Journal of Intelligent Decision Making and Information Science

www.jidmis.org.org
eISSN: xxxx-xxxx

Predicting Duration of Residential Units' Sale Using Machine Learning Techniques

Farshid Abdi[1], Shaghayegh Abolmakarem[1], Amir Karbassi Yazdi[2]
[1] Department of Industrial Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Departamento de Ingeniería Industrial y de Sistemas, Facultad de Ingeniería,Universidad de Tarapacá, Arica, Chile

ARTICLE INFO

Article history:
Received 29 October 2024
Received in revised form 25 November 2024
Accepted 1 December 2024
Available online 5 December 2024

Keywords:

Real estate; Marketing; Classification; Regression; Feature Selection; Filtering methods; Data mining.

ABSTRACT

Real estate is vital to meeting basic needs and offering solid investments. Informed investment decisions require understanding home sale factors. Estimating residential property sales time can boost rewards and decrease risks. Two steps are involved in this paradigm. Significant characteristics are identified using filter weighting and regression techniques. K-nearest Neighbour, Naïve Bayes, and Decision Trees use characteristics identified in the first stage. Determining the most efficient model requires comparing their precision. The study provides brokers with insights for improved sales forecasting and transaction management. This technique allows stakeholders to adjust their plans based on market fluctuations, potentially leading to more profitable real estate investments. This research clarifies real estate investment strategies for greater returns.

1. Introduction

Housing is one of the vital elements of the country's economy and is a basic need and an essential means of investment. This dual nature distinguishes it from other goods and adds importance to various socio-economic fields. On the one hand, it is a highly consumed product that is one of the vital needs of humans and is considered the most expensive essential product for the household. On the other hand, as an immovable, enduring product, it is a capital good and a critical investment option in many countries [1]. It would be a family's biggest asset and look attractive to economic corporations. Housing is often the most substantial household expenditure, representing a significant portion of their income. This increases the demand for housing, an essential priority for individuals and families. It is also closely related to people's health, safety and well-being. Access to suitable housing is related to improving people's health and social stability and is influential in enhancing community cohesion and individual well-being.

When deciding to buy a home, putting enough time and effort into evaluating and considering all available options is essential. It is also essential to know the housing market's current state and accurately predict its future. These factors can lead to profitable investments and open up exciting opportunities for financial growth.

Real estate has unique features, such as being illiquid assets and not easily converted into cash without a loss [2]. Real estate has limited liquidity compared to other investments [1], and liquidity varies considerably over time. Increasing liquidity (i.e., sellers typically sell their houses after short marketing times) is one of the indications of the hotness of the real estate market [3]. Days on

Market (DOM), a key metric in the real estate industry, refers to the number of days a property is active on the market. At the micro level, DOM is a suitable metric to evaluate the popularity of the estate. At the macro level, it is an essential measure of the liquidity of the real estate market and demonstrates the level of risk related to real estate investments [1].

DOM is an essential criterion in real estate. A lower DOM indicates a home is in favorable condition, while a higher value may indicate unfavorable pricing or property conditions. Knowing the DOM helps buyers and sellers evaluate the market.

The duration required to sell a property may vary significantly according to several variables, including location, pricing strategy, and market circumstances. Typically, residents who get offers soon after listing are seen to be doing well. The prolonged unsold status of a property may compel sellers to reevaluate their price or marketing approaches. A prolonged time on the market can lead to negative perceptions among potential buyers, who might question why a home hasn't sold—possibly viewing it as overpriced or in poor condition. Understanding these dynamics is essential for sellers, as timely adjustments can enhance the chances of a successful sale. DOM serves as an essential indicator of market health. In a seller's market, characterized by high demand and limited supply, homes typically sell quickly with low DOM figures. Conversely, in a buyer's market, where inventory exceeds demand, homes may linger on the market longer. This fluctuation in DOM provides insights into broader economic conditions and helps real estate professionals gauge market trends.

Knowing the duration of real estate sales and its effective factors will lead to a conscious decision. It could help real estate planners analyze and predict future conditions and choose suitable solutions. Through such information and a clear vision, people can decide whether to sell the residential and commercial units or invest in the real estate market in the coming months. In recent years, researchers have tried to use machine learning methods in different fields, such as predicting real estate and property issues. Applying machine learning to predict the time to sell a home allows for a data-driven approach that enhances decision-making. By analyzing historical sales data, machine learning algorithms can identify patterns and trends influencing sales velocity. This predictive capability helps stakeholders understand when to sell residential or commercial properties, optimizing their investment strategies based on anticipated market conditions. This research uses the three-stage data mining model to predict the duration of residential units' sales based on available variables. In the first stage, data were cleansed; in the second stage, the two-stage weighting approach was proposed to determine the effective features in predicting the residential units' sale duration. In the third stage, the optimum sets of features are entered to the data mining classification algorithm. The accuracy metric was used to evaluate the model. The proposed model in this research helps real estate consultants effectively manage the duration of real estate sales.

This article substantially enhances the real estate sector by using contemporary data mining techniques to predict the duration of residential unit sales. The study utilizes a three-phase data mining methodology, including data cleansing, feature selection, and classification algorithms to proficiently address the difficulties of predicting sales durations in an unpredictable industry. Using a feature selection method facilitates the identification of critical parameters affecting sale length, improving forecast accuracy and enabling stakeholders to refine their investment plans according to projected market circumstances. This methodological innovation enhances predictive capacities and offers practical insights for real estate professionals, facilitating informed choices about pricing and marketing.

Despite advances in real estate price prediction models, a large research vacuum remains in understanding the specific factors that influence the duration of property transactions. Previous studies have mostly focused on price forecasting, ignoring the impact of other factors on sales durations. This article highlights the importance of Days on the Market (DOM) as a crucial metric for assessing market liquidity and property attractiveness. Furthermore, it emphasizes the need to continually improve predictive models via machine learning techniques that can adapt to changing market dynamics. This research fills a gap by advancing theoretical frameworks in real estate analytics and offering practical solutions for developers and investors operating within a complex market environment. Therefore the primary objectives of this research are delineated as follows:

- To develop a predictive model for estimating the Days on Market (DOM) for residential properties through the application of advanced machine learning techniques.
- To identify and analyze critical factors that significantly influence the duration of residential unit sales, thereby enhancing the understanding of underlying market dynamics.
- To conduct a comparative analysis of the performance of various classification algorithms, including K-Nearest Neighbors, Naïve Bayes, and Decision Trees, in accurately predicting DOM.
- To provide actionable insights for real estate professionals that can inform and optimize their investment strategies and marketing approaches based on anticipated sales durations.

The paper is structured as follows: In Section 2, a literature review on the subject of real estate is presented. In Section 3 Materials and the proposed method is presented. In Section 4, experimental results are represented. In Section 5 conclusions are discussed.

2. Literature Review

In practically every country, the real estate industry has been important in driving economic progress. A vast amount of studies have been undertaken in the real estate field. Each of these studies looks at different aspects of real estate marketing [4]. Residential property value forecasting is a significant topic in the real estate industry. Many data mining approaches were used, including artificial neural networks (ANN)[5], Particle swarm optimization (PSO) using support vector machines (SVM) [6], Genetic algorithms and support vector machines [7], Decision Tree (DT)[8]. Furthermore, text mining has been used with linear regression to improve real estate price predictions [9]. Forecasting and optimizing housing market fluctuations [10] and anticipating residential construction demand [11] have also been investigated in other studies.

Numerous previous studies focused on the appraisal of real estate. A mass real estate appraisal is often used to compute property taxes and determine market value. Because of its use in marketing forecasting and sales analysis, the linear regression model was the most often used for mass evaluation. Artificial neural networks and support vector machines are further categorization approaches in real estate evaluation.[12,13,14]

In their 2023 project, Zhan et al. [15] created an advanced machine learning method for projecting residential property values. This method aims to increase the accuracy and reliability of real estate prices by combining multiple prediction algorithms. The authors use a hybrid technique that combines traditional hedonic pricing models with modern machine-learning methods such as support vector regression and decision trees to better understand the complex dynamics of housing data. The technique tries to increase prediction accuracy beyond traditional methods by using a diverse dataset, including several property qualities and market factors. The study shows that this hybrid method improves the quality of residential property evaluations and provides valuable information for real estate professionals, allowing them to make better investment decisions.

Foryś, 2022 [16] investigated the effectiveness of machine learning algorithms in projecting residential property values. This research compares the predictive capabilities of traditional regression models to neural network approaches to find which method produces more accurate price forecasts. This research uses a large dataset with information on many property attributes to assess the strengths and drawbacks of several modelling methodologies. Regression models are easier to understand and implement; nevertheless, research shows that neural networks are better at recognizing subtle associations in data, often leading to improved forecast accuracy. This article investigates the use of machine learning in the real estate industry, emphasizing the importance of predictive analytics in supporting informed decision-making. Yu et al. [17] investigated the use of data mining and machine learning to develop separate real estate pricing systems that improve the accuracy of property value. The authors examine several algorithms, such as regression models, decision trees, and neural networks, to determine their ability to estimate house prices reliably. This research shows that applying these methodologies to a broad dataset with various property attributes and market situations significantly increases price accuracy compared to standard valuation approaches. These findings show that powerful computing tools can be used to improve real estate pricing models, providing valuable information for investors, developers, and those who make policies, all of whom need to make smart choices in a constantly changing market. In their 2020 study, García-Magariño et al. [18] and their team employed machine learning and dimensionality reduction algorithms to fill in missing price data in real estate market simulations. The authors propose a new agent-based simulation framework that tackles the problem of incomplete price data. This framework utilizes advanced algorithms that precisely predict missing property values, highlighting their practical use in real-world scenarios. The findings suggest that improving the quality of data used for real estate modelling could offer a path toward more informed and reliable market analysis. Singh et al.[19] Develop into applying big data analytics to make predictions about real estate prices. The authors delve into various analytical tools and models that leverage massive datasets to enhance the precision of property valuations. The study highlights how big data can help real estate professionals understand the market better by analyzing geographical patterns, specific property features, and changing market trends. The research shows that using big data analytics in real estate pricing can make decisions more informed, boosting investment returns and making the market work more smoothly. In their 2020 paper [20], Ma et al. proposed a deep forest model specifically designed for real estate price prediction. This model effectively addresses the imbalance in real estate datasets by considering the costs associated with different predictions. To improve prediction accuracy and reduce the costs associated with misclassifications, the authors present a hybrid approach that leverages the strengths of both deep learning and ensemble methods. Integrating cost-sensitive learning within the deep forest architecture allows the model to focus on critical instances, like undervalued properties, thereby producing more precise price predictions. These results show that this special technique boosts forecasting accuracy and gives useful information to real estate market players, leading to more informed decisions about property investments.

Kamara et al. [21] proposed a sophisticated hybrid neural network model that accurately forecasts Days on Market (DOM), a vital measure of the ease with which real estate properties are bought and sold. The authors are employing a variety of neural network architectures to handle the complexities and non-linear relationships within real estate data, ultimately leading to more accurate predictions. This study utilizes a vast collection of data, including details about property features and market trends, to prove the algorithm's ability to offer up-to-the-minute information about property sales patterns. By accurately projecting DOM, stakeholders gain valuable insights

into the real estate market, empowering them to develop competitive pricing strategies and effectively position their properties in a competitive landscape. In their 2019 study, Neloy and colleagues [22] combined multiple learning algorithms to create a model that significantly boosted prediction accuracy, demonstrating the superiority of the ensemble approach over traditional methods. The study provides a deeper understanding of rental pricing by examining general features and delivering actionable insights for those involved in the real estate market. In their 2019 study, Abidoye and colleagues [23] investigated the effectiveness of various statistical and machine learning models, including ARIMA, ANN, and SVM, for high accuracy in predicting the property price index (PPI).

3. Theoretical Background

3.1. Data Cleansing

Data cleaning is the process of quality control before the data analysis. One problem that reduces the data set's quality is the presence of missing values. Missing value imputation provides estimations for missing values by using a suitable learning algorithm, such as K-nearest neighbour. [24-25]

3.2. Feature selection

Feature subset selection (FSS) methods attempt to find the best subset of features from among the original dataset. The selected subset of input variables using FSS methods has more predictive power for an output variable. Among the goals of feature selection methods are reducing the dimension of the data set, eliminating redundant and irrelevant features, increasing the speed of operations, and increasing the accuracy of data mining algorithms. Filtering methods is one of the most common feature selection techniques [26].

- **Filter methods**: These methods apply several measures such as information, distance, dependence, consistency, similarity, and statistical measures to score features or feature subsets. Examples include chi-square, correlation, chi-square, Gini Index and Relief [27].

- **3.3. Simple Linear Regression model**

  Linear regression is one of the most popular statistics and machine learning algorithms. Regression is widely used for addressing the relations among a set of variables. In particular, the linear regression model describes the dependence of a response variable on a set of predictor variables.

  Assuming there are observations on n individuals $(i = 1,2, \dots, n)$ the simple linear regression model is:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + e_i} \quad (1)$$

Where β's are the regression coefficients and e_i is a random component that is assumed to be independently normally distributed with zero mean and variance σ^2.

3.4. Classification

- **K-nearest neighbor**

  The k-nearest neighbor method was first introduced in the early 1950s. This method performs learning based on similarity by comparing testing and training records. Test records are described using n features, and each record is a point in an N-dimensional space. To classify an unknown and unlabeled record, the K- nearest neighbour classification algorithm seeks for K training record, which is most similar to the unknown record. These K training records are the K nearest neighbors of the unknown record. The similarity of records can be defined by distance metrics such as Euclidean distance

Decision Tree

The decision tree is one of the most powerful and common tools for classification and prediction. This algorithm has a tree-like structure. The highest node in the tree is the root node. Leaf nodes represent classes or class distributions. Decision Tree Induction is the learning phase from training data with class labels. The decision tree produces the law. Decision trees are used in this way for Classification. For example, a record such as X is considered whose label is completely unknown and the values of its features are tested on the decision tree. According to the values of the features of X, a path is drawn from the root to a leaf node, the leaf node that predicts the class of the X record. One of the reasons for the popularity of decision trees is that they can easily be turned into classification rules.

- **Naïve Bayes**

The Naïve Bayes is one of the most famous classification algorithms in data mining. This algorithm calculates the conditional probabilities between the input variables and the target and determines which input properties are most likely to play a role in predicting the target variable. The basis of the Bayesian classification algorithm is the Bayesian theorem. Mathematically, Bayes theorem is stated as:

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}} \quad (2)$$

where A and B are two independent events. The probability of an event A and B is denoted by $P(A)$ and $P(B)$ respectively. $P(A|B)$ denotes the conditional probability of A given B. $P(B|A)$ is the probability of event Bconcerning event A

3.5. Evaluation Metrics

Evaluation metrics such as Accuracy, Precision, and Recall are used to evaluate data mining algorithms. “Confusion Matrix” is used to calculate these metrics. Table 1 shows the confusion matrix for a multi-class classification problem.

Table 1: Confusion Matrix

		Predicted		
		A_1	A_j	A_n
Actual	A_1	N_{11}	N_{1j}	N_{1n}
	A_j	N_{i1}	N_{ij}	N_{in}
	A_n	N_{n1}	N_{nj}	N_{nn}

The classification accuracy, precision and recall are presented, respectively, in equations (3), (4) and (5):

$$\boldsymbol{Accuracy = \frac{\sum_{i=1}^{n} N_{ii}}{\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}}} \quad (3)$$

$$\boldsymbol{Precision = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ki}}} \quad (4)$$

$$\boldsymbol{Recall = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ik}}} \quad (5)$$

Where, N_{ij} denotes the number of samples actually belonging to class Ai classified as class Aj, and, and N_{ii} denotes the number of samples actually belonging to class Ai classified as class Ai

4. Materials and methods

4.1. Dataset description

The dataset used in this article deals with the information of a real estate database in Iran which includes 6836 records. It consists of 24 predicting variables and 1 target variable, which the latter is about the duration of residential unit's sale. The predictive factors include a broad variety of property parameters such as price, location, age, and other features such as the number of bedrooms, parking availability (Parking), and heating and cooling system type. These factors are useful in determining how various qualities of residential units affect their market performance, specifically how fast they sell. The variables included in the study are shown in Table 2. The objective variable indicates the time required for a property to be listed and sold, a crucial indicator for market liquidity and property attractiveness. The examination of this data may provide stakeholders with critical insights into the determinants affecting transaction length, therefore enabling them to make better educated choices on investment prospects and pricing strategies in the Iranian real estate market.

Table 2: Research Variables

Row	Variables	Description	Type
1	Fee	The price of property	Numeric
2	Region	The area where the property is located	Nominal
3	Area	Property area	Numeric
4	Bed Room	Number of bedrooms per unit	Nominal
5	No Floor	The floor where the property is located	Nominal
6	All Floor	Total number of floors available in the apartment	Nominal
7	Unit	The unit number	Nominal
8	Age	The age of unit	Numeric
9	Point	The location of apartment	Nominal
10	Front	The facing of the apartment	Nominal
11	Heating	Type of heating system	Nominal
12	Cooling	Type of cooling system	Nominal
13	Floor	Type of materials used in floor	Nominal
14	Room Floor	Type of materials used in floor of rooms	Nominal
15	Cabin	Type of cabinet in each unit	Nominal
16	Parking	The unit has a parking or not?	Binominal: 1, 0
17	Remote	Parking door has remote control or not?	Binominal: 1, 0
18	Storage	Storage available in selling unit	Binominal: 1, 0
19	Elevator	Elevator available in apartment	Binominal: 1, 0
20	Terrace	Terrace available in selling unit	Binominal: 1, 0
21	Toilet	Toilet Conditions	Binominal: 1, 0
22	Kitchen	Type of kitchen available in unit	Binominal: 1, 0
23	IPhone	Available video door-phone	Binominal: 1, 0
24	Antenna	Central antenna	Binominal: 1, 0
25	Duration1	*Target Variable:* The time taken from ad registration to the sale of the desired property	Nominal

4.2. Proposed Model

This study employs a comprehensive three-stage data mining model to predict the duration of residential unit sales effectively. Figure 1 shows the main stages of the research. Firstly, it focuses on data cleansing. In the second stage, a two-stage weighting method is proposed to select the effective features for predicting the duration of residential units' sales. In the third stage, the selected variables from the last stage are entered into the classification algorithm. Here, the Accuracy metric has been used to compare the classification algorithms and choose the algorithm with the best performance. The following provides details of the research stages.

4.2.1. Data Cleansing

The initial stage involves thorough data cleansing to ensure the reliability and quality of the dataset. This includes:

- Handling Missing Values: We identified missing values and applied imputation techniques using Rapid Miner. For numerical features, mean or median imputation was utilized based on the distribution of the data, while categorical variables were addressed using mode imputation or predictive modeling techniques to estimate missing entries.

4.2.2. Feature Selection:

In this stage, we employed a two-stage weighting approach to identify and select the most significant features influencing the duration of residential unit sales:

- Filter Weighting Method: Initially, we applied filter methods to rank features based on their correlation with the target variable. This step helped in identifying features that have a strong relationship with sales duration.
- Regression Techniques: Following the filter method, regression techniques were utilized to refine feature selection further. This involved assessing the contribution of each feature in predicting DOM and eliminating those that did not significantly enhance model performance.

4.2.3. Classification Algorithms:

The final stage involves applying various classification algorithms to predict DOM based on the selected features:

- Algorithm Selection: We compared multiple machine learning algorithms, including K-nearest Neighbor (KNN), Naïve Bayes, and Decision Trees. Each algorithm was chosen for its unique strengths in handling different types of data distributions and relationships.
- Model Evaluation: The performance of each model was evaluated using accuracy metrics, including accuracy, precision, recall, and F1-score. This evaluation process enabled us to identify the most effective algorithm for predicting sales duration.

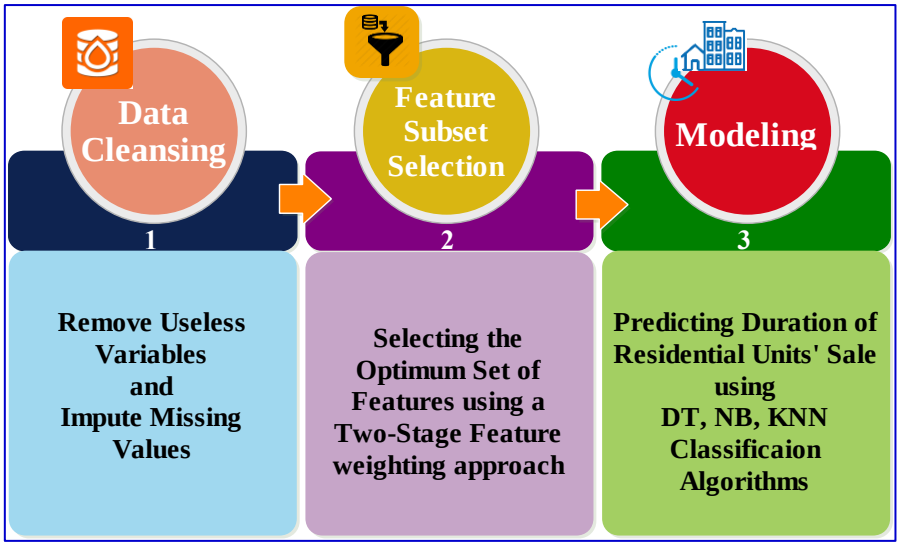


Figure 1: Research Process

5. Experimental Results

5.1. First Step: Data Cleansing

In initial reviews of the dataset, missing values were handled by imputation, and the value of missing data was estimated using a suitable supervised learning algorithm, i.e., the K-nearest neighbor algorithm.

5.2. Second Step: Two-stage weighting approach

In this section, a two-stage approach is provided to weighting the variables and also determining the relative importance of them. Figure 2 shows the stages of weighting to variables in research by using the proposed approach.

This paper presents a two-stage weighing method for analyzing and estimating the proportional influence of many factors affecting the length of sales of residential units. The 24 current features are weighted to determine their relative importance. Figure 2 shows the stages of weighting variables in research using four filtering methods: Chi-square, Correlation, Relief, and Gini Index. Every method uses many statistical ideas to evaluate the characteristics' value, producing weights between zero and one. This first analysis helps to identify the most important elements influencing the projection of sales duration.

In the second step, a regression model is developed using the average weights of the filtering techniques as response variables. The model seeks to improve the accuracy of feature weight estimations by using filtering methods.

This two-stage weighting approach seeks to find which filtering strategy is more crucial in the second step of feature selection. Using a range of filtering techniques—Chi-square, Correlation, Relief, and the Gini Index—the strategy evaluates and contrasts each method's efficacy in ascertaining the relative relevance of the variables. This systematic analysis helps to choose traits that significantly affect the length of residential unit sales, thus improving the model's prediction accuracy. Identifying and stressing these important traits will help stakeholders to grasp better the elements that significantly affect sales duration, thereby enabling more informed choices on marketing plans and pricing in the real estate industry. This all-encompassing approach enhances the forecasting capacity of the model and offers insightful analysis on the efficient handling of real estate transaction complications.

5.2.1. Weighting features by filtering methods

Four filtering methods—Chi-square, correlation, relief, and Gini Index — were applied to assign weights between zero and one to 24 available features. Table 3 shows the weights assigned to the 24 features using filtering methods.

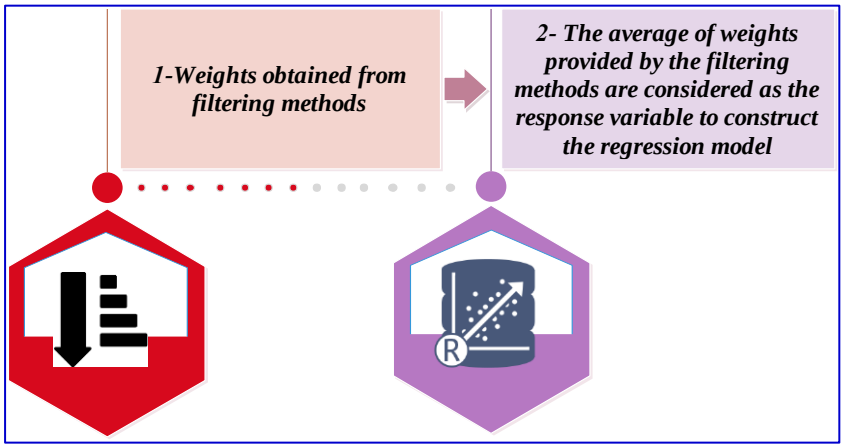


Figure2: Feature weighting Process

Table 3: Weights assigned to variables by filtering methods

Features	Chi-square	Gini	Correlation	Relief	Average of weights provided by filtering methods
Fee	0.000	0.031	0.140	0.000	0.043
Cooling	0.011	0.000	0.085	0.002	0.024
Toilet	0.015	0.005	0.219	0.077	0.079
Area	0.016	0.012	0.086	0.002	0.029
Point	0.018	0.012	0.000	0.008	0.009
Terrace	0.040	0.025	0.270	0.000	0.084
Cabin	0.042	0.037	0.003	0.007	0.022
Antenna	0.045	0.038	0.186	0.065	0.083
Front	0.046	0.035	0.147	0.008	0.059
Bed Room	0.053	0.042	0.116	0.029	0.060
IPhone	0.059	0.044	0.263	0.108	0.119
Kitchen	0.084	0.063	0.362	0.091	0.150
Region	0.086	0.057	0.185	0.118	0.112
Age	0.095	0.077	0.115	0.055	0.086
No Floor	0.104	0.078	0.020	0.072	0.068
Heating	0.122	0.115	0.061	0.016	0.078
Remote	0.125	0.083	0.521	0.193	0.230
Storage	0.157	0.099	0.600	0.180	0.259

Unit	0.201	0.161	0.359	0.109	0.207
Room Floor	0.330	0.316	0.130	0.343	0.280
Floor	0.359	0.337	0.229	0.342	0.317
Elevator	0.522	0.377	1.000	0.472	0.593
All Floor	0.526	0.464	0.356	0.269	0.404
Parking	1.000	1.000	0.035	1.000	0.759

The weights given to various factors influencing the length of time it takes for residential unit sales using the four filtering methods of chi-square, Gini index, correlation, and relief are shown in Table 3. Each characteristic's significance in predicting the target variable is represented by a weight ranging from 0 to 1. Parking is the most important component, with an average weight of 0.759, highlighting its substantial impact on customer choices and sales duration. Two noteworthy features that illustrate how facilities affect a property's attractiveness are the elevator (0.593) and the whole floor (0.404). The results highlight the different viewpoints on the importance of characteristics by exposing heterogeneity in the weights assigned by several filtering methods. For instance, if Remote's Chi-square (0.125) and Gini Index (0.083) scores are rather low, the Relief technique gives it a significant weight of 0.521. Conversely, factors with low average weights, like Point (0.009) and Fee (0.043), seem to have little effect on sales length.

We will proceed to the next stage of our investigation to ascertain which filtering strategy is most likely to affect feature selection. This involves the use of many filters to compare the weights allocated to each attribute. Examining these weights facilitates the discovery of strategies that consistently highlight the most crucial elements affecting the length of time that residential unit sales occur. Our feature selection for further modelling stages will be influenced by this research, which will assist us in determining the value of each filtering technique. By concentrating on the important elements found during comparison analysis, this step attempts to raise the model's expected accuracy.

5.2.2. Construct the regression model

This section describes the construction of the regression model. Provided weights by each filtering method are considered the predicting variable (See Table 4; columns 2 to 5), and an average of these weights is assumed to be the response variable (See Table 4; last column). The response variable is expressed as a combination of the predicting variables in the form described in Eq. (1).

Then, the mentioned variables are entered into a linear regression model to access a better prediction of the weights of variables. As a result, the regression model is obtained as Eq. (6)

$$\boldsymbol{Y = \beta_0 + \beta_1 Chisquare + \beta_2 Correlation + \beta_3 Relief + \beta_4 GiniIndex + e_i} \quad (6)$$

where, Y denotes the predicted response variable, while β_0 represents the intercept and β_1 through β_4 are the coefficients estimated for each predictor. In this study, SPSS software is used to estimate the coefficients β of the linear regression equation shown in Eq. (6). SPSS is widely used for logical batched and non-batched statistical analysis. It provides a fast and simple way to construct the linear regression model. The results of the regression model are shown in Table 4. A more reliable way to evaluate the quality of the model fit is by applying a variance analysis (ANOVA). The ANOVA results are shown in the last row of Table 4.

Table 4: Result of regression model and ANOVA

Term	Standardized Coefficients
T_1=Relief	a_1 = 0.295
T_2=Correlation	a_2=0.302
T_3=ChiSquare	a_3 =0.304
T_4=Gini Index	a_4 =0.302
Sum of squares regression= 0.812; Sum of squares residual= 0.000; Sig.000[a] a. Predictors: (Constant), Gini Index, Correlation, Relief, Chi-square Dependent Variable: Y	

Each method's standardized coefficients, which show their respective contributions to the model, are shown in Table 4. Notably, Relief has a coefficient of 0.295, although the values for Chi-square (0.304), Gini Index (0.302), and Correlation (0.302) are all rather similar in value. This resemblance implies that the four approaches have equal effects on feature weighting. Accordingly, the regression model for estimating weights of features is expressed as follows:

$$\boldsymbol{y = a_1 T_1 + a_2 T_2 + a_3 T_3 + a_4 T_4} \quad (7)$$

Table 6 shows the anticipated weights from the regression model for 24 characteristics that influence the length of residential unit sales. Higher weight values suggest a greater effect, with each weight representing the factor's importance in anticipating property selling speed. "Parking" appears as the most influential feature, with a weight of 0.913, indicating that access to parking is highly appreciated by prospective purchasers, making houses with parking facilities more likely to be sold quickly.

The "Elevator" variable follows closely, with a weight of 0.714, emphasizing the significance of elevator availability, particularly in multi-story buildings where ease of access is a top concern for purchasers.

The "All Floor" variable, which ranks third with a weight of 0.487, highlights the significance of accessibility and convenience in residential constructions. Additionally, "Floor" (0.381) and "Room Floor" (0.337) have considerable relevance, indicating that floor quality and layout impact customer choices.

On the opposite end of the spectrum, the "Point" variable has a projected weight of 0.011, suggesting a negligible influence on sale time. This shows that physical property amenities, such as parking and elevators, have a greater influence on buyer behavior than location-specific characteristics. Similarly, factors like "Area" (0.035) and "Fee" (0.052) have low weights, suggesting that, although important to buyers, they may not be major drivers in affecting the pace of property transactions.

Table 6 presents an instructive analysis of the factors that most influence the length of residential unit sales. Real estate agents should prioritize these aspects in pricing and marketing strategies by identifying crucial features like elevator access and parking availability. This study allows stakeholders to understand market dynamics better and make educated choices based on client preferences, resulting in improved sales results in a competitive real estate market.

The line chart in Figure 3 visually compares the provided weights by the filtering methods and estimated weights by the regression model.

Table 5: Estimated weights for the variables by regression model

Variable	Estimated Weight	Rank	Variable	Estimated Weight	Rank
Parking	0.913	*1*	Terrace	0.101	*13*
Elevator	0.714	*2*	Antenna	0.100	*14*
All Floor	0.487	*3*	Toilet	0.095	*15*
Floor	0.381	*4*	Heating	0.095	*16*
Room Floor	0.337	*5*	No Floor	0.082	*17*
Storage	0.312	*6*	Bed Room	0.073	*18*
Remote	0.277	*7*	Front	0.071	*19*
Unit	0.250	*8*	Fee	0.052	*20*
Kitchen	0.181	*9*	Area	0.035	*21*
IPhone	0.143	*10*	Cooling	0.029	*22*
Region	0.134	*11*	Cabin	0.027	*23*
Age	0.103	*12*	Point	0.011	*24*

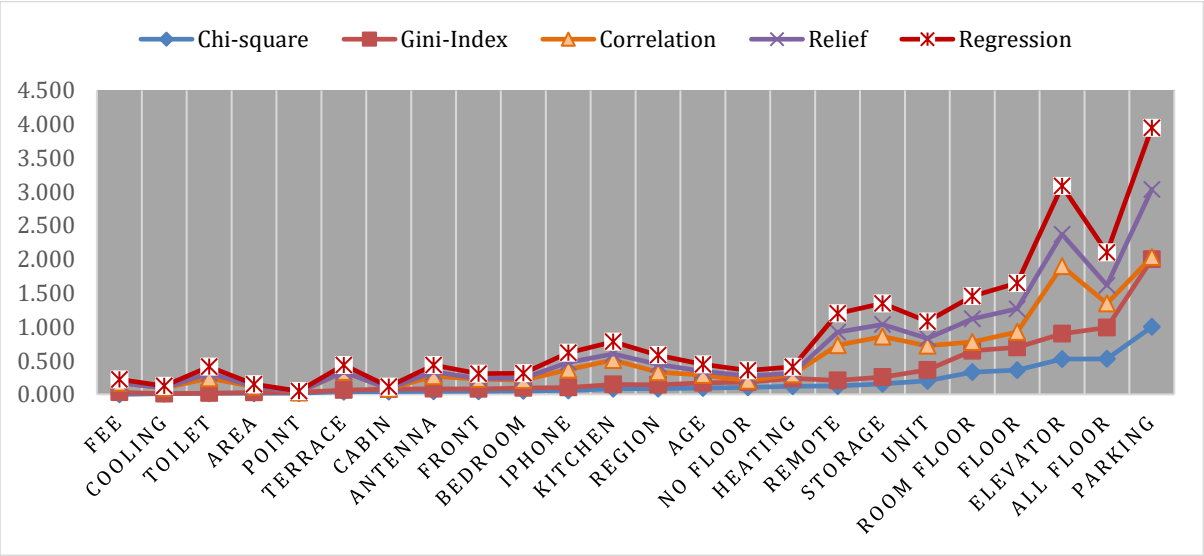


Figure 3: Comparison of the provided weights by the filtering methods and estimated weights by regression model

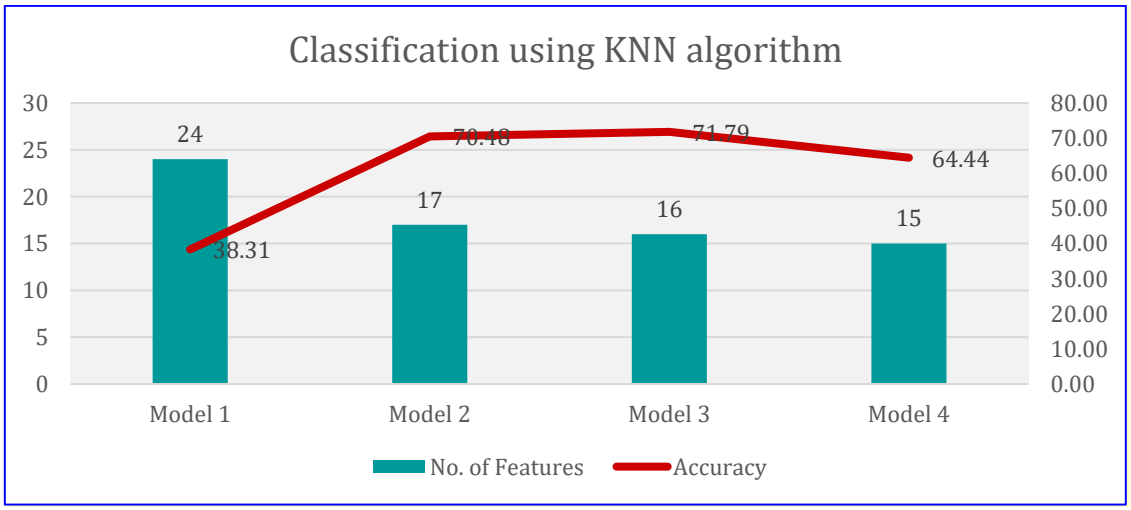
5.3. Third Step: Modeling

In this stage, we use three machine learning methods to estimate the length of residential unit sales: K-Nearest Neighbor (KNN), Decision Tree (DT), and Naïve Bayes (NB). To assess the performance of the classification models, 70% of the data was utilized for training and 30% for testing. This split validation method ensures that the model is evaluated on a separate subset of data that it has not encountered during training, providing a robust measure of its generalizability and performance. The training set is used to fit the model, allowing it to learn patterns and relationships within the data. In contrast, the testing set serves as an independent benchmark to evaluate how well the model performs when presented with data it has not previously seen. This separation helps mitigate the risk of over fitting, where a model may perform exceptionally well on training data but poorly on new data due to having

learned noise or specific details rather than underlying trends. The modeling technique was carried out on both the original and refined datasets, including the variables with the highest estimated weights produced from the FSS approach. Common machine learning measures such as accuracy, precision, and recall were utilized to assess each algorithm's classification performance. Table 6 shows the results of this modeling, which includes a comparative examination of the performance metrics for each method over a range of feature counts.

Table 6: The results of classification

Model	Algorithms	No. of Features	Accuracy	Precision	Recall	F1-Score
Model1	**KNN**	all (24)	38.31%	38.20%	38.10%	38.15%
Model2		17	70.48%	70.14%	71.25%	70.69%
Model3		16	**71.79%**	**71.39%**	**72.74%**	**72.06%**
Model4		15	64.44%	64.02%	66.33%	65.15%
Model5	**NB**	all (24)	59.32%	59.55%	61.56%	60.54%
Model6		17	60.58%	60.18%	61.28%	60.73%
Model7		16	61.93%	61.92%	62.24%	62.08%
Model8		15	62.37%	62.36%	62.80%	62.58%
Model9	**DT**	all (24)	59.13%	59.46%	63.06%	61.21%
Model10		17	65.55%	65.45%	66.35%	65.90%
Model11		16	70.19%	70.18%	71.30%	70.74%
Model12		15	67.78%	67.80%	67.89%	67.84%



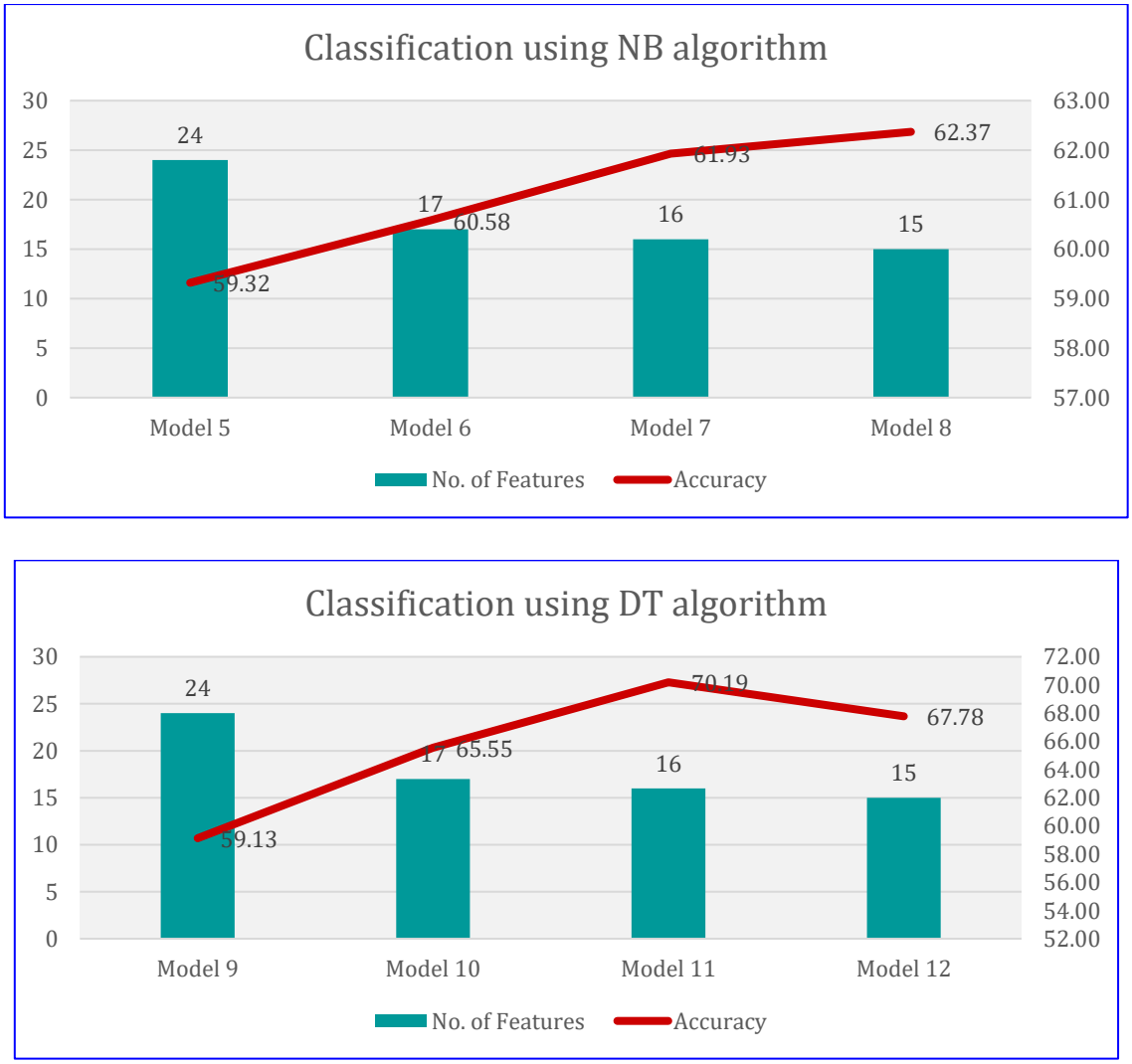


Figure 4: Classification results based on FS technique and the number of selected features for each of the applied models.

The findings reveal that the performance of various models is significantly influenced by the number of features utilized. Model 1, which employs all 24 features using the KNN algorithm, shows an accuracy of only 38.31%, indicating poor performance. However, when the number of features is reduced, particularly in Model 2 (17 features) and Model 3 (16 features), accuracy improves dramatically to 70.48% and 71.79%, respectively. This suggests that prioritizing the most relevant variables enhances the model's predictive power. Model 4, which uses 15 features, achieves a reasonable accuracy of 64.44%, but this is lower than that of Models 2 and 3. This indicates that there may be an optimal number of features for KNN, where reducing dimensionality can yield better performance without sacrificing too much information. Similarly, Naïve Bayes models exhibit a comparable trend. Model 5, which utilizes all 24 features, has an accuracy of 59.32%. As the number of features is reduced to 16 in Model 6 and further to 15 in Model 7, accuracy increases slightly to 60.58% and 62.37%, respectively. This implies that while feature selection does provide some benefits for Naïve Bayes, the improvement is not as pronounced as with KNN. The results for Decision Tree models also reflect this pattern. Model 9, which uses all 24

features, achieves an accuracy of 59.13%, while reducing the number of features to 17 in Model 10 results in a notable improvement to 65.55%. The accuracy remains relatively high at 67.78% with just 15 features in Model 12, indicating that reducing the number of features can enhance performance in Decision Trees as well.

Overall, the findings emphasize the relevance of feature selection in improving model performance across many techniques. When employing subsets of high-weight variables rather than all available features, both KNN and Decision Tree models perform much better. The Naïve Bayes model shows moderate advantages from feature reduction, but still benefits from concentrating on important variables. These results highlight the usefulness of focused feature selection in improving prediction accuracy and show that using a narrower collection of relevant factors might result in more accurate outcomes when forecasting residential unit sales duration.

5.4. Key Findings and Practical Implications

This section highlights the key findings of the research and their practical and theoretical applications within the housing market. By examining factors such as location, pricing strategies, market conditions, and the use of machine learning techniques, the study provides actionable insights for real estate professionals while contributing to the broader understanding of market dynamics.

5.4.1. Impact of Location on DOM

The results show that location significantly influences the DOM for residential properties. Properties in desirable neighborhoods tend to sell faster than those in less sought-after areas. Real estate agents can use this insight to guide clients on where to invest or list properties. For instance, agents might recommend properties in high-demand areas to sellers aiming for a quick sale.

5.4.2. Pricing Strategies and Their Influence on Sales Duration

The results of the research indicates that effective pricing strategies are crucial for reducing DOM. Homes priced competitively relative to market conditions tend to sell more quickly. Sellers can utilize predictive models to determine optimal pricing strategies based on current market data, potentially leading to faster sales and higher profits. This can be particularly beneficial in fluctuating markets where timely adjustments are necessary.

5.4.3. Market Conditions and Their Role in DOM Variability

The study highlights that broader market conditions, such as supply and demand dynamics, significantly affect DOM. In a seller's market, properties typically sell faster compared to a buyer's market. Investors can leverage this knowledge to time their purchases or sales strategically, ensuring they capitalize on favorable market conditions. For example, they may choose to sell during peak demand periods when DOM is low.

5.4.4. Machine Learning Techniques for Predictive Analytics

The application of machine learning algorithms, such as K-nearest Neighbor and Decision Trees, improved the accuracy of predicting DOM based on various influencing factors. Real estate professionals can implement these machine learning models to enhance their forecasting capabilities, allowing them to make data-driven decisions regarding property listings and marketing strategies.

6. Conclusion

Real estate is an important indicator of economic development. Modern buildings are signs of economic progress. The decision to buy a house is often the most important financial decision in any

family, and it requires the allocation of enough time and effort to evaluate all available choices and find the appropriate criteria to analyze the choices and select the best one.

Having knowledge about duration of real estates' sale and knowing the effective factors on it can help both real estate agents and buyers to make right investment decisions. Through such information and having a clear vision about the duration of real estate sales, people can make the best decision whether to sell residential and commercial units or invest in the real estate market in the coming months. In recent years, data mining techniques have been used widely for the duration of real estate sales. Knowing the effective factors can help real estate agents and buyers solve various real estate issues, including forecasting sales prices and market fluctuations.

This study significantly enhances our understanding of the variables that influence the duration of residential unit sales by employing a robust three-stage data mining methodology. The research meticulously cleans data, selects useful features, and employs multiple classification algorithms to identify significant factors influencing sales timeframes. The findings indicate that the prediction accuracy of the models is enhanced when a restricted subset of characteristics, particularly those with the highest estimated weights, is prioritized over the use of all potential variables. This customized approach not only optimizes model performance but also furnishes real estate professionals with critical information, enabling them to make more informed decisions regarding pricing and marketing.

The insights derived from the proposed predictive model offer practical applications for various stakeholders in the real estate market. For instance, real estate investors can strategically determine the optimal timing for buying or selling properties by analyzing expected Days on Market (DOM). By understanding these predictions, investors can align their transactions with periods of high demand, thereby maximizing their potential returns. Similarly, real estate agents can utilize the findings to establish competitive pricing strategies. If the model forecasts a low DOM for specific neighborhoods, agents might price properties slightly above market value to take advantage of heightened buyer interest.

Moreover, our research informs targeted marketing strategies by identifying key property features that correlate with shorter DOM. Agents can emphasize these desirable attributes in their marketing materials to attract potential buyers more effectively. Sellers can also benefit from the insights provided by our study; if the model indicates that poorly maintained homes typically experience longer DOM, they may opt to invest in repairs or renovations prior to listing their properties. Lastly, real estate professionals can leverage our findings for broader market trend analysis. By monitoring fluctuations in DOM across various regions or property types, they can make informed decisions about where to concentrate their efforts and resources, ultimately enhancing their competitive edge in the market.

This research has certain limitations that should be acknowledged despite its progress. The dataset's specificity may limit the results' applicability to other locations or markets with characteristics distinct from those of the Iranian real estate market. Although the dataset utilized for analysis is extensive, it may not encompass all relevant variables that could influence the duration of residential unit sales. For example, external factors such as economic fluctuations, changes in real estate regulations, and regional market dynamics may not be fully captured within the dataset. This limitation could impact the generalizability of our results to other markets or contexts that extend beyond those represented in our data. Additionally, the study concentrates on a specific subset of machine learning algorithms; however, other sophisticated methodologies, such as deep learning or ensemble methods, may result in even more advanced prediction accuracy. The interpretation and application of the findings could be enhanced by further investigating these

approaches and expanding the dataset to encompass various geographical regions and market scenarios. Additionally, this research fills a substantial void in the existing literature by emphasizing the importance of Days on Market (DOM) as a critical parameter for assessing the attractiveness of properties and the market's liquidity. Continuous research is necessary to enhance prediction models and adapt to changing market conditions as the real estate sector develops. Future research could investigate additional variables that influence sales duration, such as economic indicators or consumer demographics, to offer a more comprehensive understanding of market behaviour. The results of this study have the potential to assist stakeholders in more effectively navigating an increasingly complex market environment, which could lead to improved methods for administering property transactions and better investment results.

Author Contributions

Author Contributions:": resources, methodology ,writing—original draft preparation :**Farshid Abdi**; data curation, validation,formal,writing—review and editing, **Shaghayegh Abolmakarem**; supervision, project administratio;:**Amir Karbassi Yazdi**

Funding

There is no fund

Data Availability Statement

The data is available upon request

Conflicts of Interest

There is no conflict

[1] Liu, X., Ornelas, E., & Shi, H. (2022). The trade impact of the COVID-19 pandemic. *The World Economy*, *45*(12), 3751–3779. https://doi.org/10.1111/twec.13279

[2] Abate, G. (2024). *Real Assets* (pp. 495–523). https://doi.org/10.1007/978-3-031-59819-7_16

[3] Schwingshackl, C., Sillmann, J., Vicedo-Cabrera, A. M., Sandstad, M., & Aunan, K. (2021). Heat Stress Indicators in CMIP6: Estimating Future Trends and Exceedances of Impact-Relevant Thresholds. *Earth's Future*, *9*(3). https://doi.org/10.1029/2020EF001885

[4] Zhang, J., Lyu, Y., Li, Y., & Geng, Y. (2022). Digital economy: An innovation driving factor for low-carbon development. *Environmental Impact Assessment Review*, *96*, 106821. https://doi.org/10.1016/j.eiar.2022.106821

[5] Yasnitsky, L. N., Yasnitsky, V. L., & Alekseev, A. O. (2021). The Complex Neural Network Model for Mass Appraisal and Scenario Forecasting of the Urban Real Estate Market Value That Adapts Itself to Space and Time. *Complexity*, *2021*(1). https://doi.org/10.1155/2021/5392170

[7] Pai, P.-F., & Wang, W.-C. (2020). Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences*, *10*(17), 5832. https://doi.org/10.3390/app10175832

[8] Rajesh, B., Sai Vardhan, M. V., & Sujihelen, L. (2020). Leaf Disease Detection and Classification by Decision Tree. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 705–708. https://doi.org/10.1109/ICOEI48184.2020.9142988

[6] Behera, M. P., Sarangi, A., Mishra, D., & Sarangi, S. K. (2023). A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine. *Procedia Computer Science*, *218*, 818–827. https://doi.org/10.1016/j.procs.2023.01.062

[9] Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, 115019. https://doi.org/10.1016/j.eswa.2021.115019

[10] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House Price Prediction using Random Forest Machine Learning Technique. *Procedia Computer Science*, *199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

[11] Sammour, F., Alkailani, H., Sweis, G. J., Sweis, R. J., Maaitah, W., & Alashkar, A. (2024). Forecasting demand in the residential construction industry using machine learning algorithms in Jordan. *Construction Innovation*, *24*(5), 1228–1254. https://doi.org/10.1108/CI-10-2022-0279

[12] Carranza, J. P., Piumetto, M. A., Lucca, C. M., & Da Silva, E. (2022). Mass appraisal as affordable public policy: Open data and machine learning for mapping urban land values. *Land Use Policy*, *119*, 106211. https://doi.org/10.1016/j.landusepol.2022.106211

[13] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

[14] Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2021). Export sales forecasting using artificial intelligence. *Technological Forecasting and Social Change*, *163*, 120480. https://doi.org/10.1016/j.techfore.2020.120480

[15] Zhan, C., Liu, Y., Wu, Z., Zhao, M., & Chow, T. W. S. (2023). A hybrid machine learning framework for forecasting house price. *Expert Systems with Applications*, *233*, 120981. https://doi.org/10.1016/j.eswa.2023.120981

[16] Foryś, I. (2022). Machine learning in house price analysis: regression models versus neural networks. *Procedia Computer Science*, *207*, 435–445. https://doi.org/10.1016/j.procs.2022.09.078

[17] Yu, Y., Lu, J., Shen, D., & Chen, B. (2021). Research on real estate pricing methods based on data mining and machine learning. *Neural Computing and Applications*, *33*(9), 3925–3937. https://doi.org/10.1007/s00521-020-05469-3

[18] García-Magariño, I., Medrano, C., & Delgado, J. (2020). Estimation of missing prices in real-estate market agent-based simulations with machine learning and dimensionality reduction methods. *Neural Computing and Applications*, *32*(7), 2665–2682. https://doi.org/10.1007/s00521-018-3938-7

[19] Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*(S2), 208–219. https://doi.org/10.1007/s13198-020-00946-3

[20] Ma, C., Liu, Z., Cao, Z., Song, W., Zhang, J., & Zeng, W. (2020). Cost-sensitive deep forest for price prediction, *Pattern Recognition*, 107, 107499.

[21] Kamara, A.F., Chen, E., Liu, Q, & Pan, Z. (2020). A hybrid neural network for predicting Days on Market a measure of liquidity in real estate industry, *Knowledge-Based Systems*, 208, 106417

[22] Neloy, AA., Haque, HMS., & Ul Islam, MM. (2019). Ensemble learning based rental apartment price prediction model by categorical features factoring, in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pp. 350–356.

[23] Abidoye, RB., Chan, APC., Abidoye, FA., & Oshodi, OS. (2019). Predicting property price index using artificial intelligence techniques: Evidence from Hong Kong, International Journal of Housing Markets and Analysis, 12 (6), 1072-1092. https://doi.org/10.1108/ IJHMA-11-2018-0095

[24] Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing*, *73*, 816–828. https://doi.org/10.1016/j.asoc.2018.09.001

[25] Yazdi, A. K., Wang, Y. J., & Alirezaei, A. (2018). Analytical insights into firm performance: a fuzzy clustering approach for data envelopment analysis classification. *International Journal of Operational Research*, *33*(3), 413. https://doi.org/10.1504/IJOR.2018.095630

[26] F. Abdi, S. Abolmakarem, A. Karbassi Yazdi, Y. Tan and I. Andrés Marchioni Choque, "Prospective Portfolio Optimization With Asset Preselection Using a Combination of Long and Short Term Memory and Sharpe Ratio Maximization," in IEEE Access, vol. 12, pp. 144280-144294, 2024, doi: 10.1109/ACCESS.2024.3466829. keywords: {Portfolios;Optimization;Long short term memory;Predictive models;Investment;Forecasting;Data models;Stock markets;Asset pre-selection;long short-term memory;stock prediction;portfolio optimization;Sharpe ratio;cardinality},

[27] Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S., Moore, J.H., *Relief*-based *feature selection: Introduction and review*. (n.d.). *Journal of Biomedical Informatics*, 2018, 85, pp. 189–203.