Machine learning for real estate valuation: Astana, Kazakhstan case

Alibek Barlybayev[1,2], Arman Sankibayev[1], Rozamgul Niyazova[1,2], Gulnara Akimbekova[3]
[1]Department of Artificial Intelligence Technologies, Faculty of information technologies, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Higher School of Information Technology and Engineering, Astana International University, Astana, Kazakhstan
[3]Department of Microbiology and Virology named after. Sh.I. Sarbasova, Faculty of dentistry, Astana Medical University, Astana, Kazakhstan

Article Info

Article history:

Received Jan 18, 2024
Revised Mar 16, 2024
Accepted Apr 6, 2024

Keywords:

Machine learning
Neural network
Real estate valuation criteria
Real estate value forecasting
Regression learner

ABSTRACT

Purpose of this research is to investigate the accuracy of machine learning models in forecasting and evaluating house prices, and to understand the key factors that impact pricing. The study involved analyzing data scraped from real estate ads in the "sale of secondary housing" category on the website krisha.kz. The paper emphasizes the importance of understanding the factors that affect house prices, such as quality, location, size, and building materials. It was concluded that these factors have a strong correlation with house price prediction. The information available on krisha.kz was found to be a useful resource for finding good apartments. The data collected by the scraper was analyzed by models: Linear regression (LR), interactions linear regression (ILR), robust linear regression (RLR), fine tree regression (FTR), medium tree regression (MTR), coarse tree regression (CTR), linear support vector machine (LSVM), quadratic SVM (QSVM), medium gaussian SVM (MGSVM), rational quadratic gaussian process regression (RQGPR), boosted trees (BoosT), bagged trees (BagT), neural network based on the bayesian regularization algorithm (BR-BPNN). BR-BPNN showed better results than other models, with an MSE of 32.14 and R of 0.9899.

Corresponding Author:

Alibek Barlybayev
Department of Artificial Intelligence Technologies, Faculty of information technologies
L.N. Gumilyov Eurasian National University
Astana, Kazakhstan
Email: frank-ab@mail.ru

1. INTRODUCTION

Over the years, house price prediction has become a key research topic, as the demand for houses continues to sky-rocket. It is essential to develop a suitable framework that allows both buyers and sellers to make quick decisions when it comes to purchasing or selling property. Homebuyers want to be provided with comprehensive information in order to make decisions, but the volatile prices and high demand of real estate can make it difficult for them. Real estate valuations are often required for various investment purposes, such as providing collateral for secured loans from mortgage lenders. Such appraisals can provide valuable insights into the true worth of a property. Valuation reports are used for a variety of reasons like obtaining insurance premiums, determining rents, and assessing sales/purchase prices. Subsequently, it's essential to get reliable methods, which can accurate valuation figures since a wrong investment valuation could be devastating for investors that need financial information. Predicting house prices with a predictive analytics system can offer invaluable assistance to a variety of stakeholders. Lahmiri *et al.* [1] investigates the effectiveness of various machine learning algorithms in predicting housing prices. It found that some algorithms were notably more effective than others in this context. A significant discovery of the study is the role of Bayesian optimization in boosting the performance of these models. By fine-tuning the algorithms' hyperparameters, the predictive accuracy of the models was enhanced, underscoring the critical importance of algorithm selection and hyperparameter optimization in practical applications. Different technologies will help stakeholders plan and make decisions more efficiently. The house market has grown exponentially in recent years, creating a high level of demand. To address this, a new method of operation is necessary that can help to reduce the pressure. Given how much capital is involved in real estate investments, there is a pressing need to increase the accuracy of price predictions for property valuations. This is where machine learning techniques come into play and can be a great help [2]–[4]. Baur *et al.* [2] showcases the application of machine learning models to real estate valuation, highlighting their accuracy, comparisons with conventional valuation methods, and insights derived from the models' predictions. The results section outlines how these models perform in assessing property values based on descriptions, using various accuracy metrics for evaluation. In the discussion, the authors delve into the significance of their findings, placing them within the wider context of real estate valuation. They examine the strengths and limitations of their machine learning approach, consider its practical implications, and suggest potential areas for further investigation. Chou *et al.* [3] reveals that in the realm of real estate price prediction, neural networks excel above other models, primarily due to their superior capacity for modeling complex interdependencies between variables and outcomes. Random forests (RF) also emerge as strong contenders, offering competitive accuracy while providing the added benefit of greater interpretability compared to neural networks. Conversely, simpler models like decision trees (DT) and linear regression (LR) lag in performance, underscoring the importance of capturing non-linear dynamics for precise price forecasting. The scientific work [4] concludes that machine learning methods, notably RF and gradient boosting, surpass spatial interpolation techniques in predicting house prices. These machine learning approaches yield more precise forecasts and exhibit stronger resilience against dataset variations, marking a significant advancement in the accuracy and reliability of real estate price prediction. Though there are a few experts who may not be in favor of implementing new valuation models, as they believe it cannot replace the well-established conventional approaches, it is important to acknowledge non-traditional methods [5]. Sibindi *et al.* [5], the authors detail the performance of their hybrid model in predicting house prices, comparing its efficacy against both traditional and other cutting-edge models. They utilize a range of visual aids like tables and graphs, alongside statistical analyses, to illustrate their model's predictive accuracy and reliability. Even though this research doesn't suggest that the conventional method should be solely replaced with machine learning, it would still be beneficial to have an accompanying technique as a reference point for the traditional approach.

Various techniques have been proposed in literature to estimate residential real estate's market value, which can be divided into three categories: classical, statistical and advanced. Classical methods generally rely on the comparison principle for calculation. Statistical approaches enable us to establish a mathematical link between input variables and output variables, while the more advanced valuation techniques rely heavily on computer technologies [6]–[8]. Classical methods are divided into three groups such as comparison, revenue and cost method [6]. Classical methods of appraising real estate are the most widely used around the globe due to its ease of implementation. Nonetheless, they can only be applied single-handedly when there is one real estate, making them insufficient when dealing with multiple properties. Alzain *et al.* [7] present their findings on the utility of artificial intelligence (AI) in forecasting real estate prices within Saudi Arabia. They detail the precision of their AI models, share insights obtained from the prediction outcomes, and compare these results to those achieved through conventional valuation methods. This examination highlights the models' accuracy in predicting prices and underscores the potential improvements AI brings to the accuracy and efficiency of real estate valuation in the region. Santi *et al.* [8] focuses on devising a method aimed at diminishing bias in the estimation of spatial autoregressive models, particularly when confronted with incomplete geocoded data. The study's findings reveal that this novel approach significantly reduces bias in parameter estimation, outperforming traditional methods in scenarios characterized by partial geographical information. Classical methods are often inadequate when making value maps [9]. Multi-regression and nominal methods provide a mathematically related statistical approach to help better understand the situation. The value of both scenarios is determined by assigning a weight coefficient. In order to ascertain the accuracy of the method created, a multi regression method has been used in comparison with other existing studies [10]–[13]. Karamanou *et al.* [10] demonstrates a notable enhancement in the performance of the house price prediction model with the inclusion of location-based geospatial data (LOGD). This improvement is largely attributed to the model's improved contextual understanding and its capacity to discern complex interrelations among diverse factors influencing house prices. Additionally, the research underscores the utility of LOGD in offering crucial insights into the housing market, presenting a valuable tool for policymakers, investors, and homebuyers to make informed decisions. This integration of geospatial data into predictive modeling represents a significant stride forward in refining the accuracy and relevance of real estate market analyses. Jiang *et al.* [11] highlights several key factors influencing housing prices in Shenzhen, underscoring the substantial impact of distance to amenities, location, housing attributes, and spatial autocorrelation. Homes closer to schools, hospitals, and transport hubs command higher prices, reflecting the premium placed on convenience and accessibility. There's a stark variation in prices across different neighborhoods, attributed to disparities in socioeconomic status and infrastructure quality, with urban areas typically pricier than their suburban counterparts. While property characteristics such as size, age, and room count do affect prices, their influence is lesser compared to location-related factors. Additionally, the study reveals spatial autocorrelation among housing prices in Shenzhen, suggesting that economic conditions and development patterns in adjacent areas can affect local housing prices. When it comes to modeling techniques, ensemble methods like RF and gradient boosting machines were found to be more effective than LR and support vector regression (SVR) models. This superiority is due to their enhanced capability to handle complex interactions among the various factors affecting housing prices, offering more accurate predictions. This comprehensive analysis provides valuable insights for policymakers, investors, and homebuyers, emphasizing the multifaceted nature of real estate valuation. The aims of this paper [12] to uncover the intricate relationships between money supply, population dynamics, and rent fluctuations within Taiwan's real estate market. Through the identification of clusters within the data that display similar traits across these variables, the researchers provide a detailed analysis of how these factors interconnect and impact one another. This approach offers a refined insight into the dynamics of the real estate sector in Taiwan, highlighting the significant correlations that drive market behavior and trends. Such findings are invaluable for stakeholders looking to understand the complexities of the market, allowing for informed decision-making in investment, policy formulation, and market prediction.

Advanced valuation techniques which take advantage of computerized technology enable reviewing and analyzing a much larger number of data points. This also involves training computers to think like a human (utilizing artificial intelligence techniques) and achieve desired goals. Various techniques for predicting house prices, based on what is already known, have been developed. These are approaches used to guess values that are not known. In recent times, a considerable amount of research has been done using machine learning techniques, and the results have been impressive [14]. To predict user search results, google has used machine learning algorithms such as regularized regression (GLMNET), support vector machine (SVM), k-nearest neighbor (KNN), Naive Bayes, classification and regression tree (CART), bagged CART, and RF [15]–[18]. The experimental analysis of this paper [15] revealed that the artificial neural network (ANN) algorithm excelled beyond other tested algorithms in predicting house prices, showcasing remarkable performance metrics: 93% accuracy, 92% precision, 91% recall, and an F1-score of 91%. In comparison, the SVM algorithm demonstrated solid results with 87% accuracy, 86% precision, 85% recall, and an F1-score of 85%. The RF algorithm also showed good performance but lagged slightly behind with 84% accuracy, 83% precision, 82% recall, and an F1-score of 82%. The LR algorithm, while useful, proved to be the least effective among the tested methods, achieving 78% accuracy, 77% precision, 76% recall, and a corresponding F1-score of 76%. These results highlight the superiority of the ANN algorithm in handling the complexities of house price prediction, with its high accuracy and balanced performance metrics. Heidari *et al.* 16] evaluated the performance of their ensemble model using a dataset of business decisions from a manufacturing company. They compared the performance of their ensemble model to three benchmark models: logistic regression, DT, and KNN. The ensemble model outperformed the benchmark models in terms of accuracy and precision. Specifically, the ensemble model achieved an accuracy of 85% and a precision of 86%, compared to 80% and 82% for logistic regression, 78% and 79% for DT, and 75% and 77% for KNN, respectively. This study [18] introduced the KNN-MCF (missing completely at random) method, tested on a dataset of housing prices in Virginia, USA, featuring varying rates and patterns of missing data. The findings indicate that the KNN-MCF method surpasses conventional imputation techniques like mean imputation and regression imputation in both prediction accuracy and robustness, particularly in scenarios with high rates of missing data or complex missing data patterns. Additionally, when pitted against more sophisticated imputation strategies such as multiple imputation by chained equations (MICE) and MissForest (random forest-based imputation), the KNN-MCF method demonstrated comparable or even superior efficacy, notably in the context of spatial data. This suggests that the KNN-MCF method is a highly effective tool for dealing with missing data in housing price datasets, offering improved accuracy and robustness across various challenging data scenarios. Similarly in housing research studies, neural networks are being employed to predict house prices [16], [19]–[21].

Using satellite maps, some researchers collected data from public places such as schools, parks and exploited it for environmental analysis. Afterwards, several models were developed to make predictions. Several machine learning models are utilised in the current times, like gradient boosting and light gradient boosted machines, deep learning-based solutions and attention-based models [22]–[24]. Xu and Zhang [23] authors present their findings on house price forecasting using various neural network architectures. They compare these architectures, showcasing their performance through metrics like accuracy and mean squared error, to determine which model most effectively predicts house prices. The findings of this research [24] shed light on the significant spatial determinants that play a crucial role in real estate appraisals within the dutch context. Through the application of machine learning techniques, the study identifies key variables and their spatial relationships that contribute to property valuations. These insights can potentially enhance the accuracy of real estate appraisals and provide valuable information for stakeholders involved in the real estate market. Gradient boosting price prediction is highly reliant on five popular machine learning models: LR, DT, RF, Ridge and Lasso [25]. These algorithms analyze historical data of real estate market to make accurate predictions of house prices. Having carefully pre-processed the data, it was provided to a network for accurate classification. These models have proven to be highly accurate in their predictions.

ANNs have proven to be very effective in predicting real estate prices. To ensure the highest accuracy, parameters such as activation functions, weight initialization, number of hidden layers and the learning rate must be fine-tuned. A number of machine learning models such as long short-term memory (LSTM), general regression neural networks (GRNN) and backpropagation neural networks (BPNN) are efficient in accurately predicting the real estate prices. Pai and Wang [26] reveals that the random forest regression (RFR) model significantly surpasses both LR and SVR models in predicting real estate prices, as evidenced by its performance metrics. The RFR model demonstrated a mean absolute error (MAE) of 197,699 Taiwan new dollars, a root mean squared error (RMSE) of 499,268 TWD, and a coefficient of determination (R^2) of 0.937. This indicates a high level of accuracy and predictive power. In comparison, the LR model reported a MAE of 375,746 TWD, an RMSE of 874,592 TWD, and an R^2 of 0.864, while the SVR model showed a MAE of 344,875 TWD, an RMSE of 829,395 TWD, and an R^2 of 0.878. These results highlight the superior predictive capability of the RFR model in the context of real estate price forecasting, showcasing its effectiveness over more traditional or commonly used models like LR and SVR in capturing the complexities of real estate valuation. In this paper [21] explores the use of BPNN within the Keras deep learning framework for the purpose of predicting house prices. This research is pivotal as it delves into the application of cutting-edge machine learning techniques in the realm of real estate, aiming to refine the accuracy of house price predictions. The paper concludes that employing a back propagation neural network within the Keras framework marks a significant step forward in improving the precision of house price valuation. This advancement in machine learning application paves the way for better-informed decisions in the real estate market, highlighting the potential of technology to transform traditional practices.

Real estate valuation typically uses three common forms of AI: fuzzy logic, fuzzy inference system (FIS) and adaptive neuro-fuzzy inference system (ANFIS) [27], [28], neural networks and genetic algorithms (GA). Each of these techniques provides a reliable approach to obtaining property values within an acceptable margin of error. Compared to multiple regression analysis (MRA) models, GA models have been observed to be more successful in practice. Different methods have been compared in various studies, with the fuzzy logic approach being the one that got closest to the real market value. For example, it was found that the model created using FIS was more successful than the model created using ANN. This scientific paper [29] delves into the creation and comparative analysis of two distinct models for residential real estate valuation: the knowledge-based FIS and ANFIS. The FIS model leverages expert knowledge in real estate to establish a set of fuzzy rules for property value estimation, emphasizing the incorporation of expert insights into its decision-making framework. On the other hand, the ANFIS model merges the adaptive learning capabilities of neural networks with the logical reasoning of fuzzy systems, optimizing its rules and parameters through data-driven learning, making it highly adaptable to real estate valuation. The paper meticulously outlines the development, training, and validation processes for both models, comparing their performance in terms of accuracy, efficiency, and practical relevance. This comparison aims to elucidate the models' operational strengths and weaknesses in accurately predicting real estate values. The study's significance lies in its examination of different computational intelligence techniques for real estate valuation, offering valuable insights into the comparative effectiveness of these models. This knowledge assists real estate professionals, investors, and policymakers in making better-informed decisions about property valuation, investment strategies, and market dynamics, contributing to the advancement of strategic decision-making in the real estate sector.

The research problem at hand is the development of a reliable and efficient framework for predicting house prices in a market characterized by soaring demand and volatile prices. The traditional methods for appraising real estate comparison, revenue, and cost are widely used due to their simplicity and ease of implementation, but they fall short when dealing with multiple properties or constructing value maps. This limitation highlights the necessity for more sophisticated methodologies that can manage the complexity of today's real estate market. The burgeoning field of machine learning offers promising techniques for enhancing the accuracy of property valuations. These machine learning methods range from classical statistical models to advanced computerized technologies, including AI. Yet, the integration of these innovative approaches with the established methods has been met with resistance from some experts who doubt the ability of machine learning to replace or even complement conventional valuation models. This study explored the ability of machine learning models to accurately forecast and evaluate house prices. The main hypothesis is to understand what criteria are important in pricing. The main purpose of the research is to go into detail about the process involved in preparing data for training and predicting by using real examples. To gain a comprehensive understanding of the data, feature selection will be conducted based on the importance of each feature as well as several analytical techniques will be adopted. The remainder of this paper is organized as follows. Section 2 details the data and methods used for our evaluation. Section 3 presents the experimental studies and the results and section 4 discusses our findings.

2. METHOD

2.1. Criteria for calculating real estate value

The research delineated in this paper is predicated on a comprehensive analysis of data derived from advertisements listed under the ‘sale of secondary housing’ category on krisha.kz, a prominent online marketplace in Kazakhstan. This investigation is driven by the hypothesis that key variables such as quality, location, size, and material significantly influence house price estimations, and it empirically examines the strength and nature of these correlations. The uniqueness of this study lies in its methodological approach, employing advanced data scraping techniques to collect a large dataset that serves as the foundation for robust statistical analysis.

To construct the dataset, a specialized web scraping tool was developed using the Python programming language, leveraging the BeautifulSoup library [30] for parsing HTML and extracting the desired information. This tool systematically navigated through the website’s listings, targeting advertisements explicitly marked as “sold by the owners” to ensure the inclusion of direct seller-to-buyer transactions. On January 29, 2023, a total of 19,750 advertisements for apartment sales in Astana were meticulously collected, encompassing a wide range of attributes deemed relevant for the study. Each advertisement was analyzed for a comprehensive set of parameters, including but not limited to room count, price, location, square footage, type of home, floor number, condition of the apartment, availability of a mortgage, and amenities such as balcony type, internet access, parking facilities, telephone connectivity, toilet facilities, furnishings, and security features. A description of all parameters and their examples is presented in Table 1.

Table 1. Variables present in the initial data set

Variable	Type	Description	Example
room	Numeric	Room	1
urlid	Character	krisha.kz website link ID	/a/show/681645544
price	Numeric	Price	3,500,000
location	Factor	Administrative subdivision of the city of Astana	Yesil District, Astana
square	Numeric	Living area in square meters	13.2
hometype	Factor	Main building material	brick
floor	Numeric	Floor	2
basefloor	Numeric	Total number of floors	5
complex	Character	Name of the residential complex	Talterek
year	Numeric	Year of completion	2021
renovation	Factor	Condition	good
mortgaged	Factor	If it is pledged	pledged
balcony	Factor	Balcony	balcony and loggia
balconyg	Factor	If the balcony is glazed	yes
door	Factor	Door type	metallic
flooring	Factor	Flooring type	linoleum
ceiling	Numeric	Ceiling height	2.7
priv_dorm	Factor	if it is a former hostel	no
has_change	Factor	If exchange is possible with additional payment	exchange is possible
inet	Factor	Internet access type	via TV cable
parking	Factor	Parking type	parking
phone	Factor	If there is a landline telephone connection	possible to connect
toilet	Factor	Toilet type	separated
furniture	Factor	If there is furniture	fully
security	Factor	Safety conditions	video intercom, video surveillance, alarm, security, combination lock, intercom, concierge

These variables were initially cataloged as per their representation in the advertisements, with categorical and numerical data meticulously recorded. For analytical purposes, the raw data underwent a rigorous preprocessing stage. This involved the conversion of certain categorical variables into numerical values to facilitate quantitative analysis. Variables with significant impact on price estimation, such as the price per square meter, were calculated using direct formulas (e.g., price ÷ square footage). Additionally, variables with limited predictive value or complexity in numeric conversion, such as ‘complex’ and ‘urlid’, were excluded from subsequent analyses to refine the focus on more impactful determinants. This preprocessing yielded a dataset optimized for statistical modeling, allowing for the exploration of the relationships between house prices and the selected variables. The nuanced approach of converting categorical variables into numerical codes enabled the application of various statistical and machine learning models to predict house prices based on the identified determinants.

2.2. Statistical analysis when calculating the value of real estate

Prior to the initiation of statistical analysis, a thorough data cleaning process was undertaken to ensure the quality and reliability of the dataset. This preliminary step led to a reduction in the number of observations from the initial dataset to 18,992 valid entries. The refined dataset was subsequently subjected to a detailed descriptive statistical analysis, the outcomes of which are systematically presented in Table 2. This table elucidate the dataset’s characteristics through a range of summary statistics designed to offer insights into the distributional properties of the variables under study. The summary statistics include measures of central tendency (average, median, and mode), measures of variability (standard deviation, sample variance, and standard error), measures of skewness (excess and skewness), and measures of range (interval, minimum, and maximum). The average number of rooms is 1.94, while the median is 2. The standard deviation of the number of rooms is 0.96, indicating that the number of rooms varies greatly among the observations. The mode of the number of rooms is 1. The descriptive analysis shows that the average room capacity in the city is usually 2, but tends to decrease towards one-room apartments. The average price of a property is 33,277,050, with a standard deviation of 39,065,890. The median price is 26,000,000 and the mode is 25,000,000. The skewness of the price variable is 19.53, which indicates that the distribution is positively skewed (i.e., the tail of the distribution extends further to the right than to the left). The mathematical mode of the price shows a value of 25,000,000, which indicates a possible reduction in prices by more than 30% in accordance with the standards of the state mortgage program “7-20-25”, although if we take the average value of the set of prices, thus removing the increases from the average price, the decrease will be only 4%. This suggests that many sellers of real estate keep the price too high, which does not meet the capabilities and expectations of buyers. It can also be assumed that the difference of 30% is the pledged profit of realtors to the real price of the apartment. The descriptive analysis revealed notable insights, particularly regarding the real estate market’s dynamics. The observed central tendency measures indicate a market with a substantial presence of smaller properties, predominantly one to two-room apartments. Furthermore, the significant variability in the number of rooms and the pronounced positive skew in property prices suggest disparities in the housing market, potentially influenced by various factors including location, property features, and market demand. The price analysis, underscored by a considerable positive skewness, hints at a market segment that maintains prices well above the median, potentially misaligned with buyer capabilities and expectations. This discrepancy could be attributed to several factors, including speculative pricing or the expectation of negotiating down from a higher listing price. Moreover, the analysis suggests a potential misalignment with the "7-20-25" state mortgage program’s price reduction targets, indicating a gap between listed prices and those accessible to buyers utilizing this program.

The average pricepersquare meter is 516,232, with a standard deviation of 476,695. The median pricepersquare meter is 486,284. The mode of the pricepersquare meter is 500,000. The skewness of the pricepersquare meter variable is 114.74, which indicates a heavily positive skewed distribution. The average size of a property is 61.77 square meters, with a standard deviation of 36.06 square meters. The median size is 54 square meters. The range of the size variable is from 10 square meters to 800 square meters. The mathematical mode of the square area signals a decrease of 35%, this is the second statistical confirmation of the tendency to decrease the roominess of apartments towards one-room apartments. The average floor level is 6.33, with a standard deviation of 4.13. The median floor level is 5, and the mode is 3. The skewness of the floor level variable is 1.40, which indicates a slightly positively skewed distribution. The average basefloor level is 10.81, with a standard deviation of 4.84. The median basefloor level is 10, and the mode is 9. The skewness of the basefloor level variable is 1.34, which indicates a slightly positively skewed distribution. The average year of construction is 2014.49, with a standard deviation of 11.91 years. The median year of construction is 2019, and the range of years is from 1948 to 2025. The skewness of the year variable is -2.41, which indicates a negatively skewed distribution. The average ceiling height is 2.84 meters, with a standard deviation of 0.87 meters. The median ceiling height is 2.7 meters, and the mode is 2.7 meters. The skewness of the ceiling height variable is 27.83, which indicates a heavily positive skewed distribution. The average ceiling height is 2.83699 meters, which indicates that middle-class comfort apartments prevail on the market. This table provides a comprehensive summary of the different variables in the data set, and gives us a general idea of their distribution and variability. The small number of missing values in criteria such as location, hometype, mortgage, door, priv_dorm, suggests that this information is important for buyers when making a decision to purchase real estate.

Figure 1 shows the graphs of criteria called the plotmatrix function. The graphs in the 1:5 square show a positive relationship between the price and the area of the apartment, which proves the correctness of the correlation. We can also see from the squares 1:7 and 1:8 that the price is inversely correlated with the number of floors of the house. This dependence has not been established by the correlation method. It was noticed that apartments are more expensive if the floor is below the 5th floor, but apartments in high-rise buildings, on the contrary, have a lower price than in low-rise buildings. The graph from the 1:9 square shows that the year of construction of the building also affects the price of the apartment, the newer the house, the higher the price. Also, from the 1:24 square, we can find that the price is affected by the level of security description. The higher the security, the higher the price. From the 1:4 square, we can see that prices in the Yesil district are higher than in the rest. From the 1:6 square, it can be understood that apartments in houses made of monolith and brick are more expensive than ones made from other materials. Also, the prices are affected by the material from which the floor is made, from the 1:15 square, it can be understood that the better the floor material, the higher the price of the apartment. To our great surprise, we recorded according to the graph from the 1:16 square that the ceiling height negatively affects the price of the apartment; that is, the lower the ceiling height, the higher the apartment price. The rest of the charts regarding the price did not provide any interesting information. From the 2:9 square, it was noticed that one-room apartments are built more often in new houses. From the squares 2:7 and 2:8, it is noticeable that the higher the number of floors of the house, the more often one-room apartments are located on them. From the 2:2 square it can be seen that the market is dominated by advertisements for sale of 1-2 room apartments. From the 4:4 square, it is noticed that apartments are most often sold in the Yesil and Almaty districts. The 6:6 square shows that apartments in houses with brick or monolithic construction are more often sold on the market. Squares 7:7 and 8:8 reflect the fact that apartments on the lower floors are more often sold.

Table 2. Descriptive statistics of numeric basic housing characteristics

Value/Criteria	Room	Price	Price per square	Square	Floor	Base floor	Year	Ceiling
Average	1.942555	33277050	516232	61.77392	6.331613	10.81139	2014.487	2.83699
Standard error	0.006956	283473.4	3459.044	0.261639	0.029944	0.035147	0.086437	0.006331
Median	2	26000000	486284.3	54	5	10	2019	2.7
Mode	1	25000000	500000	40	3	9	2022	2.7
Standard deviation	0.958579	39065890	476695.9	36.05687	4.126683	4.843637	11.91198	0.872499
Sample variance	0.918874	1.53E+15	2.27E+11	1300.098	17.02951	23.46082	141.8954	0.761255
Excess	2.454064	859.786	14796.55	63.78346	3.442933	3.387208	5.882879	814.7383
Skewness	1.075821	19.52917	114.7413	5.176789	1.396552	1.344595	-2.4111	27.83083
Interval	10	2.45E+09	62126633	790	39	42	77	31.3
Minimum	1	3500000	98373.56	10	1	1	1948	2
Maximum	11	2.45E+09	62225006	800	40	43	2025	33.3
Sum	36893	6.32E+11	9.8E+09	1173210	120250	205330	38259139	53880.11
Count	18992	18992	18992	18992	18992	18992	18992	18992

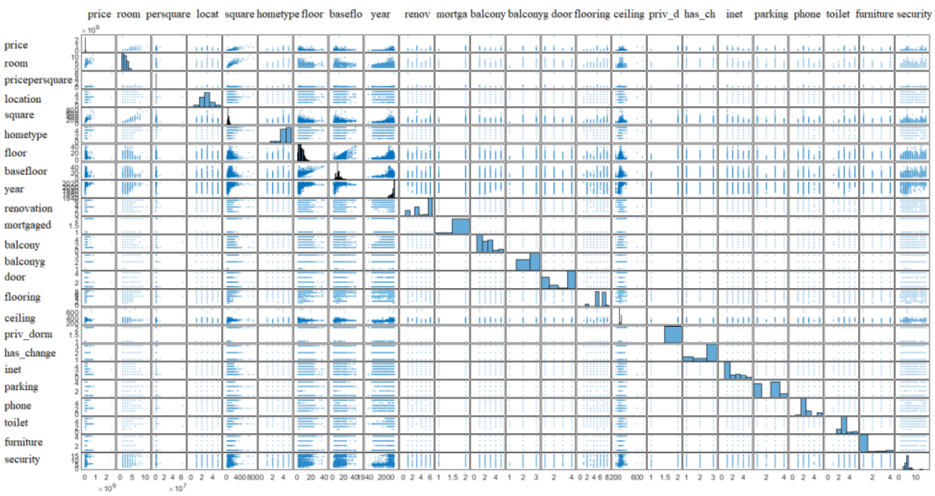

Figure 1. Plotmatrix charts

3. RESULTS AND DISCUSSION

3.1. Regressions in determining the value of real estate

For a better regression calculation, we divided the price value by one million, for example, 23,000,000 = 23 million. As an output parameter, we also set the values of the price column. In the calculations, we used 12 forecast models, the results of calculating the error of forecast models are presented in Table 3. The results of the models in the Table 3 show the performance of each model on a regression task. The models are compared based on four performance metrics: RMSE, R-squared, MSE, and MAE. Based on the results, the models with the best performance in terms of lowest RMSE and MAE values are boosted trees (BoosT) and bagged trees (BagT), with values of 22.102 and 22.181 respectively. These models also have the highest R-squared values of 0.68, indicating that they explain a high proportion of the variation in the target variable. The linear models such as linear, interactions linear, and robust linear have relatively higher RMSE and MAE values compared to the tree-based models and SVM models. The R-squared values of these linear models are also relatively lower compared to the tree-based and SVM models. The SVM models have varying performance with the medium gaussian SVM having the highest RMSE and MAE values and the lowest R-squared value. The linear SVM and quadratic SVM have lower RMSE and MAE values compared to the medium gaussian SVM but are still higher compared to the tree-based models. Overall, the results of the models indicate that the tree-based models, particularly BoosT and BagT, have the best performance in terms of the four metrics considered.

Table 3. Regression models in determining the value of real estate

Model	RMSE	R-squared	MSE	MAE
LR	26.031	0.56	677.62	8.4775
ILR	25.44	0.58	647.21	7.9248
RLR	29.379	0.43	863.12	7.0961
FTR	25.712	0.57	661.09	6.0571
MTR	23.873	0.63	569.93	5.8746
CTR	24.172	0.62	584.27	6.015
LSVM	28.123	0.48	790.89	6.9569
QSVM	25.599	0.57	655.3	6.0408
MGSVM	29.771	0.42	886.3	6.1388
RQGPR	23.209	0.65	538.65	6.8929
BoosT	22.102	0.68	488.49	5.7518
BagT	22.181	0.68	491.98	5.1441

3.2. Neural networks in forecasting real estate prices

To build a neural network, we will use MATLAB [31], which has a very user-friendly interface [32]. The first neural network model for the forecast will contain price on the output parameter, and the second model will contain pricepersquare data. The neural network will consist of 70 layers and work on the Bayesian regularization algorithm. Based on this data, we will train Bayesian regularization-backpropagation neural network (BR-BPNN) [33]. Bayesian regularization [34] is a helpful strategy for training neural networks. It uses a prior distribution over the network weights, which can be updated during training by Bayesian inference. This helps reduce the chances of overfitting and ensures better performance. Training a deep neural network with 70 layers can be a challenging experience due to the presence of hyperparameters such as learning rate, batch size and regularization strength. In order to achieve successful results, all these parameters need to be judiciously selected and Bayesian regularization needs to be employed. When setting up an AI network, one must take into account the size and complexity of the data set being used in order for it to be effective and efficient. To effectively train a network with 70 layers using Bayesian regularization, it is recommended to use a powerful computing platform such as a high-performance graphics processing units (GPU), as the training process can be computationally intensive. This process can be quite resource intensive and require considerable amounts of memory and processor power. It is also important to carefully monitor the training process, regularly checking for signs of overfitting or underfitting and adjusting the hyperparameters accordingly. The second model with pricepersquare output data showed MSE=6.6483e+09 R=0.9873. The first model with price output data showed RMSE=7.146, MSE=32.14 and R=0.9899. The results of comparing machine learning models are shown in Figure 2.

Figure 2 presents a comparative analysis of various machine learning models applied to the estimation of real estate values. It encompasses a spectrum of models ranging from simple LR to more sophisticated non-linear and ensemble approaches, illustrating their respective efficacies in property value prediction. The figure also highlights the trade-offs between model simplicity and predictive power, demonstrating that while simpler models may offer ease of interpretation and faster computation times, more complex models, including ensemble methods, tend to provide superior accuracy by effectively capturing non-linear relationships and interactions within the data. The analysis underscores the importance of model selection in real estate valuation, suggesting that the choice of algorithm should be guided by the specific characteristics of the dataset and the objectives of the analysis.

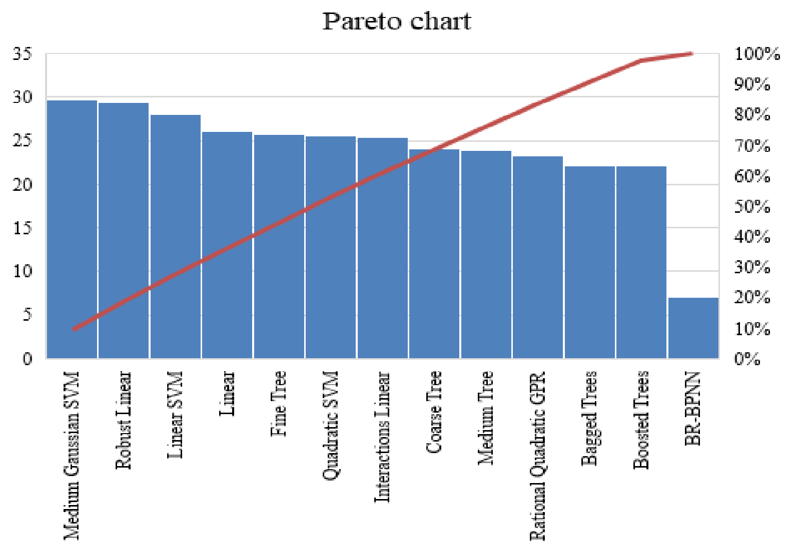

Figure 2. Comparison of machine learning models

The linear model has a moderate R-squared value of 0.56 and the highest errors across RMSE, MSE, and MAE compared to other models, indicating a lesser fit to the data. The interaction linear model shows a slight improvement in all metrics, suggesting that including interactions between variables provides a better fit. The robust linear model, designed to be less sensitive to outliers, shows higher error metrics and a lower R-squared value, indicating that it may not be the best model in this context. The fine tree model improves on the linear models, but it is the medium tree model that has the lowest RMSE and MSE among the tree-based models, combined with a high R-squared value, which suggests a good fit. The performance of SVM models varies with the kernel used. The linear SVM has high error metrics, whereas the Quadratic SVM performs better, as indicated by the lower error metrics. The medium gaussian SVM model does not perform as well as other SVM models, with the highest RMSE and MSE in the table, and a lower R-squared value. The rational quadratic gaussian process regression model shows better performance than many of the other models, with the second-best R-squared value of 0.65. Ensemble methods such as BoosT and BagT show good predictive performance, with low error metrics and high R-squared values, suggesting they are robust for predicting real estate values. Remarkably, the BR-BPNN outperforms all other models by a significant margin, exhibiting the lowest RMSE, MSE, and MAE values, along with the highest R-squared value of 0.9899. This indicates that the BR-BPNN model has excellent predictive capabilities and provides the best fit to the real estate valuation data among the models compared.

This study investigates the efficacy of machine learning algorithms in accurately predicting and evaluating residential property values. The results of the study confirmed the main hypothesis, which is aimed at identifying the most important factors influencing the valuation of real estate.

4. CONCLUSION

The dataset acquired through web scraping was subjected to a comprehensive analysis using a suite of models, including descriptive statistics, correlation analysis, LR, ILR, RLR, FTR, MTR, CTR, LSVM, QSVM, MGSVM, RQGPR, BoosT, BagT, BR-BPNN. Descriptive statistical analysis revealed an average room count of two, with a trend towards smaller, single-room apartments. In light of the "7-20-25" state mortgage program, the modal price of 25,000,000 suggests potential for a 30% reduction in rates. However, an analysis of average prices, adjusted for increases, shows only a 4% decrease, indicating a discrepancy between owner-set prices and market expectations, with the 30% gap potentially representing expected profit margins. Further, the analysis indicates a 35% decrease in square meterage, aligning with the trend towards smaller apartments. The average ceiling height was found to be 2.837 meters, suggesting a prevalence of middle-class comfort apartments. Strong correlations were observed between price and factors such as room size, price per square meter, and total area, following expected positive trends. Construction patterns suggest phased building activities, leading to uneven dilapidation levels, with location and construction year showing significant correlations.

A notable correlation was found between property age and renovation frequency, indicating a tendency to sell without updating. Conversely, a negative relationship was identified between the presence of furniture and improvements, suggesting these factors are independent. The plot matrix in Figure 1 illustrates a positive correlation between apartment price and area, with inverse correlations observed between price and floor level, suggesting lower-floor apartments command higher prices, contrasting with higher floors in high-rise buildings. Security features and construction year were positively correlated with apartment prices, with higher security districts and newer buildings commanding premium prices. The Yesil district emerged as a higher-priced area, with construction materials like monolith and brick influencing prices. Interestingly, analysis revealed an inverse relationship between ceiling height and price, contrary to expectations. A peculiar trend was observed in newly constructed buildings favoring one-room apartments, with an increase in one-bedroom units as buildings added more floors. Market analysis showed a predominance of 1-2 room apartment listings, particularly in Yesil and Almaty districts, with brick and monolithic constructions being more sought after. Lower-floor units showed higher purchase rates. The neural network-based model, specifically the BR-BPNN, outperformed SVM, BoosT, and other models in predicting prices, demonstrating RMSE=7.146, MSE=32.14, and R=0.9899. These findings offer valuable insights for the real estate market, aiding in buying/selling decisions and guiding construction companies in developing marketable real estate. This methodology is applicable not only to Kazakhstan's market but can also be adapted for international markets and different sectors.

REFERENCES

[1] S. Lahmiri, S. Bekiros, and C. Avdoulas, "A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization," *Decision Analytics Journal*, vol. 6, p. 100166, Mar. 2023, doi: 10.1016/j.dajour.2023.100166.

[2] K. Baur, M. Rosenfelder, and B. Lutz, "Automated real estate valuation with machine learning models using property descriptions," *Expert Systems with Applications*, vol. 213, p. 119147, Mar. 2023, doi: 10.1016/j.eswa.2022.119147.

[3] J. S. Chou, D. B. Fleshman, and D. N. Truong, "Comparison of machine learning models to provide preliminary forecasts of real estate prices," *Journal of Housing and the Built Environment*, vol. 37, no. 4, pp. 2079–2114, Dec. 2022, doi: 10.1007/s10901-022-09937-1.

[4] J. Kim, Y. Lee, M. H. Lee, and S. Y. Hong, "A comparative study of machine learning and spatial interpolation methods for predicting house prices," *Sustainability (Switzerland)*, vol. 14, no. 15, p. 9056, Jul. 2022, doi: 10.3390/su14159056.

[5] R. Sibindi, R. W. Mwangi, and A. G. Waititu, "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices," *Engineering Reports*, vol. 5, no. 4, Apr. 2023, doi: 10.1002/eng2.12599.

[6] M. Castelli, M. Dobreva, R. Henriques, and L. Vanneschi, "Predicting days on market to optimize real estate sales strategy," *Complexity*, vol. 2020, pp. 1–22, Feb. 2020, doi: 10.1155/2020/4603190.

[7] E. Alzain, A. S. Alshebami, T. H. H. Aldhyani, and S. N. Alsubari, "Application of Artificial Intelligence for Predicting Real Estate Prices: The Case of Saudi Arabia," *Electronics (Switzerland)*, vol. 11, no. 21, p. 3448, Oct. 2022, doi: 10.3390/electronics11213448.

[8] F. Santi, M. M. Dickson, D. Giuliani, G. Arbia, and G. Espa, "Reduced-bias estimation of spatial autoregressive models with incompletely geocoded data," *Computational Statistics*, vol. 36, no. 4, pp. 2563–2590, Dec. 2021, doi: 10.1007/s00180-021-01090-7.

[9] J. Zaki, A. Nayyar, S. Dalal, and Z. H. Ali, "House price prediction using hedonic pricing model and machine learning techniques," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 27, Dec. 2022, doi: 10.1002/cpe.7342.

[10] A. Karamanou, E. Kalampokis, and K. Tarabanis, "Linked open government data to predict and explain house prices: the case of scottish statistics portal," *Big Data Research*, vol. 30, p. 100355, Nov. 2022, doi: 10.1016/j.bdr.2022.100355.

[11] X. Jiang, Z. Jia, L. Li, and T. Zhao, "Understanding housing prices using geographic big data: a case study in Shenzhen," *Sustainability (Switzerland)*, vol. 14, no. 9, p. 5307, Apr. 2022, doi: 10.3390/su14095307.

[12] C. H. Yang, B. Lee, and Y. Da Lin, "Effect of money supply, population, and rent on real estate: a clustering analysis in Taiwan," *Mathematics*, vol. 10, no. 7, p. 1155, Apr. 2022, doi: 10.3390/math10071155.

[13] M. Y. Huang, "Analyzing the effects of green building on housing prices: case study of Kaohsiung, Taiwan," *Environment, Development and Sustainability*, vol. 25, no. 2, pp. 1205–1235, Feb. 2023, doi: 10.1007/s10668-021-02089-2.

[14] S. Kumar and M. H. Syed, "House pricing prediction based on composite facility score using machine learning algorithms," in *Lecture Notes in Networks and Systems*, vol. 421, 2023, pp. 235–248.

[15] Z. Liu and Y. Wang, "City's house pricing prediction based on machine learning algorithms," in *International Symposium on Robotics, Artificial Intelligence, and Information Engineering (RAIIE 2022)*, Nov. 2022, p. 25, doi: 10.1117/12.2658679.

[16] M. Heidari, S. Zad, and S. Rafatirad, "Ensemble of supervised and unsupervised learning models to predict a profitable business decision," in *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*, Apr. 2021, pp. 1–6, doi: 10.1109/IEMTRONICS52119.2021.9422649.

[17] A. A. Neloy, H. M. Sadman Haque, and M. M. Ul Islam, "Ensemble learning based rental apartment price prediction model by categorical features factoring," in *ACM International Conference Proceeding Series*, Feb. 2019, vol. Part F148150, pp. 350–356, doi: 10.1145/3318299.3318377.

[18] K. Sanjar, O. Bekhzod, J. Kim, A. Paul, and J. Kim, "Missing data imputation for geolocation-based price prediction using KNN-MCF method," *ISPRS International Journal of Geo-Information*, vol. 9, no. 4, p. 227, Apr. 2020, doi: 10.3390/ijgi9040227.

[19] O. K. Di, N. F. M. Radzuan, and M. N. A. Rani, "Prototype of predictive Kuala Lumpur house pricing system through neural network technique," in *AIP Conference Proceedings*, 2022, vol. 2617, p. 020007, doi: 10.1063/5.0119705.

[20] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House price prediction using a machine learning model: A survey of literature," *International Journal of Modern Education and Computer Science*, vol. 12, no. 6, pp. 46–54, Dec. 2020, doi: 10.5815/ijmecs.2020.06.04.

[21] Z. Jiang and G. Shen, "Prediction of house price based on the back propagation neural network in the keras deep learning framework," in *2019 6th International Conference on Systems and Informatics, ICSAI 2019*, Nov. 2019, pp. 1408–1412, doi: 10.1109/ICSAI48974.2019.9010071.

[22] N. Chen, "House price prediction model of Zhaoqing City based on correlation analysis and multiple linear regression analysis," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–18, May 2022, doi: 10.1155/2022/9590704.

[23] X. Xu and Y. Zhang, "House price forecasting with neural networks," *Intelligent Systems with Applications*, vol. 12, p. 200052, Nov. 2021, doi: 10.1016/j.iswa.2021.200052.

[24] E. Guliker, E. Folmer, and M. van Sinderen, "Spatial determinants of Real Estate appraisals in the Netherlands: a machine learning approach," *ISPRS International Journal of Geo-Information*, vol. 11, no. 2, p. 125, Feb. 2022, doi: 10.3390/ijgi11020125.

[25] C. H. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House price prediction using regression techniques: a comparative study," in *6th IEEE International Conference on and Smart Structures and Systems, ICSSS 2019*, Mar. 2019, pp. 1–5, doi: 10.1109/ICSSS.2019.8882834.

[26] P. F. Pai and W. C. Wang, "Using machine learning models and actual transaction data for predicting real estate prices," *Applied Sciences (Switzerland)*, vol. 10, no. 17, p. 5832, Aug. 2020, doi: 10.3390/app10175832.

[27] S. A. Abdymanapov, M. Muratbekov, S. Altynbek, and A. Barlybayev, "Fuzzy expert system of information security risk assessment on the example of analysis learning management systems," *IEEE Access*, vol. 9, pp. 156556–156565, 2021, doi: 10.1109/ACCESS.2021.3129488.

[28] S. A. Abdymanapov, A. Barlybayev, and B. A. Kuzenbaev, "Quality evaluation fuzzy method of automated control systems on the LMS example," *IEEE Access*, vol. 7, pp. 138000–138010, 2019, doi: 10.1109/ACCESS.2019.2943000.

[29] S. Yalpir and G. Ozkan, "Knowledge-based FIS and ANFIS models development and comparison for residential real estate valuation," *International Journal of Strategic Property Management*, vol. 22, no. 2, pp. 110–118, Mar. 2018, doi: 10.3846/ijspm.2018.442.

[30] A. Omarbekova, A. Sharipbay, and A. Barlybaev, "Generation of test questions from RDF files using PYTHON and SPARQL," *Journal of Physics: Conference Series*, vol. 806, no. 1, p. 012009, Feb. 2017, doi: 10.1088/1742-6596/806/1/012009.

[31] A. Barlybayev, L. Zhetkenbay, D. Karimov, and B. Yergesh, "Development neuro-fuzzy model to predict the stocks of companies in the electric vehicle industry," *Eastern-European Journal of Enterprise Technologies*, vol. 4, no. 4(124), pp. 72–87, Aug. 2023, doi: 10.15587/1729-4061.2023.281138.

[32] A. Sharipbay, A. Barlybayev, and T. Sabyrov, "Measure the usability of graphical user interface," in *Advances in Intelligent Systems and Computing*, vol. 444, 2016, pp. 1037–1045.

[33] T. Nguyen, J. Shiau, and D. K. Ly, "Enhanced earth pressure determination with negative wall-soil friction using soft computing," *Computers and Geotechnics*, vol. 167, p. 106086, Mar. 2024, doi: 10.1016/j.compgeo.2024.106086.

[34] P. Wu, Z. Zhang, X. Peng, and R. Wang, "Deep learning solutions for smart city challenges in urban development," *Scientific Reports*, vol. 14, no. 1, p. 5176, Mar. 2024, doi: 10.1038/s41598-024-55928-3.

BIOGRAPHIES OF AUTHORS

Alibek Barlybayev received the B.Eng. degree in information systems from L.N. Gumilyov Eurasian National University, Kazakhstan, in 2009 and the M.S. and Ph.D. degrees in computer science from L.N. Gumilyov Eurasian National University, Kazakhstan, in 2011 and 2015, respectively. Currently, he is a Director of the Research Institute of Artificial Intelligence, L.N. Gumilyov Eurasian National University. He is also an Associate Professor of the Department of Artificial Intelligence Technologies, L.N. Gumilyov Eurasian National University, and Higher School of Information Technology and Engineering, Astana International University. His research interests are NLP, the use of neural networks in word processing, smart textbooks, fuzzy logic, stock market price prediction, and information security. He can be contacted at email: frank-ab@mail.ru.

Sankibayev Arman Temirchanovich Received the B.Sc. and M.Sc. Degree in computer science from the L.N. Gumilyov Eurasian National University. Worked in the IT department of the Kazakh Railways in administrative posts. In 2010-2019 he worked as a database programmer, was an Oracle database system expert. Also, he has been developing web applications. Then worked as a systems analyst. In 2019-2022, he studied at doctoral studies at L.N. Gumilyov Eurasian National University. Master's research thesis was "Detection of pseudo-enterprises using tax reporting". The task was to identify false enterprises that evaded taxes, paid little taxes or carried out other criminal economic activities. His vast experience in the field of data analytics have enabled him to become systems analyst. He is currently works as a senior lecturer at the Department of Information Systems and Technologies at the Esil University. His research interests include data science, neural networks, image classification and analysis, and cybersecurity. He can be contacted at email: armanjana1980@gmail.com.

Rozamgul Niyazova is graduated from the Faculty of Information Technology of Atyrau State University named after Kh. Dosmukhamedov with a degree in physics and computer science. In 2010, she defended candidate of Technical Science, specialty 05.13.11 – "Mathematical and software of computers, complexes and computer networks" on the topic "Methods for improving the performance of a distributed corporate information system". In 2011 she present L.N. Gumilyov Eurasian National University, Department of Theoretical computer science, then an associate professor of artificial intelligence technologies. Her research interests include artificial intelligence, machine learning, and intelligent learning systems. She can be contacted at email: rs.niyazova@gmail.com.

Gulnara Akimbekova is senior lecturer at the Department of Microbiology and Virology named after Sh.I. Sarbasova of Astana Medical University. Having graduated from the Faculty of Natural Sciences of L.N. Gumilyov ENU with a special diploma (with honors), in September 2009 she was accepted to the department as an assistant; transferred to the position of senior lecturer in 2010. Conducts practical classes for 2nd year students in the state and Russian languages; takes part in the development of educational and methodological complexes in the disciplines taught at the department; responsible for document flow, the responsibility for maintaining minutes of departmental meetings is also fixed. Author and co-author of a number of scientific publications, including in European publications. She can be contacted at email: akimbekova.g@amu.kz.