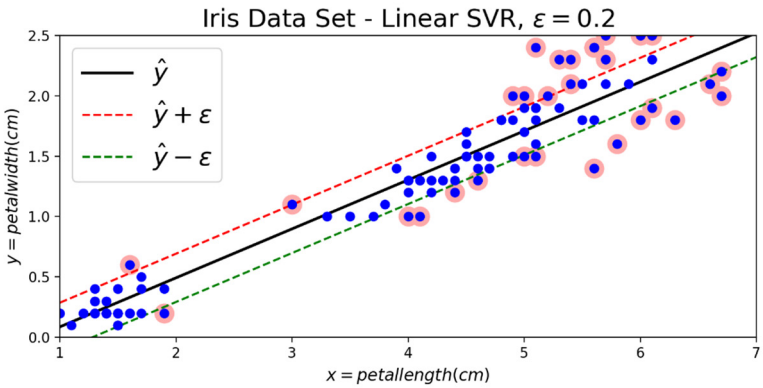


Figure 1 Linear SVR.

The traditional linear regression model is to find a curve to minimize the residual. However, as shown in the figure above, the SVR uses a strip to fit the data. The width of this strip can be set by yourself and controlled by parameter ϵ. The researchers defined the residual error of the data point within the dotted line as 0, and the distance from the data point outside to the boundary is the residual error (ζ). Like linear models, the research wants minimal residuals (ζ) [10].

$$\zeta(x,y) = \begin{cases} 0, & | yi - w \cdot \phi(xi) - b | \leq \epsilon \\ | y - w \cdot \phi(x) - b | - \epsilon, & | yi - w \cdot \phi(xi) - b | > \epsilon \end{cases} \tag{3}$$

For non-linear models, the kernel function is used to map the feature space and then perform the regression.

3. Results and Discussion

3.1. Datasets

The essay selected two data sets this time, one is the public data set in Boston which has been widely used, and the other one is the actual data set in Beijing.

The Boston data set comes from the UCI machine learning knowledge base. It began to be counted in 1978, with 506 pieces of housing price information covering 14 features in different suburbs of Boston, Massachusetts. After this data set was included in Kaggle as a competition, it gradually became a classic machine learning project. The features explanation of the Boston dataset is shown in Table I.

Table 1 The features explanation of the Boston dataset.

Features	Description	Features	Description
CRIM	per capita crime rate by town	NOX	nitric oxides concentration (parts per 10 million)
ZN	the proportion of residential land zoned for lots over 25,000 sq. ft.	RM	the average number of rooms per dwelling
INDUS	the proportion of non-retail business acres per town	AGE	the proportion of owner-occupied units built before 1940
CHAS	Charles River dummy variable (= 1 if tract bounds river; = 0 otherwise)	DIS	weighted distances to five Boston employment centres
LSTAT	% lower status of the population	MEDV	The median value of owner-occupied homes is $1000's
RAD	index of accessibility to radial highways	PTRATIO	pupil-teacher ratio by town
TAX	full-value property-tax rate per $10,000	B	1000 (Bk-0.63)^2 where Bk is the proportion of blacks by town

The Beijing data set comes from Lianjia, a well-known housing trading platform in China. It records more than 20,000 pieces of housing price information and 25 features of each house. This data set was recorded from 2002 to 2018, with a period of up to 16 years. The detailed description of the features is as Table 2.

Table 2 The features explanation of the Beijing dataset.

Features	Description	Features	Description	Features	Description
Area	Area of the house	Bighouse0	Stacked townhouse	Car	Numbers of parking lots
Release Time	Time after the house was released on the website	Bighouse1	Townhouse	Room	Numbers of rooms
TotalPrice	The total price of the house	Bighouse2	Single-family townhouse	Office	Numbers of offices
Dist	District in Beijing	Bighouse3	Semi-detached townhouse		

3.2. Assessment indicators

To evaluate the validity and precision of the models developed by these four different regression methods. The evaluation standards are R^2 score and *MSE* (Mean Squared Error) [11].

The *MSE* is generally used to measure the deviation between the predicted value and the true value of the model. For instance, the researchers obtain *N* samples in the training set and *M* samples in the test set to establish the model and get the M predicted values $\widehat{y_m}$. Consequently, the essay calculated *MSE* as:

$$MSE = \frac{1}{M}\sum_1^M (y_m - \widehat{y_m})^2. \quad (4)$$

The lower the value of *MSE* is, the better the model fits. The R^2 also measures the deviation between the predicted value and the true value of the model, calculated as:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2}. \quad (5)$$

The maximum value of R^2 is 1, which means the predicted value is equal to the true value. The closer the value of the R^2 score to 1, the better the model fits.

3.3. Feature selection

Not all features are closely related to housing prices from our life experience, making it necessary to extract features. The researchers first draw a statistical distribution map of all features for the Boston dataset, and found that the distribution of several features was very close to that of prices.

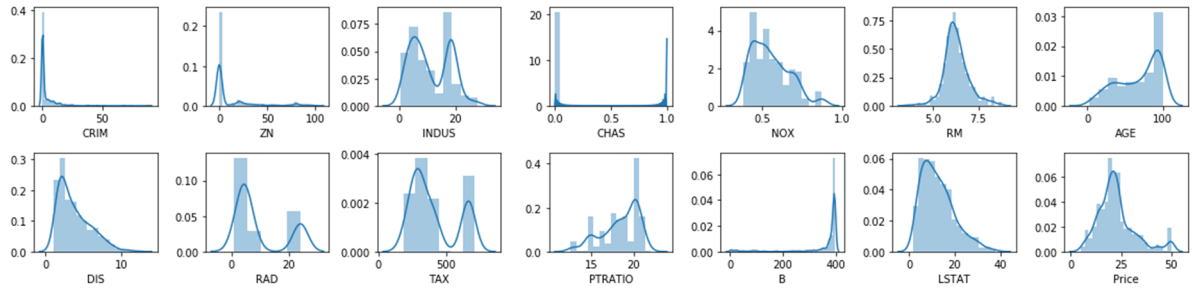
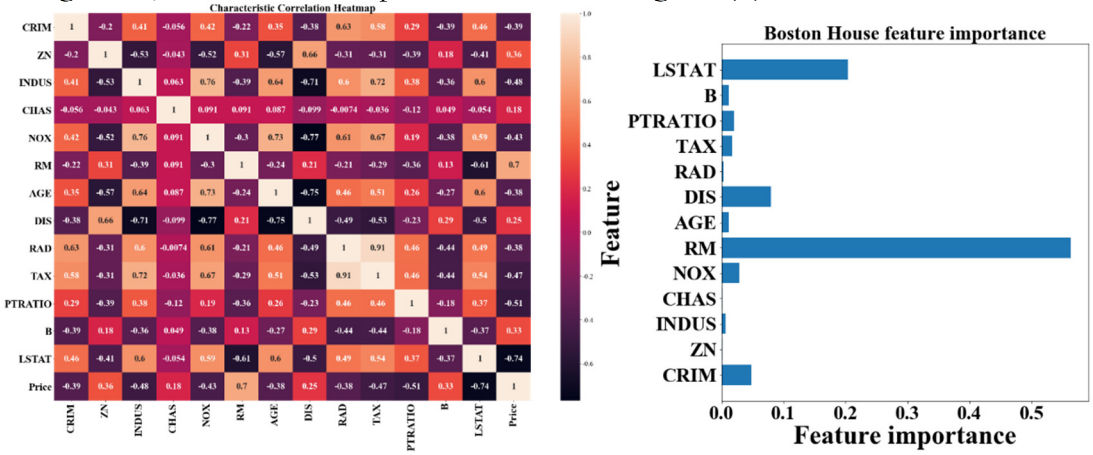


Figure 2 Distribution of Boston data set.

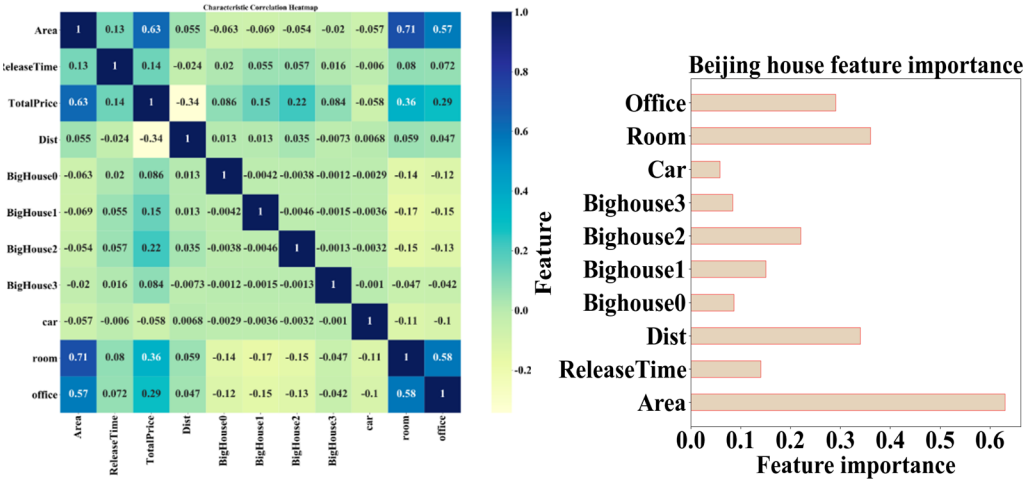
Then, the researchers drew a heat map of the correlation coefficient between all features for Boston data, shown in Figure 3(a). It can be found that LSTAT has the highest correlation with price, followed by RM, PTRAIO, TAX, NOX. Considering that many features are not strongly correlated with housing prices, bringing in all the data may introduce difficulties and operation time of machine learning. To solve this problem, the essay selected the five most important features based on the decision tree with the ID3 algorithm, and the feature importance is shown in Figure 3(b).



(a) (b)

Figure 3 The feature importance of the Boston dataset. (a) Heatmap of Boston dataset; (b) Feature importance of Boston dataset.

For the Beijing data set, the features selection has become much more difficult since more characteristics affect prices, such as house type, unit price, profile, and community. Based on the feature selection of the decision tree with the ID3 approach, the researchers selected the 7 most important features. The selected features for the Beijing dataset are shown in Figure 4.



(a) (b)

Figure 4 The feature importance of the Beijing dataset. (a) Heatmap of Beijing dataset; (b) Feature importance of Beijing dataset.

3.4. Housing price prediction

To verify the validation of SVR with Gaussian kernel, the essay compared the methods with the *K*-Nearest Neighbor (KNN), the SVR with linear kernel, and the decision tree, and the prediction results are shown in Figure 5. Here, the essay used the grid search to obtain the optimal parameters for these models [12].

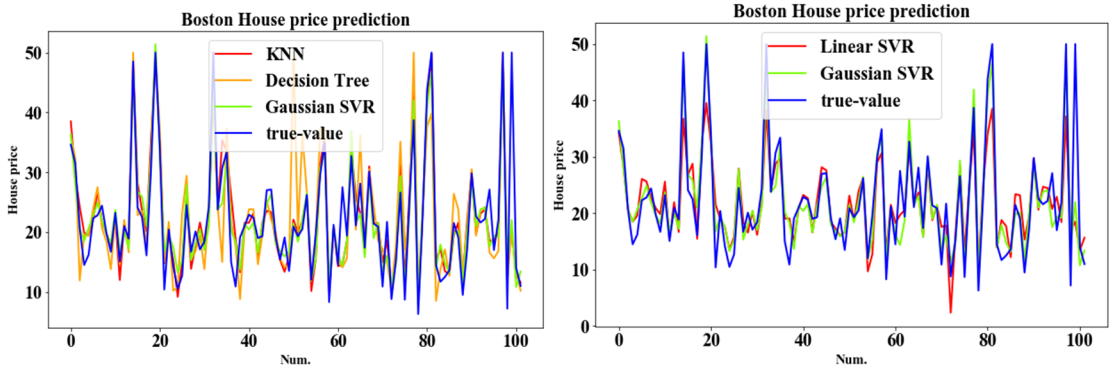


Figure 5 Comparison of prediction results of KNN, the SVR with linear kernel, and decision tree, and our method.

The experimental results from Figure 5 show that the SVR with Gaussian kernel achieves superior fitting results. The further R^2 score and *MSE* results are shown in Figure 6. It can be found that the SVR with Gaussian kernel obtains the best performance, with the best R^2 score ($R^2 = 0.81$) and the best *MSE* (*MSE* = 18.13).

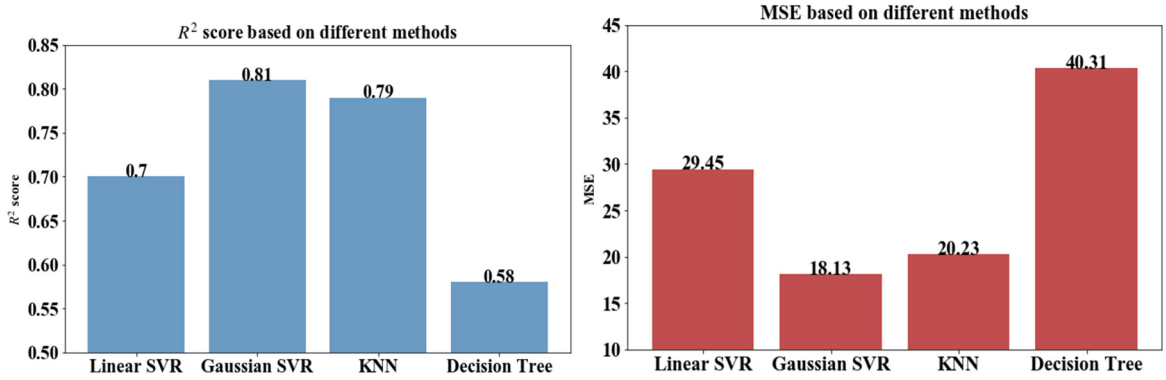


Figure 6 R^2 score and MSE of models.

After finishing the previous process, the essay has completed an excellent housing price prediction model. However, if this prediction model is only applied to the public data set in Boston in 1978, it obviously cannot prove its application capabilities nowadays. Therefore, the researchers applied our prediction model to the actual Beijing dataset [13]. After removing obvious meaningless features, such as URLs, trade numbers, and then explore the remaining 7 features in the next step.

Based on the experience of the Boston housing prediction model and observing the similarity between the two data sets, the researchers compared the Gaussian SVR with the linear SVR on this dataset. The fitting result is shown in Figure 7. The experimental results show that the SVR with Gaussian kernel gains optimal fitting performance. Figure 8 presents the further R^2 score and *MSE* comparison results. It can be found that the SVR with Gaussian kernel achieves optimal performance ($R^2 = 0.77$ and $MSE = 43.13$), which is significantly better than the SVR with linear kernel.

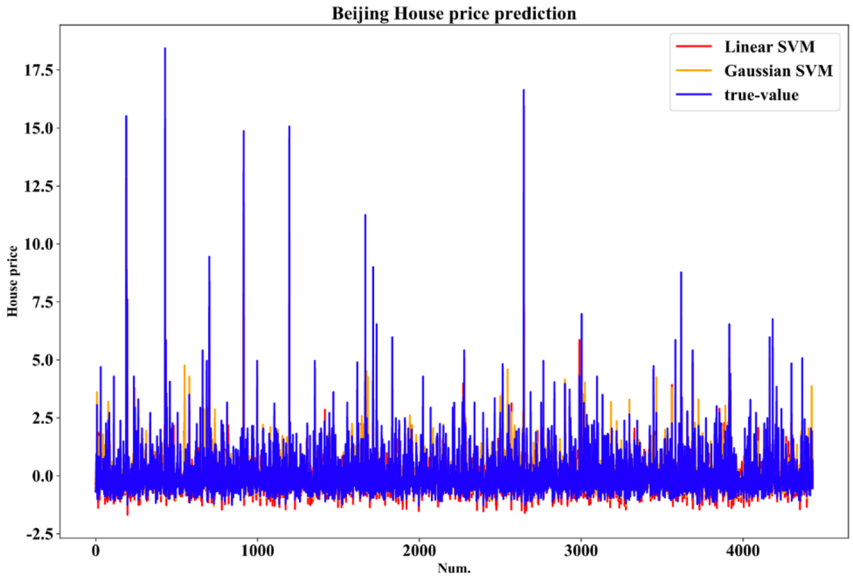


Figure 7 Prediction of Beijing under Linear and Gaussian SVR

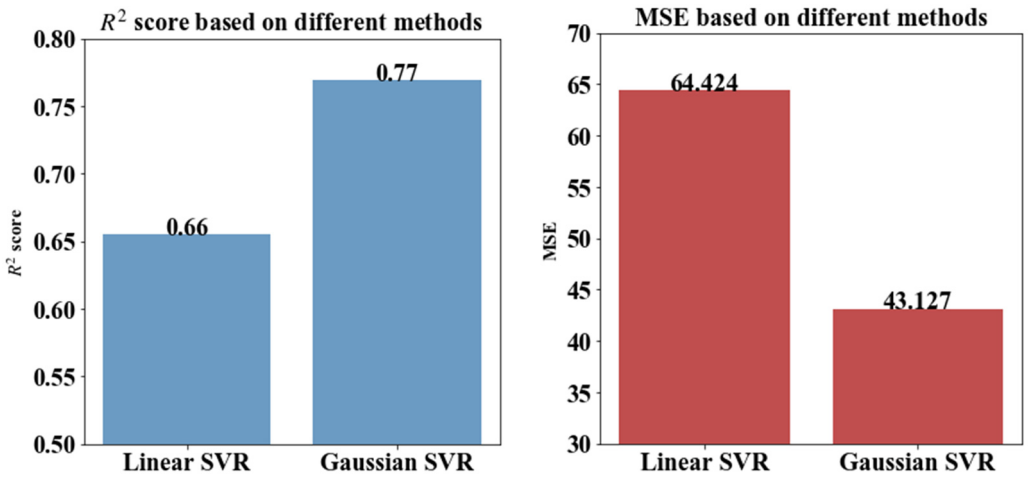


Figure 8 R^2 score and MSE of models.

4. Conclusion

The essay developed the SVR with Gaussian kernel to predict housing prices to provide house traders with a fair valuation in this work. The researchers selected the Boston public data set and the Beijing actual data set for approach verification.

First of all, the essay performed feature selection based on the decision tree with ID3 algorithm and selected effective features for subsequent housing price prediction, which is very important for filtering out invalid features to help effective model prediction. Furthermore, based on the effective features of the screening, the reseachers designed the SVR with a Gaussian kernel function for housing price prediction. To verify the effectiveness of the methods, the essay compared the methods with KNN, decision tree, and SVR with a linear kernel function. The experimental results show that our method achieves the best performance.

Through this model, both buyers and sellers can estimate the price based on the comprehensive information of the house, which will effectively protect their rights. In addition, an excellent housing price prediction model can make the market more transparent and greatly improve the housing market's fairness. At the same time, housing brokers can also form a virtuous circle. However, our model still has some limitations. The most obvious is that some factors that affect house prices cannot be quantified, but they are also important at certain times, such as decoration style, community cultural tolerance and so on.

References

[1] Glaeser E L, Nathanson C G. An Extrapolative Model of House Price Dynamics[J]. *Journal of Financial Economics*, 2017, 126(1):147-170.

[2] Myoungsub, Choi, Sehil, *et al.* Comparison on Forecasting Performance of Housing Price Prediction Models in Seoul[J]. *Seoul Studies*, 2016(3):75-89.

[3] Avanija J, Sunitha G, Madhavi K R, *et al*. Prediction of House Price Using XGBoost Regression Algorithm[J]. *Turkish Journal of Computer and Mathematics Education* (*TURCOMAT*), 2021, 12(2):2151-2155.

[4] Zhang, Jing. House Price Expectations: Unbiasedness and Efficiency of Forecasters[J]. *Real Estate Economics*, 2016, 44(1):236–257.

[5] Haurin D, Ma C, Moulton S, *et al*. Spatial Variation in Reverse Mortgages Usage: House Price Dynamics and Consumer Selection[J]. *Journal of Real Estate Finance & Economics*, 2016, 53(3):392-417.

[6] Hu Y, Oxley L. Exuberance, Bubbles or Froth? Some Historical Results using Long Run House Price Data for Amsterdam, Norway and Paris[J]. *Working Papers in Economics*, 2016.

[7] X. H. Xu and X. Luo, "Information entropy risk measure applied to large group decision-making method," *Soft Computing,* vol. 23, no. 1, 2019.
[8] V. Viswanathan, S. Ramakrishnan, and K. Sk, "Diabetic and Kidney Disease Prediction in Human based on Their Age Group using C4.5 Decision Tree Algorithm in Python," *Test Engineering and Management,* vol. 82, no. 1, pp. 8335-8342, 2020.
[9] H. Zhao, Y. Gao, H. Liu, and L. I. Lang, "Fault diagnosis of wind turbine bearing based on stochastic subspace identification and multi-kernel support vector machine," *Journal of Modern Power Systems and Clean Energy,* 2019.
[10] Z. Mei, W. Zhang, L. Zhang, and D. Wang, "Real-time multistep prediction of public parking spaces based on Fourier transform–least squares support vector regression," *Journal of Intelligent Transportation Systems,* pp. 1-13, 2019.
[11] D. Preethi and N. Khare, "Sparse auto encoder driven support vector regression based deep learning model for predicting network intrusions," *Peer-to-Peer Networking and Applications,* no. 1, 2020.
[12] L. Yao, Z. Fang, Y. Xiao, J. Hou, and Z. Fu, "An Intelligent Fault Diagnosis Method for Lithium Battery Systems Based on Grid Search Support Vector Machine," *Energy,* vol. 214, p. 118866, 2021.
[13] https://www.kaggle.com/ruiqurm/lianjia.