A Novel Hybrid House Price Prediction Model

Süreyya Özöğür Akyüz

Faculty of Engineering and Natural Sciences, Department of Mathematics, Bahçeşehir University, Istanbul, Turkey

Birsen Eygi Erdogan

Faculty of Arts and Sciences, Department of Statistics, Marmara University, Istanbul, Turkey

Özlem Yıldız

Big Data Analytics Program, Institute of Science, Bahçeşehir University, Istanbul, Turkey

Abstract

The real estate sector is evolving and changing rapidly with the increase in housing demand, and new luxury housing projects appear every day. The reliability of housing market investments is largely dependent on accurate pricing. The aim of this study is to introduce a dynamic pricing procedure that estimates housing prices using the most important attributes of a house. To this end, a hybrid modeling system is proposed employing linear regression, clustering analysis, nearest neighborhood classification, and the Support Vector Regression (SVR) method. The housing data of the Kadıköy area in Istanbul, collected via manual web scraping, was used for the training and validation of the proposed algorithm. The results of the hybrid model were compared using multiple linear regression, ridge regression, and Support Vector Machines (SVMs). The experimental results show that the proposed model is superior, both in terms of Residual Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) measures. Therefore, the proposed dynamic hybrid modelling structure can be successfully used for predicting house pricing.

Keywords: `Ensemble Models, Housing Pricing, Support Vector Regression, K-means clustering, K-NN classification`

2010 MSC: 00-01, 99-00

1. Introduction

Recent research has demonstrated a rapid increase in housing investments in the construction sector – a sector which plays a significant role in countries' economic development. For this reason, evaluating the factors affecting the housing market and predicting the price of a house based on its characteristics have gained increasing academic attention. The last decade has seen rapid growth in the construction sector in Turkey with major investments in infrastructure and urban renewal projects. Meanwhile, the demand for housing in Turkey has also increased due to the increase in the population, changes in lifestyle, and increase in living standards. For this reason, nowadays, improving the existing pricing mechanism, a complicated task which has always been carried out by the real estate agent, and developing more modern mechanisms to price real estate in Turkey is critical for all parties involved in the real estate market.

In [11], property prices are determined by two different approaches, referred to as traditional and advanced valuation methods. Traditional valuation methods include the comparable method, investment/income method, profit method, development/residual method, contractor's method, cost method, multiple regression method, and stepwise regression method. Advanced valuation methods, for their part, are used to simulate the decision processes of the market players and predict the change in these decisions. Artificial Neural Networks (ANNs), hedonic pricing method, spatial analysis method, Fuzzy Logic (FL), and autoregressive integrated moving average (ARIMA) methods can be given as examples of advanced valuation methods.

In many studies, the hedonic approach and multiple regression analysis are used, which are also known as hedonic regression. For a 'one-stop' reference of hedonic approaches, one can look at the review study of [7]. In [14] earlier hedonic house price model is reassessed by adding the average age of homes in

Boston. Using the statistical average age of the houses, his results demonstrated that there is heteroscedasticity in housing data. Stevenson referred to study in [5], who provided evidence that the age of the dwelling is a primary cause of heteroscedasticity. The heteroscedastic nature of the housing data inspired us to use a clustering approach as a preprocessing step of the housing data before we use an appropriate model for price prediction of the houses.

In [2], a model using Geographic Information System (GIS) data is developed to take heteroscedasticity into account using the location and characteristics of the houses. In this study, a hedonic price function was estimated using semi-parametric regression. The performance of the semi-parametric regression model was observed to be more efficient in comparison to the traditional parametric models. Another study in [3], applied a decision tree approach on the Singapore resale public housing market. The results showed the usefulness of this method in determining the relationship between house prices and housing characteristics.

Over the last decade, artificial neural networks and fuzzy logic methods have also been used to predict house prices. One recent study on housing pricing estimation in Eskişehir, Turkey employed a fuzzy logic approach in which the distances from the house to cultural, educational, medical buildings, transportation systems, in addition to other environmental attributes were taken as the main features of the data, [10].

Beyond these features, it has also been observed that house selling prices vary due to real estate agencies' marketing capabilities. According to Kuşan's view, the hedonic method and multiple regression methods are not enough to estimate house prices and cannot deal with problems such as outliers, nonlinearity, discontinuity, and fuzziness. Therefore, in their study, house unit prices were estimated using fuzzy logic where environmental, transportational and regional socio-economic factors were used as independent variables.

In [13] the determinants of house prices in Turkey for the whole country are considered, including both urban and rural areas, using both hedonic regression and ANN methods. Some hybrid and smarter approaches in the prediction

of housing prices have also arisen over the last decade. For example, in [4] the performance of Adaptive neuro-fuzzy (ANFIS) with grid partition and sub clustering models is compared. It was shown that ANFIS with grid partition models performed better than ANFIS with sub clustering models. In [12] various classification methods are used such as C4.5, RIPPER, Naïve Bayesian, and, AdaBoost to predict if a townhouse would be sold for less or more than the list price using the house's characteristics. They compared their classification accuracy performances to other classification methods, demonstrating the RIPPER algorithm's superiority.

In this paper, we propose a novel hybrid method which, unlike existing methods such as hedonic regression or machine learning approaches, integrates both hedonic regression and machine learning techniques, including clustering and classification methods. To the best of our knowledge, there has not been such a hybrid method that combines those two approaches to the problem of house pricing. House pricing data from the Kadıköy area of Istanbul, Turkey were collected via the website [17], which serves as a real estate database for the public. The experimental results show that the proposed approach outperforms the existing literature methods.

The rest of the paper is organized as follows: In the next section, we provide an abstract definition of the methods used in our hybrid approach. In Section 3 we describe our hybrid method, which is followed by Experiments and Results in Section 4 and Section 4.2, respectively. Finally, we conclude in Section 5 with a summary and discussion of the advantages of the proposed method.

2. Background Material

2.1. Multiple Regression

Linear regression is a well-known method with a wide scope of application areas. The fundamental aim of multiple regression is to learn the relationship between several independent variables and a dependent variable. For instance, a real estate specialist may want to see how the price of a house is affected by the

size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of the appeal of the house. One might learn that the size of the flat is a better predictor of the price than the floor that it is on.

Multiple linear regression considers the model of the distribution of a continuous type quantitative response variable Y_i of the $i-th$ observation for given explanatory variables $X_{i1}, \ldots, X_{ip}$ as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \tag{1}$$

where β_0 refers to intercept, $\beta_1, \ldots, \beta_p$ are regression coefficients and ϵ_i denotes the error term which has zero mean and which captures the residual variability. Estimation of these paramaters β_i is done by well-known methods such as least square or maximum likelihood estimation [16].

In a multicollinearity case, Ridge regression is one of the commonly used methods to estimate the coefficients of variables [8]. In this study, we compared the prediction performance of the proposed hybrid model with ridge regression because of the multicollinear structure of the data. The most related two independent variables, causing the multicollinearity, are the area of the house and the number of the rooms. Researchers dealing with housing data often observe the effect on the parameter estimation in their attempt to use multiple linear regression for price prediction. Because of the very well-known effect of multicollinearity, the coefficient of the room number may appear negative while the coefficient of the area of the house is positive. Logically, we expect both of them to be positive. That is why we suggest the use of ridge regression instead of multiple linear regression. Besides, it is known that even though multicollinearity effects the variances of the coefficients, still may be good at prediction. Therefore, a researcher may also use multiple linear regression instead of ridge regression if the primary aim of the study is prediction rather than interpretation of the coefficients.

2.2. K-means Clustering

Clustering is a method that groups similar objects by minimizing the distance within the cluster while maximizing the distance between the clusters. K-means is one of the most preferable methods of clustering techniques, as it minimizes the distance between the centroids of the clusters. During the clustering process, observation/examples are added iteratively until the smallest distance is achieved [9]. Here, K refers to the number of clusters given initially to the algorithm as a parameter. In our study, the appropriate number of clusters is determined using the so called Elbow method that uses a grid search algorithm included in scikit learn. The overview of the clustering algorithm can be explained in the following three steps:

1. The input space is divided in to K clusters and examples are randomly assigned to the clusters,
2. **For** each data point
   - Find the distance between the example and centroid of the cluster,
   - **If** the example has the shortest distance to its own cluster **then** leave it **else** select another cluster.
3. **Repeat** steps 1 and 2 untill there is no observation left to be moved from one to another cluster.

The common distance measures in K-means algorithm are the Euclidean distance, the Euclidean squared distance and the Manhattan or City distance. In this paper, we used Euclidean distance for K-means method.

2.3. K-NN classifier

K Nearest Neighbor (K-NN) method is the simplest algorithm that can be used both for regression and classification. It is preferred due to its low calculation time and ease of interpretation.

For given training data $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, and a test point (x_t, y_t), K-NN algorithm predicts the class labels of the test examples

by looking at K most similar training examples. In classification cases, it assigns the majority class label (majority voting). Similarly, it assigns the average response for regression. K-NN is also called a non-parametric method as it does not learn an explicit mapping f from the training data but rather uses the training data at the test time to make predictions. It needs an input of K nearest neighbors and the distance function to compute the similarities between the examples. Here we used K as 1. There are several ways to compute distances based on the type of features. For example, Euclidean distance is commonly used for real-valued features, whereas Hamming distance is preferred for binary valued features [1]. We have used the Euclidian distance.

2.4. Support Vector Regression (SVR)

Support Vector Machines are developed for classification problems in general [15], and have been extended to regression analysis with a new heading called Support Vector Regression (SVR). The method is linear regression in the sense of hyperplanes determined in SVR, but it is nonlinear in the sense of interpreting the data points in output space using kernel functions. SVR has quickly become popular because it does not have any assumption for the data.

For a given training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, SVR fits the data points into a (hyper-) tube of width 2ϵ with $\epsilon \geq 0$. This hypertube can be considered as a regression model which fits the data points with a hyperplane positioned in its center. There are many possible ways to locate a hypertube of width 2ϵ. However, there exists an optimal hypertube that contains as many training points as possible. The optimal hypertube can be found by maximizing the distance of observations from the center hyperplane which has the same idea of maximizing margin in SVM principle [6]. SVR has two parameters C and ϵ to be optimized where epsilon can be considered as a parameter that affects the accuracy of the solution, but in most cases the solution is required to be as accurate as possible. In order to overcome this problem, ϵ can be included as a part of the optimization problem which turns into a $\nu-$SVR problem where $0 \leq \nu \leq 1$ is the regularization constant in the

objective function of the optimization problem below:

$$\begin{aligned} \min_{w,b,\xi,\xi^*,\epsilon} \quad & \frac{1}{2}w^T w + C(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \\ & (w^T\phi(x_i) + b) - y_i \leq \epsilon + \xi_i, \\ & y_i - (w^T\phi(x_i) + b) \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l, \ \epsilon \geq 0. \end{aligned} \tag{2}$$

Here, w is the normal of the hyperplane, l refers to the number of observations in the training set, ξ_i is the $i-th$ slack variable ($i = 1, \ldots, l$) and $\phi(.)$ is a nonlinear mapping from input space to output space.

In ν-SVR, the parameter ν is used to determine the proportion of the number of support vectors we desire to keep in our solution with respect to the total number of samples in the dataset. In this study, we used $\nu-$ SVR and the parameters of $\nu-$SVR were determined on the training set by using the well-known classical model selection method "k-fold cross validation".

3. Hybrid Method

For the real estate market, it is very important to determine the value of a house. For the people who are trying to automate this determination, the predetermined features are important. In this study, the first step of the hybrid model consists of partitioning the data into train and test folders, where regression is applied on the training folder and prediction errors are obtained on the test folder. It was observed that regular application of regression revealed errors in approximately three categories, shown by simple scatter diagram in Figure 1 (error versus y prediction). Here, the first category corresponds to the house prices that the regression model underestimated, the second category stands for the house prices that are closely predicted, and the third category includes the house prices that are overestimated.

These breaks in the modeling, inspired the use of a clustering approach on errors, to discover the training subsets. Indeed, the clustering idea came from

the nature of the data at hand. It is known that housing data have some sub clusters due to regional and/or physical factors.

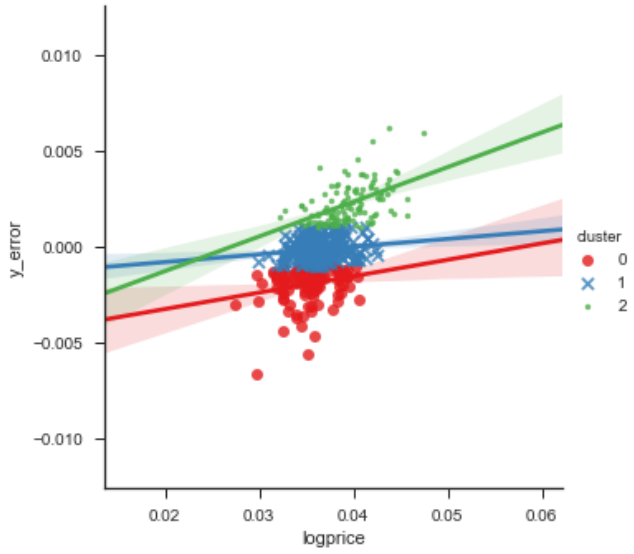

Figure 1: Error based clustered data visualization

In the second step, the K-means clustering algorithm was applied to *residual error vector*, which constitutes single dimension data, in order to find the number of classes which coincided with the visual results shown by Figure 1. The aim of using a clustering method is to automatically determine the break points of the errors revealed by regression analysis. The training folder labels were updated with the new clustering labels using K-means. In the third step, K-NN classification was applied on the updated training data to predict the class labels on the test folder. Once all class labels were predicted, updated training labels using K-means and predicted test labels using K-NN model were concatenated. In the final step, nu-SVR was applied to each cluster separately to predict the house prices. With this proposed hybrid model, it is possible to first classify the houses and then estimate the prices for each classes independently. All these steps are presented in Algorithm 3 and Figure 2.

Algorithm 1 Hybrid House Price Prediction Algorithm

Input: Data: (X, Y)

Output: Prediction Performance

1: split the data into Train and Test
2: do prediction on train data using linear regression
3: compute residual error vector on train data
4: do K-Means on residual error vector
5: set y train labels as the cluster of residual error vector determined using K-Means
6: do K-NN on train data
7: do prediction of y test labels on test data using K-NN
8: update y labels as concatenation of step 5 and step 7
9: **for all** cluster **do**
10: apply nu-SVR
11: **end for**
12: find the best parameters of nu-SVR by K-Fold cross validation
13: do prediction on test set

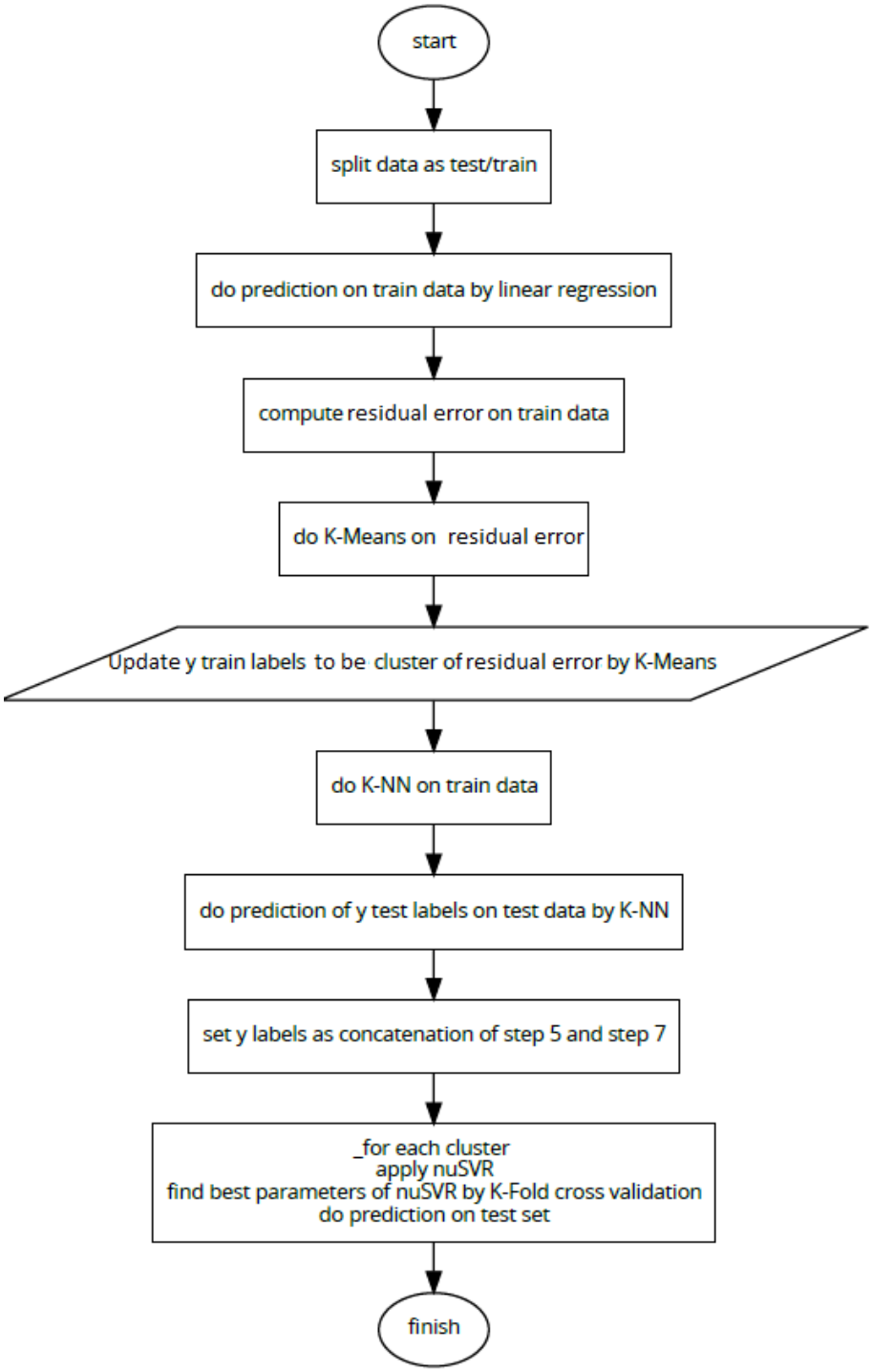

Figure 2: Flow chart of Algorithm 3

As a summary, first, the class labels were determined for all samples included in the data set by integration of K-means and K-NN, then nu-SVR was applied to each cluster for a final estimation. To find the best parameters of nu-SVR K-Fold cross-validation was used independently for each cluster. In this stage, a holdout validation was also used to test nu-SVR on the test folder.

The proposed hybrid approach was compared with the standard nu-SVR in terms of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Rsquare values. The results are discussed in the next section.

4. Experiments

4.1. Data

We collected 744 observations via manual web scrapping from the web site [17], which serves as a commercial platform to real estate agents, homeowners, and buyers. In data analysis, a typical pre- processing step was performed to scale variables, since variables measured at different scales will likely skew the analysis. For instance, the price of the house, i.e. the dependent variable, was very positively skewed and was corrected by a natural logarithm function. The house features that we collected include: square meters of the house (logm2), number of rooms (nroom), number of bathrooms (nbathroom-tr), floor number of the flat (floor), total number of floors in the building (nfloor), a dummy variable indicating if the house is in an enclosed residence or not (insite), and the year of the selling (year).

At the beginning of the study, the data were monitored by bivariate correlations and error terms of the multiple linear regression. It was observed that the parameter estimations were unstable and gave the wrong signs due to the correlations between features. This well-known, so called multicollinearity problem appeared when "the house size" and "the number of the rooms" were used in the same model. The relationship between the features can be seen in the heat map which reflects the correlations given in Figure 3.

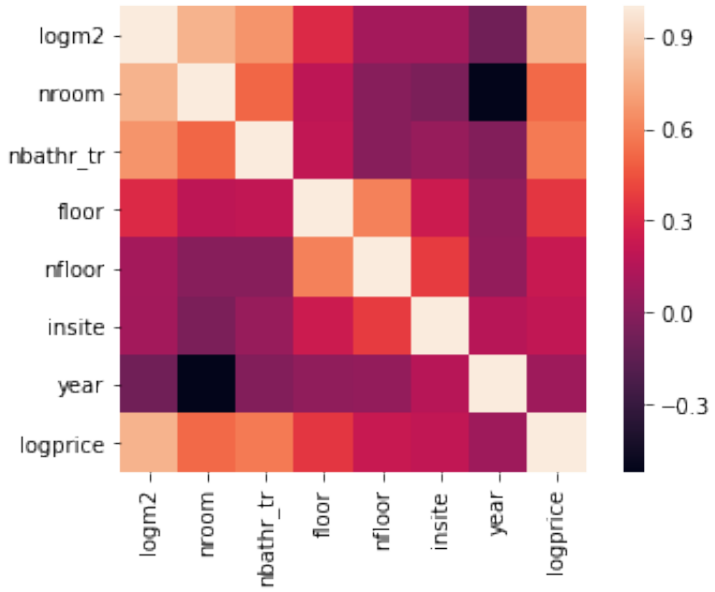

Figure 3: Heat map of multiple correlations

Here, the brighter the color, the stronger the correlation between the features. It is clear in Figure 3 that the features most correlated with price are house size (logm2), number of bathrooms, and the number of rooms.

4.2. Results

As is explained in Section 3, the data was first clustered using K-means algorithm and next nu-SVR was employed for each cluster. For the construction of the hybrid model Python 3 was used. Below, experimental results are presented for two different situations.

In the first situation, we summarized the results for multiple linear regression, ridge regression and nu-SVR obtained using 5 fold cross validation in Table 1. The reason for considering ridge regression is the multicollinearity problem caused by the use of two highly correlated variables, house size and the number of the rooms, at the same time. It was observed that although ridge regression made a slight correction on parameter estimation, the prediction performance

did not change significantly. Moreover, it was noted that nu-SVR performed better than linear regression and ridge regression in terms of both MAPE and Rsquare. The smallest MAPE and the largest Rsquare with nu-SVR were obtained, as is shown in Table 1.

Table 1: Standard Regression Results using 5 fold CV

Model	RMSE	MAE	MAPE	Rsquare
Linear Regression	0.0017	0.0013	3.50	0.65
Ridge Regression	0.0018	0.0014	3.63	0.62
nu-SVR	0.0017	0.0013	**3.42**	**0.67**

Using the proposed hybrid model, very competitive results were achieved, as given in Table 2, where MAPE values decreased from 3.46 to 2.40 and Rsquare value increased from 0.66 to approximately 0.78. Here, the results presented in Table 2 are the means of 10 random runs of the hybrid model.

Table 2: The Proposed Hybrid Algorithm vs Standard nu-SVR

Model	MSE	MAE	MAPE	Rsquare
nu-SVR	0.00168	0.0013	3.46	0.66
Std Dev*	0.00002	0.00001	0.04	0.008
Hybrid Model	0.00127	0.00090	**2.40**	**0.779**
Std Dev*	0.00003	0.00002	0.05	0.008

* Means and standard deviations were given for 10 runs

$$C = [0, 10, 100, 1000]$$

$$\nu = [0.1, 0.3, 0.5, 0.7, 1.0]$$

$$\gamma = [1e-4, 1e-3, 0.01, 0.1, 0.2, 0.5]$$

5. Conclusion and Discussion

The housing market is crucial for the country's economy, and the estimation of housing prices is very important for buyers and sellers. In this study, a hybrid model was proposed for modeling the relationship between the properties of houses and the housing prices. This hybrid modeling consisted of two stages. In the first stage, K-means clustering was employed on the residual error vector to update the training labels. The updated training folder was then classiffied using K-NN algorithm. In the second stage, house prices in each cluster were then estimated by nu-SVR. We compared the prediction performance of the proposed hybrid model with ridge regression because of the multicollinear structure of the data. The experimental results show that more successful results were obtained by the proposed hybrid approach based on nu-SVR in comparision to classical nu-SVR or ridge regression. This hybrid approach can be used to create a dynamic statistical learning-based system for house price prediction.

6. Credit Author Statement

Süreyya Özöğür Akyüz: Conceptualization, Methodology, Writing- Original draft preparation. **Birsen Eygi Erdogan:** Investigation, Data Collection Methodology, Reviewing and Editing. **Özlem Yıldız:** Software, Validation.

References

[1] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*, 175–185. doi:10.1080/00031305.1992.10475879.

[2] Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, *13*, 68–84.

[3] Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies*, *43*, 2301–2315.

[4] Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction*, *41*, 33–39.

[5] Goodman, A. C., & Thibodeau, T. G. (1997). Dwelling-age-related heteroskedasticity in hedonic house price equations: An extension. *Journal of Housing Research*, *8*, 299–317.

[6] Hamel, L. H. (2011). *Knowledge discovery with support vector machines* volume 3. John Wiley & Sons.

[7] Herath, S., & Maier, G. (2010). *The hedonic price method in real estate and housing market research. A review of the literature*. Technical Report WU Vienna University of Economics and Business.

[8] Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, *4*, 105–123.

[9] Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* volume 344. John Wiley & Sons.

[10] Kuşan, H., Aytekin, O., & Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. *Expert systems with Applications*, *37*, 1808–1813.

[11] Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, *21*, 383–401.

[12] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, *42*, 2928–2934.

[13] Selim, H. (2009). Determinants of house prices in turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*, 2843–2852.

[14] Stevenson, S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics*, *13*, 136–153.

[15] Vapnik, V. (1995). The nature of statistical learning theory springer new york google scholar, .

[16] Vapnik, V. (1998). *Statistical learning theory*. Wiley.

[17] www.sahibinden.com (2019). Housing data. URL: `https://www.sahibinden.com/`.