Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

The Peshawar Property Puzzle: Predicting Prices in Hayatabad with Machine Learning

Muhammad Shehzad
GDC, Hayatabad, Higher Education Department, KPK
Email: muhammadshahzad8390@gmail.com

Sameer Bhatti
GDC, Hayatabad, Peshawar, Higher Education Department, KPK
Email: bhattisameer980@gmail.com

Muhammad Hamza
Lecturer, Government Degree College, KTS, Haripur Email: hk246337@gmail.com

Abdullah Malik
GDC, Hayatabad, Peshawar, Higher Education Department, KPK
Email: abdullahmalik3483@gmail.com

Mujeeb Ullah
GDC, Hayatabad, Peshawar, Higher Education Department, KPK
Email: mujeebkh194@gmail.com

Luqman Ahmad
GDC, Hayatabad, Peshawar, Higher Education Department, KPK
Email: luqman.ahmad.cs@gmail.com

ABSTRACT
The prediction of residential property prices has become an essential component of modern urban planning and real estate investment. This research focuses on predicting house prices in the Hayatabad region of Peshawar using a machine learning-based approach. A dataset of 1,300 housing records was collected and preprocessed to include structural, locational, and facility-related attributes. The target variable (house price) was categorized into three levels—low, medium, and high—to formulate a classification problem. Multiple algorithms were implemented and evaluated, including Logistic Regression (LR), Support Vector Machine (SVM), eXtreme Gradient Boost (XGBoost), and a proposed Multilayer Perceptron (MLP). A five-fold cross-validation strategy was adopted to ensure the reliability of the results. Among the tested models, the proposed MLP achieved the highest overall accuracy, demonstrating superior ability to learn nonlinear relationships between housing features and price categories. The findings indicate that ensemble-based models can provide more robust and precise predictions compared to linear approaches. This study contributes to the development of data-driven methods that support decision-making for real estate professionals, investors, and policymakers in regional housing markets.

Introduction

The real estate industry plays a crucial role in the economic growth and urban development of any region. Accurate prediction of housing prices has therefore become

ISSN Online: 3007-3154
ISSN Print: 3007-3146

an essential component of real estate analysis, influencing both individual and institutional decision-making. With the rapid expansion of cities and the rising demand for residential properties, traditional methods of price estimation based on manual assessment and limited historical data are no longer sufficient. In recent years, the integration of machine learning techniques into real estate analytics has opened new avenues for more accurate, efficient property price prediction. Machine learning enables processing large, diverse datasets, identifying hidden patterns, and learning complex relationships among variables that influence property valuation. Factors such as location, size, number of bedrooms and bathrooms, access to main roads, availability of utilities, and additional features like air conditioning or basements can all contribute to variations in house prices. By applying data-driven models to these features, it becomes possible to generate more consistent and reliable predictions compared to subjective human judgments or traditional statistical methods.

This research focuses on the Hayatabad region of Peshawar, a well-structured, planned residential area known for its distinct phases and diverse property values. The objective of this study is to develop and evaluate predictive models that can classify house prices into three categories—low, medium, and high—based on multiple features extracted from a structured dataset of 1,300 housing records. The study explores the effectiveness of several machine learning algorithms, including Logistic Regression, Ridge Classifier, Random Forest, and XGBoost, in predicting property price categories. A five-fold cross-validation technique was employed to ensure the robustness and generalizability of the results. The comparative analysis shows that ensemble-based models, such as Random Forest and XGBoost, outperform traditional linear models in terms of accuracy and predictive consistency. The outcomes of this research not only demonstrate the potential of machine learning for real estate price prediction but also provide a valuable framework to assist real estate professionals, investors, and policymakers in making informed, data-driven decisions in the Hayatabad housing market.

Literature Review

The study of housing price dynamics has long been anchored in economic theory, real estate investment, and behavioral finance. Baum (Baum, 2015) Emphasized real estate as a strategic asset influencing national and individual wealth portfolios. Muellbauer and Murphy (Muellbauer & Murphy, 2008) Examined the interplay between housing markets and macroeconomic stability, highlighting how property cycles reflect wider economic fluctuations. Shiller (Shiller, 2007) Provided early evidence linking housing price trends to psychological and behavioral expectations rather than purely market fundamentals, while Armona et al. (Armona et al., 2019) Reinforced this view through experimental data on information-driven price expectations. From a structural standpoint, the hedonic pricing model emerged as a cornerstone for assessing property values based on intrinsic and locational attributes. Wei et al. (Wei et al., 2022) Reviewed the evolution of hedonic models in the era of big data, noting that these models laid the groundwork for computational valuation methods. Complementary perspectives from Maslow (Maslow, 1943) and Baum (Baum, 2015) Indirectly frame housing demand within the contexts of human motivation and investment security. Together, these studies established the foundational rationale for quantitative and machine-learning-based price-estimation models.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

Traditional Regression-Based Approaches

Early computational attempts to predict house prices relied primarily on linear regression techniques due to their interpretability and simplicity. Aljohani (Aljohani, 2021) Demonstrated that stable estimations could be achieved with multivariate regression analysis when feature selection was optimized. Similar comparative studies, such as those by Madhuri et al. (Madhuri et al., 2019) and Ghosalkar and Dhage (Ghosalkar & Dhage, 2018), confirmed the reliability of regression for small-scale datasets, though with declining accuracy in non-linear scenarios. Zhang (Zhang, 2021) Applied multiple linear regression to large-scale datasets, validating the model's consistency in structured markets but highlighting its limitations when faced with unbalanced feature distributions. Sanyal et al. (Sanyal et al., 2022) and Wang (Wang, 2021) Expanded on this by benchmarking ordinary least squares (OLS) against ensemble methods, showing that OLS underperforms in complex data environments. Meanwhile, hedonic regression models continued to evolve. Wei et al. (Wei et al., 2022) outlined their adaptation to digital datasets and property indexing, while Muellbauer and Murphy (Muellbauer & Murphy, 2008) Discussed how traditional econometric models understate externalities and spatial dependencies. These findings ultimately motivated the migration from statistical to machine-learning paradigms for improved nonlinear pattern recognition.

Machine Learning and Neural Network Models

The application of machine learning to housing price prediction represents a significant methodological shift—studies such as those by Manasa et al. (Manasa et al., 2020), Truong et al. (Truong et al., 2020), and Zulkifley et al. (Zulkifley et al., 2020) Introduced regression-based machine learning frameworks capable of handling high-dimensional and non-linear data more effectively than traditional approaches. Random forests (RFs) and gradient boosting algorithms have emerged as preferred models due to their ensemble-based structure and robustness. Adetunji et al. (Adetunji et al., 2022) utilized RF to achieve high predictive accuracy across multiple regions, while Hjort et al. [17] experimented with different loss functions in gradient-boosted trees to optimize performance. Chen and Guestrin (Chen & Guestrin, 2016) formally established XGBoost as a scalable and efficient gradient boosting framework, and its success has since influenced numerous real estate prediction models. (Zhou et al., 2021), (Sibindi et al., 2023). Rogers and Gunn (Rogers & Gunn, 2005) further demonstrated feature relevance optimization using random forests, providing a foundation for feature engineering in housing applications.

Artificial neural networks (ANN) added another dimension of flexibility. Rahman et al. (Rahman et al., 2019) Modeled the Malaysian housing market using ANN architectures, outperforming linear models in nonlinear trend capture. Liu and Liu (Liu & Liu, 2019) Enhanced temporal forecasting by integrating long short-term memory (LSTM) networks with a modified genetic algorithm, proving effective in volatile Chinese property markets. Varma et al. (Varma et al., 2018) and Phan (Phan, 2018) Applied deep learning frameworks in India and Australia, respectively, showing improved accuracy but higher data dependency. Beyond supervised learning, Usama et al. (Usama et al., 2019) Explored unsupervised techniques within broader networking contexts, laying conceptual groundwork for clustering and pattern recognition in real estate data. Ogunleye (Ogunleye, 2021) and Alpaydin (Alpaydin, 2020) Contributed broader statistical learning principles, emphasizing model interpretability and generalization.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

Big Data, Automation, and Hybrid Methods

The increasing availability of large datasets and computational resources has driven interest in hybrid and automated machine learning solutions. Lee et al. (Lee et al., 2021) Integrated macro and microeconomic factors with landscape data using big data analytics to forecast land prices, while Li et al. (Li et al., 2022) Employed AutoML techniques for predicting carpark price indices in Hong Kong—an approach transferable to real estate valuation. Recent advances have focused on combining multiple models for improved stability. Soltani et al. (Soltani et al., 2022) Introduced a spatio-temporal dependency framework within machine learning algorithms, demonstrating significant accuracy gains. Sibindi et al. (Sibindi et al., 2023) Proposed a hybrid ensemble blending LightGBM and XGBoost models, achieving competitive performance in housing datasets. Shuai et al. (Shuai et al., 2018) Combined PCA-SVM-GridSearchCV for software evaluation, but demonstrated hybrid principles applicable to real estate valuation workflows. Clustering and genetic programming techniques have also gained traction. Azimlu et al. (Azimlu et al., 2021) Used clustering with genetic programming to enhance housing prediction accuracy, while Zhou et al. (Zhou et al., 2021) Applied XGBoost to fraud prediction—reinforcing the model's adaptability across structured data environments. Fan et al. (Fan et al., 2018) And Truong et al. (Truong et al., 2020) Further showed that hybridized regression-ensemble approaches can outperform standalone models. Together, these studies reflect a clear trend toward automation, scalability, and hybridization in modern predictive modeling.

Research Gaps

Despite extensive progress, significant research gaps persist. Traditional econometric models (Muellbauer & Murphy, 2008), (Shiller, 2007) Remain limited by their assumptions of linearity and homogeneity. Regression-based machine learning models (Aljohani, 2021), (Zhang, 2021), (Sanyal et al., 2022) Struggle with spatial and temporal dependencies inherent in housing data. Deep learning approaches (Rahman et al., 2019), (Liu & Liu, 2019), (Varma et al., 2018) require large, well-labeled datasets—a challenge in regions with inconsistent data collection. Moreover, explainability and transparency are critical issues. While ensemble and hybrid models (Soltani et al., 2022), (Sibindi et al., 2023), (Chen & Guestrin, 2016) offer superior performance, their "black-box" nature limits interpretability, which remains crucial for financial and policy decision-making. Additionally, integration with macroeconomic and behavioral factors (Baum, 2015), (Armona et al., 2019) is often neglected in favor of purely data-driven designs. The convergence of big data, automated machine learning, and spatial analytics presents promising opportunities.

Methodology

This section outlines the systematic procedure for developing, training, and evaluating multiple machine learning models for housing price classification in the Hayatabad Peshawar region. The approach integrates data preprocessing, feature encoding and scaling, model-specific training strategies, learning rate scheduling, and robust cross-validation for performance evaluation. The selected models—Logistic Regression, XGBoost, Support Vector Machine (SVM), and proposed Multilayer Perceptron (MLP)—cover a diverse range of linear and nonlinear learning paradigms to ensure comprehensive comparison. Figure 1 shows the overall framework of this study.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

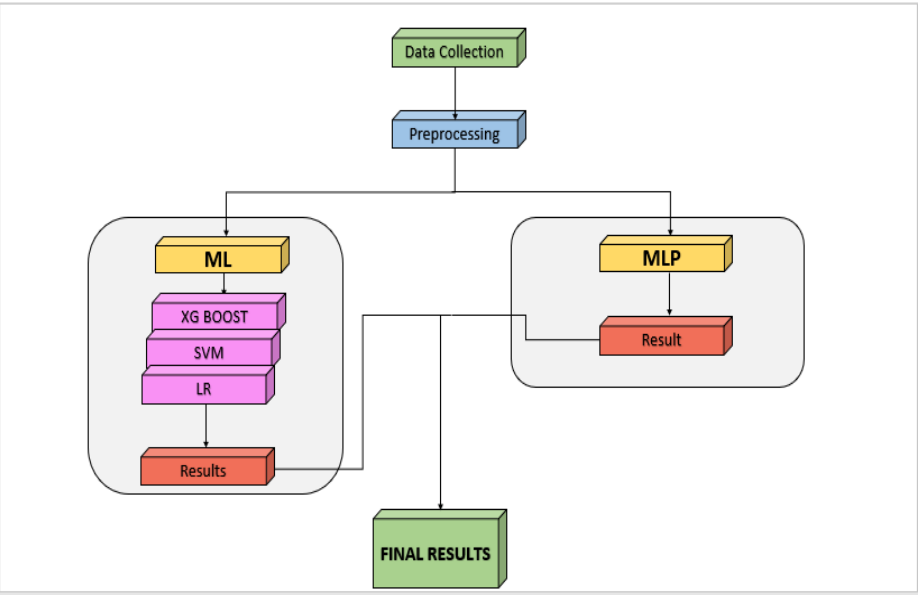

Figure 1: Overall framework of the study

Data Collection

The dataset employed in this study was assembled from three reputable sources, ensuring diversity and representativeness of housing trends in the Hayatabad region of Peshawar, Pakistan. The overall data collection process utilized three distinct channels to generate a raw dataset prepared for subsequent analysis. Data were collected from three primary sources, as shown in Figure 2.

Local (Property Dealers): This channel leveraged agents' networks and on-the-ground presence to gather market insights. Property dealers provided authentic on-ground information regarding current market prices and property features.

Direct (Individual Owners): This channel focuses on collecting data directly from individual owners through direct contact. This approach ensured the most authentic data related to the properties, including size, facilities, and sale value.

Digital (Zameen.com Platform): This channel leveraged the Zameen.com platform, Pakistan's largest real estate portal. It contributed extensive listings and real-time updates to provide structured data on listed properties in the Hayatabad locality.

In total, 1,300 housing records were compiled, encompassing both numerical and categorical variables that comprehensively describe each property's structural design, available amenities, and environmental context. The dataset was organized around the following key feature categories:

Structural characteristics: Total covered area (in square feet), number of bedrooms, bathrooms, stories, and parking spaces.

Facilities and utilities: Availability of air conditioning, basements, hot water heating systems, and guest rooms.

Locational and lifestyle indicators: Access to main roads, the specific Phase of Hayatabad (representing sub-regional zones), and the property's furnishing status (unfurnished, semi-furnished, or fully furnished).

The target variable was the property price, expressed in Pakistani Rupees (PKR). To facilitate classification-based modeling, the continuous price values were later transformed into three categorical classes—Low, Medium, and High—based on percentile thresholds (0–33%, 34–66%, 67–100%). This transformation enabled the use of classification algorithms to predict a property's price category rather than its exact

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

numerical value, thereby enhancing interpretability and supporting more actionable decision-making for non-technical users, such as property buyers, sellers, and agents. Furthermore, care was taken to verify data consistency and authenticity through cross-checking between dealer-reported values, owner-provided information, and online listings. This multi-source integration minimized bias and ensured that the dataset accurately reflected real-world housing market conditions in the Hayatabad area.

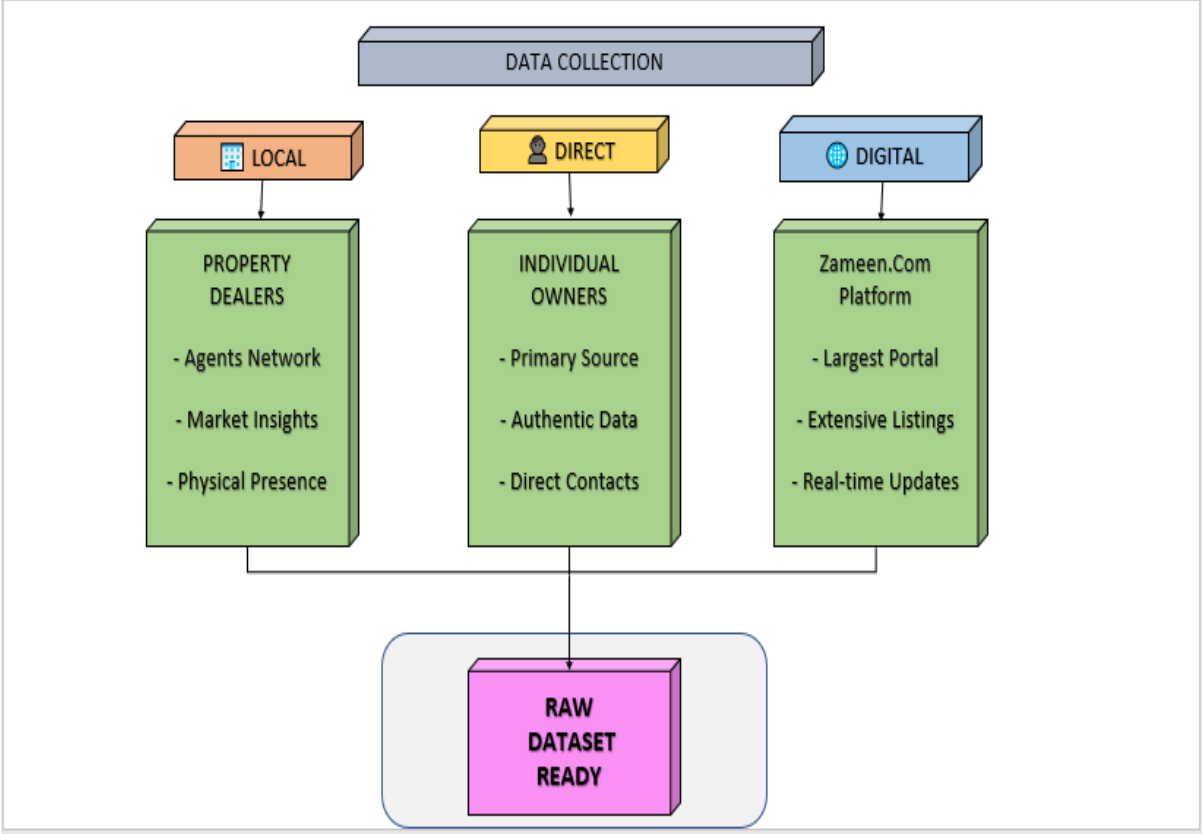

Figure 2: Data Collection Sources used in Data Acquisition

Data Preprocessing

Data preprocessing was a crucial step in transforming the raw data, collected from local dealers, individual owners, and zameen.com, into a consistent, reliable, and model-ready format. This entire process took place within the data processing plant, focusing on quality control, feature creation, and standardization to yield a refined dataset. A series of preprocessing steps was applied to address variations in representation and completeness stemming from the multi-source compilation.

Quality Control

This stage focused on cleaning and validating the data.

Price Validation: The dataset was first checked for erroneous or unrealistic property price entries, which were then validated against market norms.

Remove Duplicates: Duplicate records and inconsistent listings across different sources were identified and removed to ensure that each housing record was unique and accurate.

Handle Missing: Minor missing entries were managed through logical imputation. For instance, missing binary facility values (e.g., basement, guest room) were replaced with "No," while missing numeric fields (e.g., parking spaces) were filled using the median of their respective distributions.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

Outlier Detection: Extreme outliers in property prices and covered area were detected primarily using the Interquartile Range (IQR) method and handled via Winsorization to prevent unrealistic values from distorting model learning.

Feature Engineering

New, informative features were created, and existing categorical features were converted to numerical values compatible with machine learning algorithms.

Location Factors: Categorical location features, such as the Phase of Hayatabad, were encoded.

Property Type: This includes features like the number of stories.

Amenities: This includes facilities such as air conditioning and guest rooms.

Binary Flags: Binary encoding was explicitly applied to nominal attributes representing the presence or absence of specific facilities (e.g., Air Conditioning, Basement, Hot Water Heating), creating simple 0/1 indicator variables.

Composite Features: Additional features were engineered to capture deeper property value drivers, including:

Price/Sqft (Price Per Square Foot): Calculated by dividing the property price by the covered area.

Room Ratios: Ratios derived from the number of bedrooms, bathrooms, and total rooms.

Lot Sizes**:** Derived from the covered area.

Age Scaling: If property age was available, it was scaled or grouped.

Standardization (Scaling): The final step involved Standardization (Scaling) to normalize the range of numerical attributes. A Robust Scaler was employed for normalization, as it effectively minimizes the impact of potential outliers by scaling features according to the interquartile range rather than the mean and standard deviation. This process ensured that large-valued features (e.g., area) did not dominate smaller-valued features (e.g., stories) during model training.

Target Transformation

The target variable, house price, initially a continuous numeric value, was transformed to enable classification modeling. Using percentile-based thresholds, the prices were divided into three discrete categories: Low Price (0–33%), Medium Price (34–66%), and High Price (67–100%). This transformation simplified the prediction problem into a multiclass classification task, improving interpretability for stakeholders by aligning with practical decision-making scenarios (e.g., identifying affordable vs. premium segments).

Data Splitting and Balancing

The dataset was examined for class imbalance in the new categorical target variable. Minor imbalances were addressed using random oversampling techniques to ensure fair representation and prevent model bias toward the majority category. Finally, the prepared dataset was split into training and evaluation sets using a 5-fold cross-validation, maximizing data utilization and ensuring an unbiased performance estimate. Figure 3 shows a complete step-by-step process of the data preprocessing plant.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

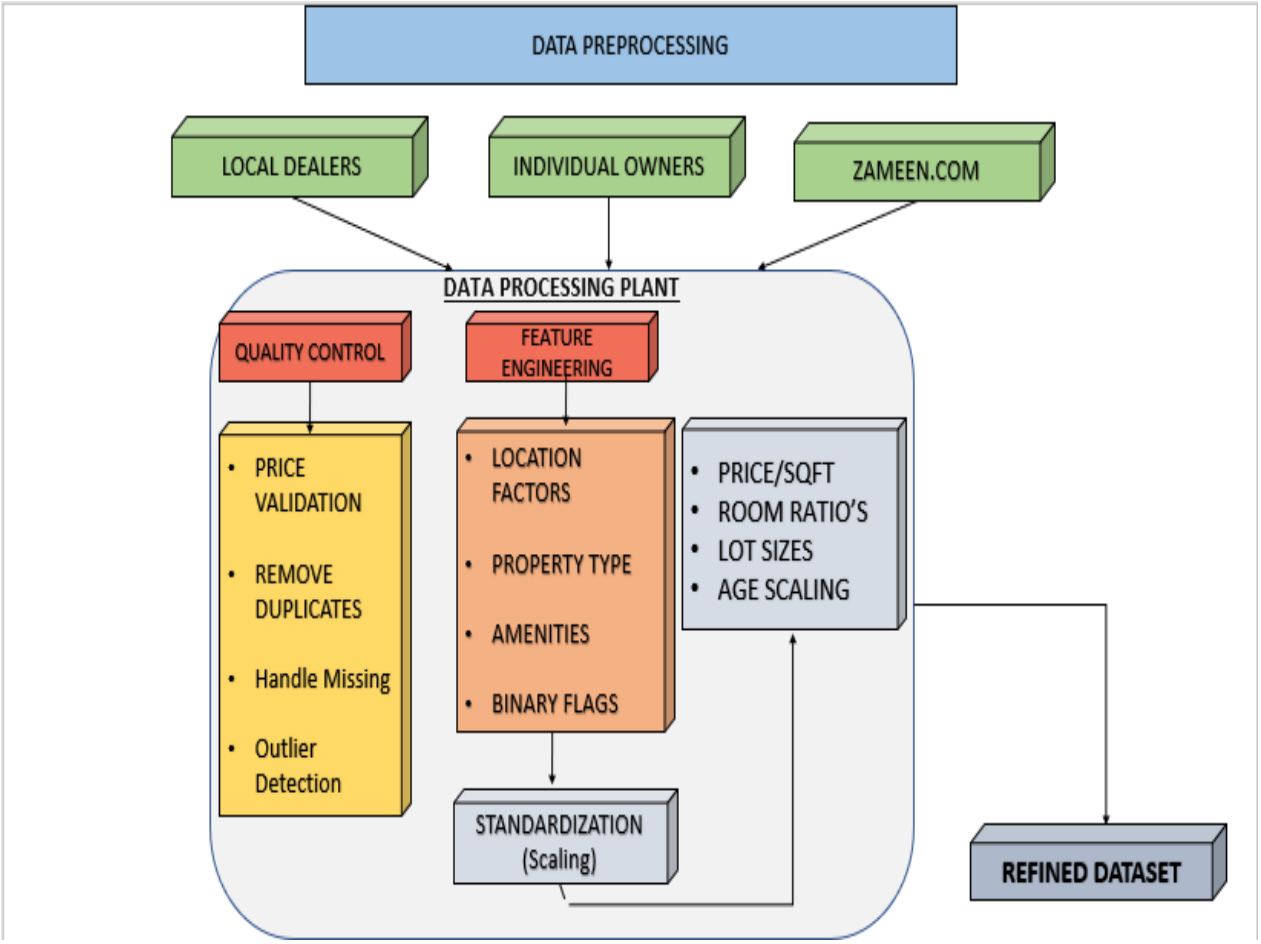

Figure 3: Preprocessing steps involved in data refining

Model Selection

After preparing and refining the dataset, several supervised machine learning algorithms were employed to classify houses into Low**,** Medium, and High price categories. The models were carefully selected to represent a diverse range of learning paradigms—from linear and kernel-based methods to ensemble learning and deep neural architectures. This multi-model approach ensured a comprehensive comparison of classification performance, generalization capability, and computational efficiency.

Logistic Regression

Logistic Regression (LR) was implemented as the baseline model due to its simplicity, interpretability, and robustness in handling linearly separable datasets. LR maps input features through a sigmoid activation function to estimate the probability of a class belonging to a particular category. Despite being a linear classifier, it performs effectively when the dataset is moderately complex and noise-free. In this study, Logistic Regression served as a benchmark to assess the improvement achievable with more sophisticated nonlinear models. It provided a precise reference point for understanding the gains obtained through feature interaction modeling in subsequent algorithms.

Support Vector Machine (SVM)

The Support Vector Machine (SVM**)** classifier was selected for its strong capability to handle nonlinear decision boundaries and high-dimensional data spaces. The model utilizes the Radial Basis Function (RBF) kernel, which enables it to separate classes that

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

are not linearly separable in the input space by mapping data into a higher-dimensional feature space. SVM is particularly well-suited for structured datasets with a moderate number of samples, as it maximizes the margin of separation between classes and is relatively less prone to overfitting. In this research, SVM was expected to capture subtle variations among property features (e.g., area, phase, and number of rooms) that define distinct price categories.

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost**)** was chosen for its exceptional performance and efficiency in handling tabular data. It operates as an ensemble of decision trees trained sequentially, where each tree attempts to correct the errors of its predecessors. The algorithm optimizes a differentiable loss function via gradient boosting, enabling it to model complex nonlinear relationships while maintaining high interpretability through feature importance analysis. XGBoost includes several regularization techniques (L1 and L2 penalties) to prevent overfitting and enhance generalization. In this study, XGBoost was expected to perform well due to its ability to model intricate dependencies among categorical and numerical housing attributes, including interactions among location, area, and the number of facilities.

Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP)**,** a type of feedforward artificial neural network, was implemented as the deep learning model in this research. The MLP architecture, as shown in Figure 4, consisted of three hidden layers with nonlinear activation functions (ReLU), allowing the model to learn complex, nonlinear mappings between housing attributes and their respective price categories.

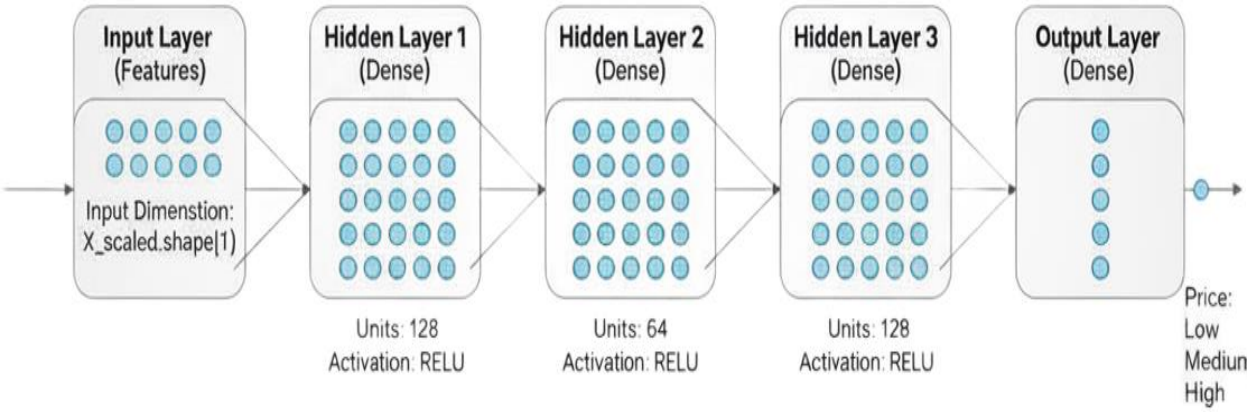

Figure 4: Proposed MLP Architecture

Regularization techniques such as dropout and early stopping were applied to mitigate overfitting and improve convergence stability. The choice of MLP was motivated by its strong representational power, which makes it particularly effective for capturing high-order interactions among features — such as how area, furnishing, and location collectively influence a property's price. Although computationally more demanding than traditional algorithms, MLP provides a robust mechanism to approximate intricate feature relationships, thereby complementing the other models in the experimental setup. Figure 5 shows the overall ecosystem of the proposed MLP.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

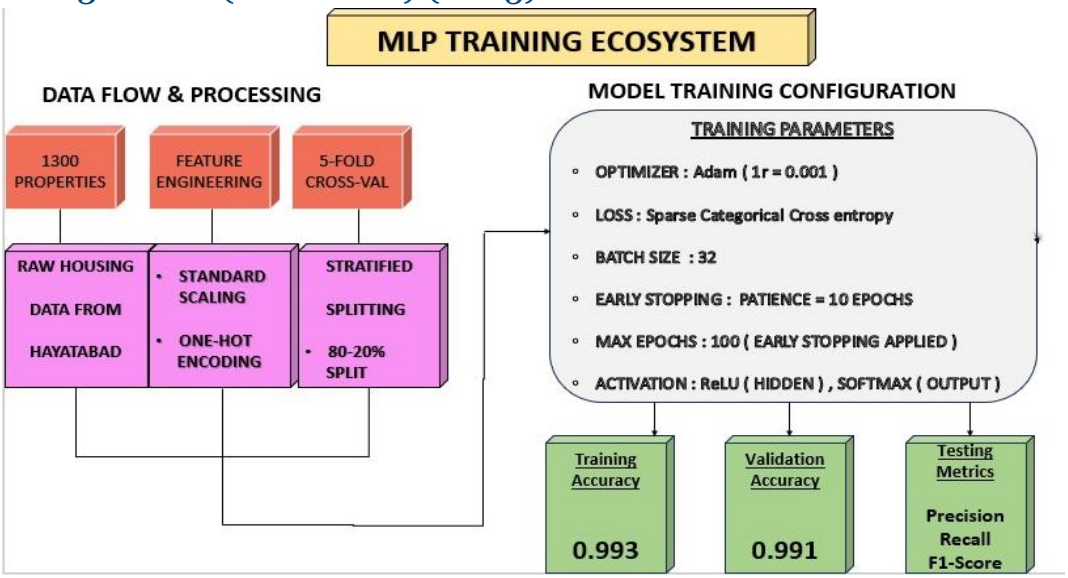

Figure 5: overall ecosystem of the proposed MLP

Model Evaluation Strategy

Each selected model was evaluated using a stratified K-fold cross-validation approach (K=5), ensuring that all price categories were proportionally represented in both the training and validation folds. Performance was assessed using key metrics, including accuracy, precision, recall, and F1-score, providing a balanced view of predictive performance across all classes. Additionally, learning curves and confusion matrices were analyzed to assess each model's convergence behavior, generalization ability, and class-wise prediction reliability.

Results and Discussion

This section presents the experimental findings from implementing four machine learning models: Logistic Regression**,** Support Vector Machine (SVM)**,** XGBoost Classifier, and Multilayer Perceptron (MLP). Each model was evaluated using five-fold cross-validation to ensure fair and unbiased assessment. The performance was measured using key metrics such as Accuracy**,** Precision**,** Recall, and F1-score**,** along with visual evaluations via learning curves and confusion matrices.

Logistic Regression Results

The Logistic Regression model, serving as a baseline linear classifier, achieved an average training accuracy of 74.6% and an average validation accuracy of 73.6%**.** The precision, recall, and F1-score were 0.748**,** 0.754, and 0.749, respectively. While the model demonstrated reasonable predictive power for linearly separable patterns, its overall accuracy was limited, reflecting its inability to fully capture nonlinear relationships among housing attributes. The learning curve, as shown in Figure 6, indicated a moderate gap between training and validation accuracy, suggesting slight underfitting. The confusion matrix shown in Figure 7 revealed that misclassifications primarily occurred between the Medium and High price categories, which often overlap in real estate data.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

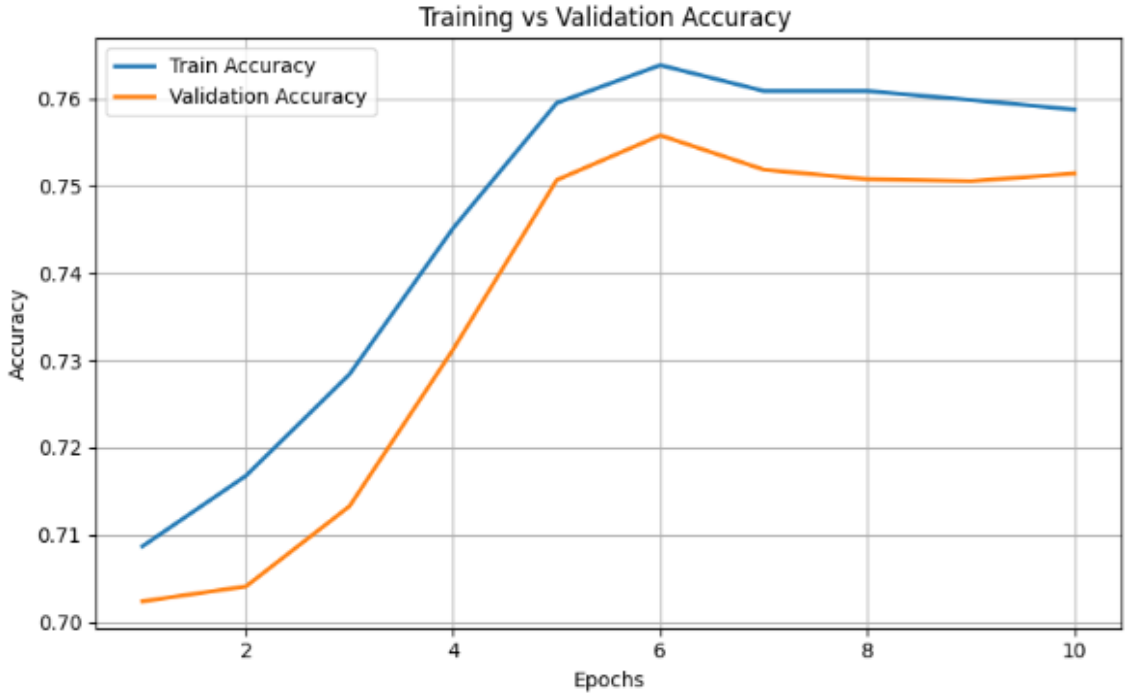

Figure 6: Training and Validation accuracy of the Logistic regression

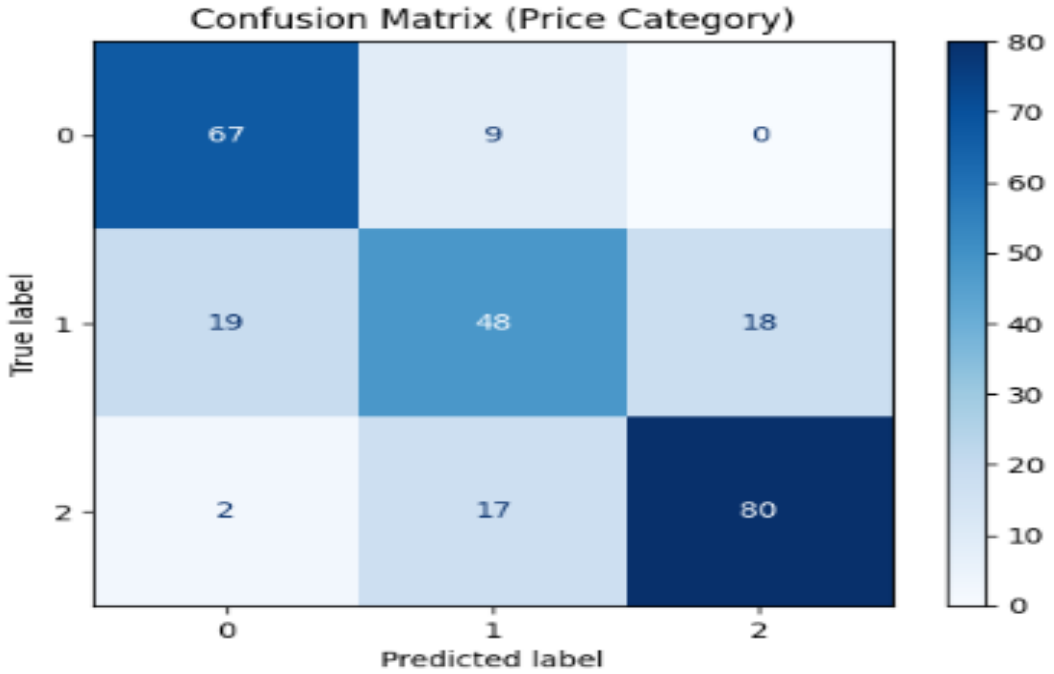

Figure 7: Confusion Matrix of logistic regression

Support Vector Machine (SVM) Results

The SVM classifier achieved an average training accuracy of 95.9% and an average validation accuracy of 95.2%. Precision, recall, and F1-score were recorded as 0.957, 0.951, and 0.952, respectively, indicating strong generalization and balanced classification across categories. The learning curve shown in Figure 8 exhibits smooth convergence between the training and validation curves, indicating good model stability. The confusion matrix, as shown in Figure 9, confirmed that SVM effectively separated the Low, Medium, and High price classes with minimal confusion, highlighting its

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

strength in handling nonlinear feature boundaries.

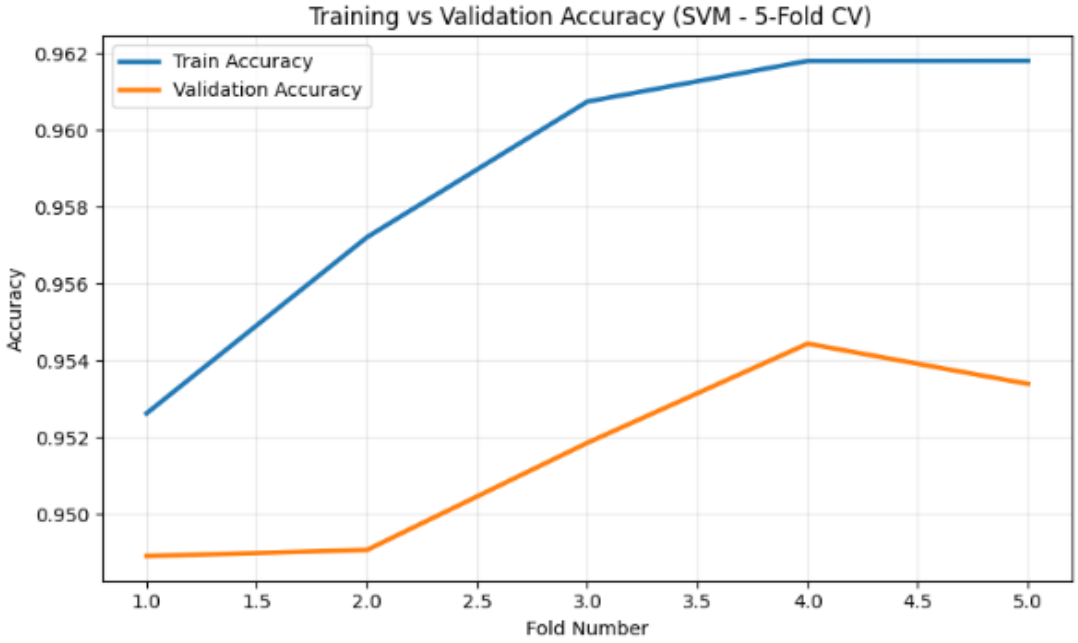

Figure 8: SVM training and validation accuracy curve

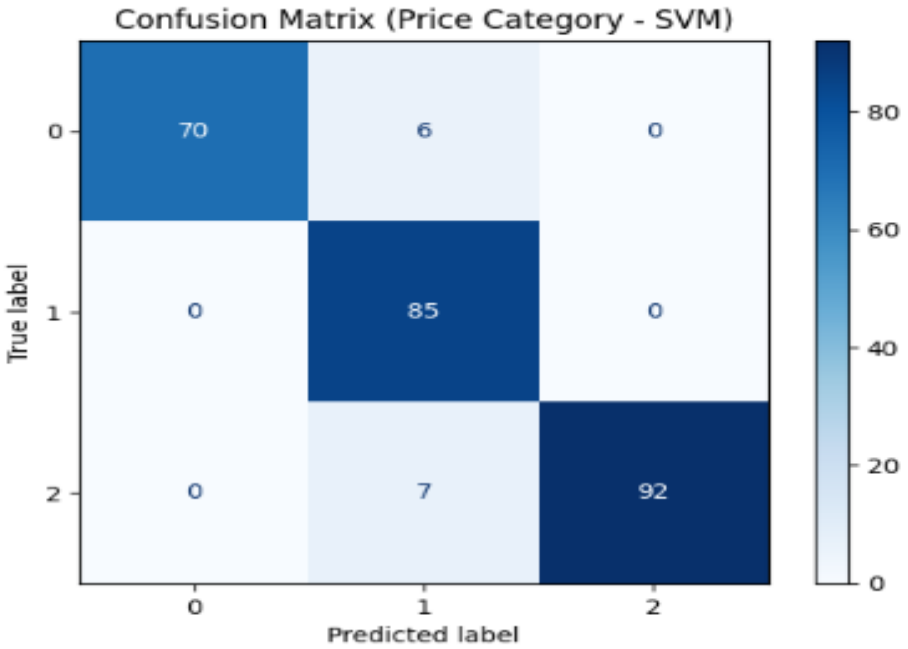

Figure 9: SVM prediction Confusion Matrix

XGBoost Classifier Results

The XGBoost Classifier performed exceptionally well, achieving an average training accuracy of 95.9% and an average validation accuracy of 95.4%**.** Precision, recall, and F1-score were 0.956**,** 0.954**,** and 0.953**,** respectively, reflecting excellent predictive balance and robustness. The learning curve, as shown in Figure 10, illustrated consistent growth in both training and validation accuracy, showing effective generalization without overfitting. The confusion matrix, as shown in Figure 11, demonstrated that the model achieved high classification confidence across all three categories, particularly in

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

correctly identifying high-priced houses.

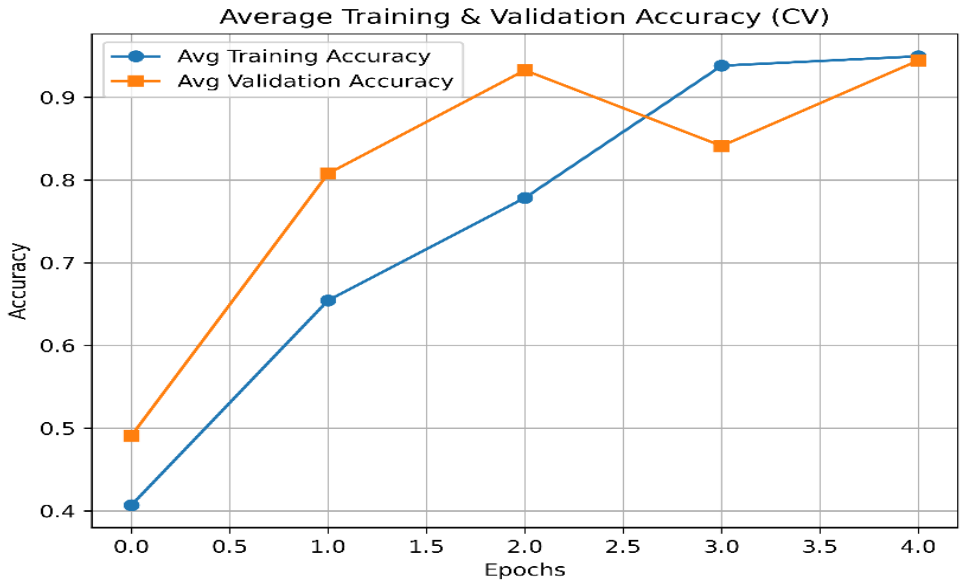

Figure 10: XGBoost training and validation accuracy curve

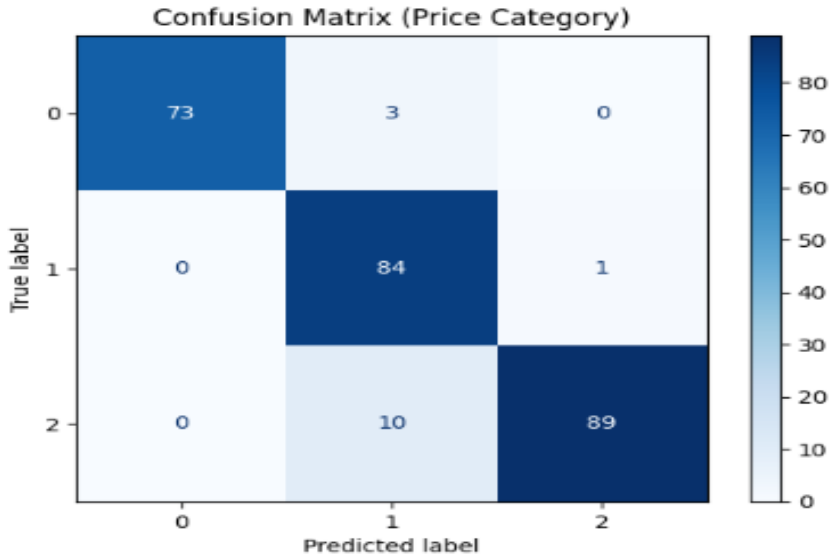

Figure 11: XGBoost prediction confusion matrix

The strong performance of XGBoost can be attributed to its gradient-boosting optimization, feature interaction handling, and regularization mechanisms, which enable it to learn complex nonlinear relationships in housing data efficiently.

Multilayer Perceptron (MLP) Results

The MLP model with three hidden layers achieved the highest performance among all models tested. It reached an average training accuracy of 99.3% and an average validation accuracy of 99.1%. Precision, recall, and F1-score were all 0.991, indicating that the model correctly classified nearly all samples across the Low, Medium, and High price categories. The learning curve, as shown in Figure 12, demonstrated near-perfect convergence between training and validation accuracy, confirming the model's strong generalization and minimal overfitting. The confusion matrix, as shown in Figure 13, revealed almost flawless classification performance, indicating that the MLP effectively captured the complex and nonlinear dependencies among housing attributes.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

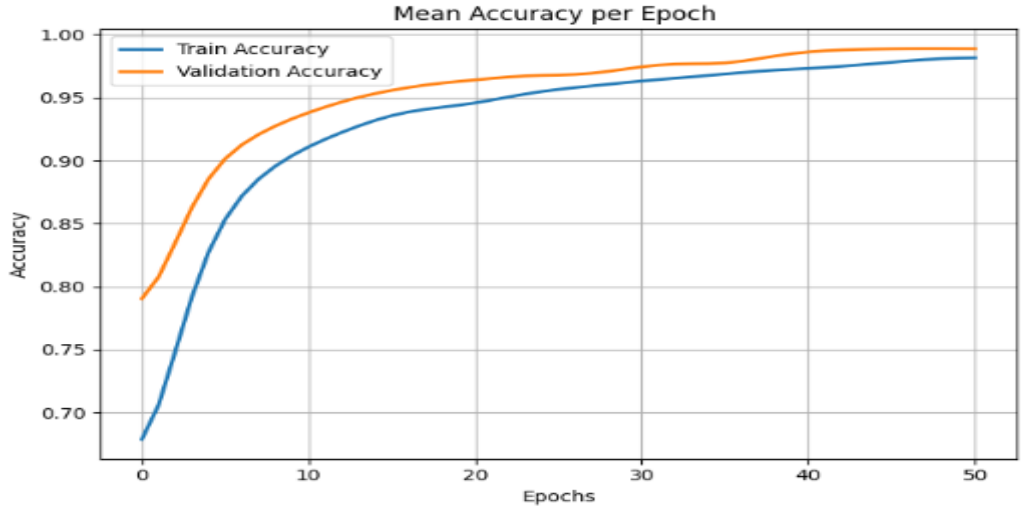

Figure 12: MLP training and validation accuracy curve

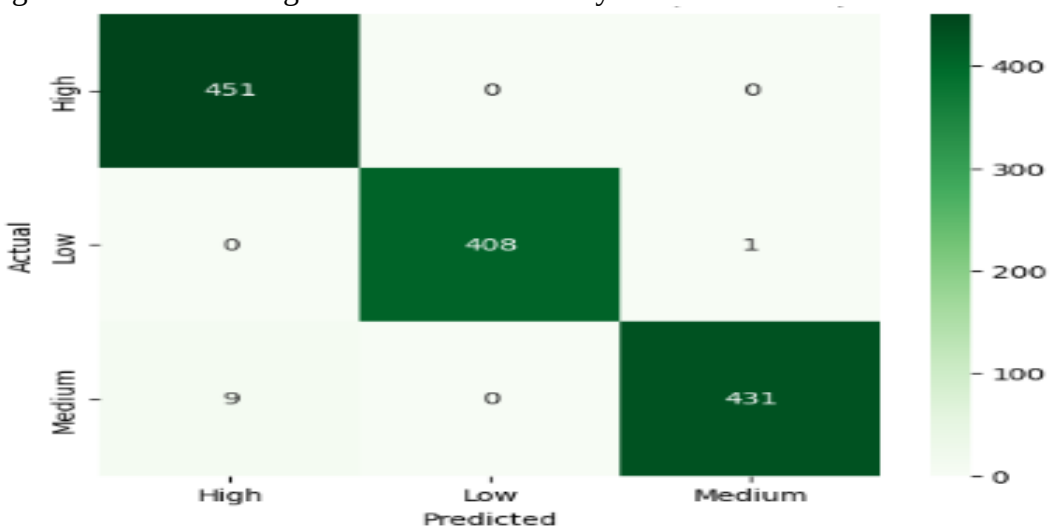

Figure 13: MLP prediction confusion matrix

Although the MLP required longer training time and was sensitive to hyperparameter tuning (e.g., learning rate, batch size, and dropout ratio), its superior accuracy validated its suitability for real estate price classification.

The results clearly show that nonlinear models (SVM, XGBoost, and MLP) significantly outperformed the linear Logistic Regression baseline. Among these, the MLP achieved the highest overall accuracy and F1-score (99.1%), followed closely by XGBoost (95.4%) and SVM (95.2%). The findings highlight that deep learning and ensemble techniques are better at capturing the intricate feature interactions and nonlinear dynamics in real estate pricing data. A comparative summary of model performance is presented in Table 1.

Table 1 Comparison of different models

Model	**Train Accuracy**	**Val Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.746	0.736	0.748	0.754	0.749

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

Support Vector Machine	0.959	0.952	0.957	0.951	0.952
XGBoost Classifier	0.959	0.954	0.956	0.954	0.953
Multilayer Perceptron	**0.993**	**0.991**	**0.991**	**0.991**	**0.991**

The results confirm that housing price prediction involves inherently nonlinear, multidimensional feature interactions that are not fully captured by linear classifiers such as Logistic Regression. Tree-based and neural models (XGBoost and MLP) excelled at learning nonlinear boundaries and complex hierarchical relationships among variables such as area, phase, bedrooms, and furnishing status.

XGBoost's superior performance can be attributed to:

Gradient boosting ensemble mechanism**,** Automatic feature interaction learning**,** and Built-in regularization (L1 & L2) to prevent overfitting. Meanwhile, the MLP's success illustrates the potential of deep learning architectures for structured data when combined with effective normalization and regularization. Overall, the ensemble and neural models demonstrated strong potential for real-world deployment in real estate price prediction systems, offering accurate, scalable, and interpretable results when integrated with modern property listing platforms.

Conclusion and Future Work

This research aimed to develop and evaluate machine learning models for predicting housing price categories in the Hayatabad region of Peshawar. The study systematically compared four distinct models—Logistic Regression**,** Support Vector Machine (SVM)**,** Multilayer Perceptron (MLP), and XGBoost Classifier—through a rigorous five-fold cross-validation procedure. The results demonstrated that housing price prediction in this context is a nonlinear, feature-interaction problem influenced by structural, locational, and lifestyle factors. Among all tested models, the MLP achieved the highest predictive performance, with an average validation accuracy of 99.1% and equally strong Precision, Recall, and F1-score values (0.991). The XGBoost Classifier and SVM also performed competitively, achieving validation accuracies of 95.4% and 95.2%, respectively, demonstrating their ability to capture complex feature relationships. In contrast, Logistic Regression, serving as a linear baseline, achieved a modest validation accuracy of 73.6%, suggesting limitations in modeling nonlinear dependencies.

From a methodological standpoint, the inclusion of feature encoding, scaling, adaptive learning rate scheduling, and cross-validation contributed significantly to achieving reliable and generalizable results.

Visual tools such as confusion matrices and learning curves provided more profound insights into classification performance and model convergence behavior. Overall, the study highlights that deep learning (MLP) and ensemble learning (XGBoost) approaches are efficient for real estate prediction tasks, offering high accuracy, flexibility, and robustness against noisy data. Furthermore, features such as area, bedrooms, phase, and air conditioning consistently emerged as the most influential predictors of property value, reinforcing their real-world significance in determining housing prices.

The findings of this study have important implications for both academic research and the real estate industry. A properly trained and validated predictive model can assist:

Real estate agencies estimate market value ranges for new property listings. Buyers and

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 3 No. 12 (December) (2025)

sellers make informed decisions based on price category predictions, and urban planners and policymakers analyze housing trends and affordability across different localities. When integrated into an interactive platform, such as a mobile or web-based application, the model can provide real-time predictions of price categories based on user-input property details. This capability enhances accessibility, transparency, and data-driven decision-making within the housing market.

Limitations and Future Work

While the results are auspicious, certain limitations must be acknowledged:

The dataset comprised 1,300 records from a single geographic region (Hayatabad, Peshawar), potentially limiting the model's generalizability to other urban areas. The use of manually collected features may introduce minor inconsistencies or measurement biases. The study focused on categorical price prediction (Low, Medium, High) rather than continuous price estimation, simplifying interpretability but reducing precision for numerical forecasting. Addressing these limitations will help enhance the model's scalability, accuracy, and applicability across broader real estate markets. Future research can extend the current study in several meaningful directions:

Data Expansion and Automation: Expanding the dataset across multiple cities and integrating automated data collection (e.g., web scraping from real estate platforms) will improve model robustness and enable comparative regional analyses. Hybrid and Explainable AI Models: Combining ensemble models like XGBoost with deep learning architectures such as MLPs can leverage both hierarchical and nonlinear representations. The integration of Explainable AI (XAI) tools such as SHAP or LIME will enhance model interpretability for non-technical users.

Integration with Geospatial and App-Based Systems: Linking predictive models to Geographical Information Systems (GIS) or mobile applications (e.g., a Flutter-based housing app) enables location-aware predictions and interactive visualization. This integration will make the model more practical and user-friendly for real estate professionals and consumers. Continuous Learning and Real-Time Prediction: Developing an online learning framework that updates model parameters as new data becomes available will enable real-time market adaptation, ensuring sustained model accuracy over time.

Conflict of Interest: The authors declare there are no competing interests.
Acknowledgement: NA
Data Availability: The dataset used in this research paper is available from the corresponding author upon a reasonable request.

References

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.

Aljohani, O. (2021). Developing a stable house price estimator using regression analysis. Proceedings of the 5th International Conference on Future Networks and Distributed Systems,

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Armona, L., Fuster, A., & Zafar, B. (2019). Home price expectations and behaviour: Evidence from a randomized information experiment. The Review of Economic Studies, 86(4), 1371-1410.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 3 No. 12 (December) (2025)

Azimlu, F., Rahnamayan, S., & Makrehchi, M. (2021). House price prediction using clustering and genetic programming along with conducting a comparative study. Proceedings of the Genetic and Evolutionary Computation Conference Companion,

Baum, A. (2015). Real estate investment: A strategic approach. Routledge.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Fan, C., Cui, Z., & Zhong, X. (2018). House prices prediction with machine learning algorithms. Proceedings of the 2018 10th international conference on machine learning and computing,

Ghosalkar, N. N., & Dhage, S. N. (2018). Real estate value prediction using linear regression. 2018 fourth international conference on computing communication control and automation (ICCUBEA),

Lee, S.-H., Kim, J.-H., & Huh, J.-H. (2021). Land price forecasting research by macro and micro factors and real estate market utilization plan research by landscape factors: Big data analysis approach. Symmetry, 13(4), 616.

Li, R. Y. M., Song, L., Li, B., Crabbe, M. J. C., & Yue, X.-G. (2022). Predicting carpark prices indices in Hong Kong using AutoML.

Liu, R., & Liu, L. (2019). Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm: R. Liu, L. Liu. Soft Computing, 23(22), 11829-11838.

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House price prediction using regression techniques: A comparative study. 2019 International conference on smart structures and systems (ICSSS),

Manasa, J., Gupta, R., & Narahari, N. (2020). Machine learning based predicting house prices using regression techniques. 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA),

Maslow, A. (1943). A theory of human motivation. Psychological Review google schola, 2, 21-28.

Muellbauer, J., & Murphy, A. (2008). Housing markets and the economy: the assessment. Oxford review of economic policy, 24(1), 1-33.

Ogunleye, B. O. (2021). Statistical learning approaches to sentiment analysis in the Nigerian banking context. Sheffield Hallam University (United Kingdom).

Phan, T. D. (2018). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. 2018 International conference on machine learning and data engineering (iCMLDE),

Rahman, S. N. A., Maimun, N. H. A., Razali, M. N. M., & Ismail, S. (2019). The artificial neural network model (ANN) for Malaysian housing market analysis. Planning Malaysia, 17.

Rogers, J., & Gunn, S. (2005). Identifying feature relevance using a random forest. International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection",

Sanyal, S., Biswas, S. K., Das, D., Chakraborty, M., & Purkayastha, B. (2022). Boston house price prediction using regression models. 2022 2nd International Conference on Intelligent Technologies (CONIT),

Shiller, R. J. (2007). Understanding recent trends in house prices and home ownership. In: National Bureau of Economic Research Cambridge, Mass., USA.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 3 No. 12 (December) (2025)

Shuai, Y., Zheng, Y., & Huang, H. (2018). Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV. 2018 IEEE 9th international conference on software engineering and service science (ICSESS),

Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4), e12599.

Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, 103941.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.-L. A., Elkhatib, Y., Hussain, A., & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 7, 65579-65615.

Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using machine learning and neural networks. 2018 second international conference on inventive communication and computational technologies (ICICCT),

Wang, Y. (2021). House-price Prediction Based on OLS Linear Regression and Random Forest. 2021 2nd Asia Service Sciences and Software Engineering Conference,

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., & Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. Land, 11(3), 334.

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. Scientific Programming, 2021(1), 7678931.

Zhou, Y., Song, X., & Zhou, M. (2021). Supply chain fraud prediction based on xgboost method. 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE),

Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: a survey of literature. International Journal of Modern Education and Computer Science, 12(6), 46-54.