Exploring the Rental Market Dynamics of the Guadalajara Metropolitan Area

1st Cuauhtemoc Guerrero Ramírez
Masters in computer science
Tecnológico de Monterrey
Guadalajara, Jal.

2nd Diego Ramos García de Alba
Masters in Computer Science
Tecnológico de Monterrey
Guadalajara, Jal.

3rd Javier Cerriteño Magaña
Masters in Computer Science
Tecnológico de Monterrey
Guadalajara, Jal.

***Abstract*—The research focuses on analyzing the rental housing market in the Guadalajara Metropolitan Area by leveraging data from various real estate websites. The study aims to create a custom dataset through web scraping techniques and analyze it to understand factors influencing rental pricing and property listing duration. Key objectives include identifying significant features affecting rental prices and listing duration, predicting rental prices based on these features, and enhancing predictive models with new features. The research combines traditional data analysis with machine learning methodologies to uncover trends and provide insights for tenants and property owners, aiming to contribute to a more transparent and efficient rental housing market in Guadalajara.**

***Index Terms*—Rental housing market, Guadalajara Metropolitan Area, Web scraping, Property attributes, Data-driven decision-making, Pricing strategies, Predictive modeling**

I. Introduction

The contemporary housing market is influenced by geographical factors, socio-economic dynamics, and evolving consumer preferences. In this era of data-driven decision-making, leveraging insights extracted from real-estate sources such as websites is important for tenants and property owners. This research aims to provide an in-depth analysis of the rental housing market in the Guadalajara Metropolitan Area, by studying an extensive dataset encompassing listings of houses for rent.

The contribution of this research is twofold: First, we will implement web-scrapping techniques in order to create a custom dataset consisting of rental houses listings within the Guadalajara Metropolitan Area. Secondly, we will analyze this dataset to answer relevant questions such as optimal rental pricing and the importance of different features.

Our custom dataset includes features of each house listing such as location, neighborhood, square meters, number of bedrooms, number of bathrooms, commodities, neighborhood characteristics, temporal dynamics, etc. In order to create this dataset, we will explore popular real-estate websites such as inmuebles24.com, propiedades.com, casas.trovit.com.mx, lamudi.com.mx and vivanuncios.com.mx, and extract the meaningful features from this websites using web-scrapping techniques.

We will implement the use of machine learning models to extract meaningful information from this custom dataset. We implement a model that is able to estimate the optimal rental pricing of a house considering its features, using regression techniques. We also implement a model capable of estimating how long will a house remain listed for rental considering its relevant characteristics. We also explore what are the features that have a higher correlation with both rental pricing and listing duration. Finally, we explore whether the use of feature engineering is able to enhance the prediction accuracy of our model.

Our approach combines traditional data analysis with machine learning methodologies to extract meaningful information to provide an understanding of the factors shaping the rental landscape in Guadalajara. Through this in-depth investigation, our objective is to inform tenants and property owners about the most important factors that influence the rental housing market for them to make strategic decisions, optimize pricing strategies, and contribute to a more transparent and efficient rental property in the city of Guadalajara.

A. Objective and research question

To conduct a comprehensive analysis of the rental housing market in the Guadalajara metropolitan area with the following aims:

- What features have the most significant impact on rental prices?
- What features have the highest impact on the time a property spends on the market?
- Can we predict rental prices in Guadalajara based on the most significant features identified?
- Can we create new features that enhance the predictive power of our models, such as combining existing features or extracting relevant information?
- Can you predict the duration a property will stay on the market before being rented based on its features?

II. Related work

The exploration of rental market dynamics, as observed in various housing markets, sheds light on crucial aspects influencing housing affordability, investment decisions, and policy formulation. In a study analyzing the rent/price ratio across diverse English housing sub-markets, findings indicate a nuanced relationship between property types, neighborhood characteristics, and rental values [3]. This research underscores the significance of property attributes, such as the number of

bedrooms and proximity to amenities, in shaping rental market dynamics. Moreover, it suggests that disparities in wealth and neighborhood conditions contribute to variations in the rent/price ratio, with implications for housing affordability and market stability. Similarly, a study focusing on housing market segmentation in Shanghai, China, emphasizes the importance of considering temporal dynamics and functional proximity in sub-market delineation [5]. By integrating hedonic models, geospatial analysis, and machine learning techniques, the study reveals distinct trajectories in sales and rental submarkets, underscoring the need for differentiated policy interventions [2]. Additionally, efforts to forecast rental market trends using system dynamics modeling offer promising insights into the feasibility of predicting market movements and informing decision-making for stakeholders [8]. Such endeavors contribute to a comprehensive understanding of rental market dynamics, facilitating evidence-based strategies for addressing housing challenges and promoting sustainable urban development.

III. Method and Data

A. Dataset creation

For the dataset creation process, we devised a Python web scraping script to navigate through the initial 50 pages of house rental listings and another 50 pages for apartment rentals. The script saved the HTML code of each page into .txt files, which were then processed using regular expressions (RegEx) to extract crucial property details such as location, zone, rental price in MXN, number of rooms, number of bathrooms, and area in m^2, we used this code for extraction in various real-estate websites to obtain the most values possible.

Following this data extraction phase, we segregated the obtained information into separate data frames for houses and apartments, appending the respective property labels "House" or "Apartment" to each entry. After undergoing initial cleaning and value redistribution procedures, we merged these individual data frames into a unified dataset. This consolidated dataset, encompassing both house and apartment listings, was exported as a CSV file for preliminary validation, analytical exploration, and method implementation to address our research inquiries.

Key variables within our dataset include the property's location, rental price, number of rooms, number of bathrooms, area, address, colony, and property type.

B. Exploratory Data Analysis (EDA)

We initiated our exploratory data analysis by importing our raw data and ensuring label consistency. This involved identifying unique values for each column in our dataset. As part of enhancing clarity, we opted to rename the 'Colony' column to 'Municipality' and 'Address' to 'Neighborhood'. Notably, upon observing 'Jalisco' entries in the municipality column, we generalized them to 'Other' since all municipalities are situated within Jalisco.

Following these adjustments, we conducted a check for duplicate entries in our dataset. Instances of duplication, often stemming from properties advertised on multiple websites, were promptly removed to maintain data integrity.

Our initial data visualization endeavors included a pie chart (Fig. 1) illustrating occurrences by municipality. Similarly, to portray the distribution of the number of rooms across municipalities, we employed a histogram (Fig. 2).

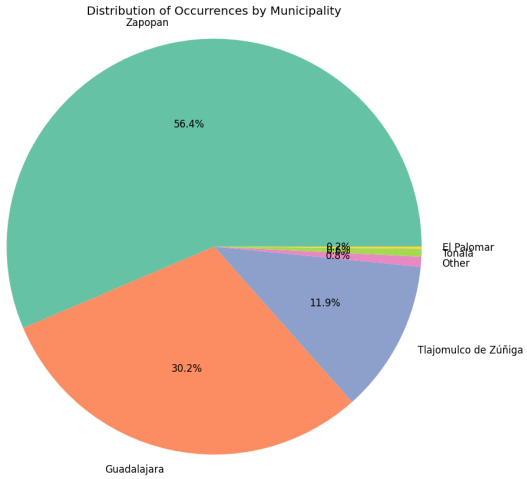

Fig. 1. First visualization of the distribution of Occurrences by Municipality

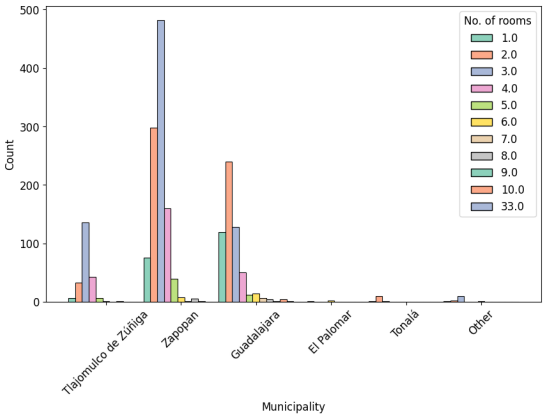

Fig. 2. Visualization of the initial distribution of the number of rooms by Municipality

Additionally, we utilized a boxplot to visualize the price distribution per number of rooms. However, the presence of outliers, markedly deviating from the median, hindered effective visualization and analysis. Hence, we opted to remove outliers exceeding 1.5 times the interquartile range (IQR) and re-ran our visualizations (Figs. 3, 4, 5), furthermore, a boxplot (Fig. 6) was employed after the removal of outliers to illustrate the price distribution based on property type.

To tackle missing values, notably in the number of rooms and bathrooms, we implemented interpolation techniques. Our approach involved sorting the dataset by area and interpolating the number of rooms and bathrooms according to their area. Subsequently, we filled any missing values at the beginning or end of the data frame with the first or last available value.

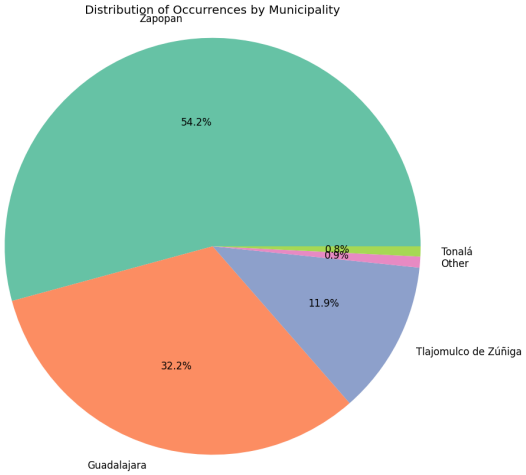

Fig. 3. Visualization of the distribution of Occurrences by Municipality after removal of outliers

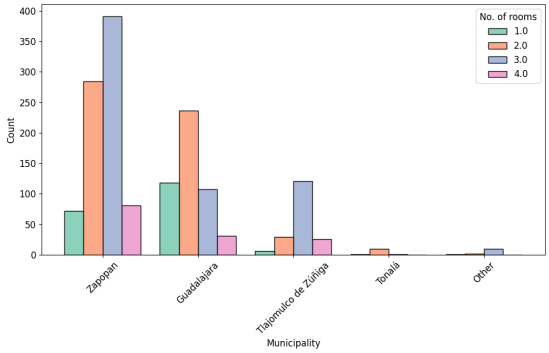

Fig. 4. Visualization of the distribution of number of rooms by Municipality after removal of outliers

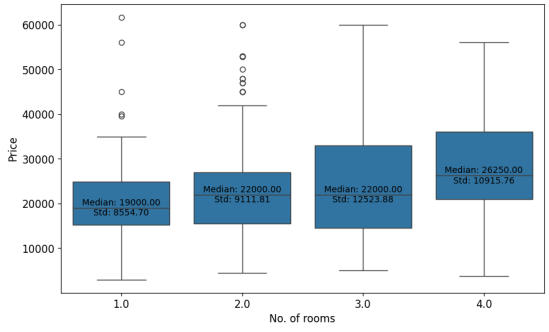

Fig. 5. Visualization of the distribution of price per number of rooms after removal of outliers

To provide a visual representation of the effects of interpolation, we utilized a boxplot (Fig. 7). This visualization aids in assessing the efficacy of the interpolation method in approximating missing values and maintaining the integrity of the dataset.

Moreover, we generated a correlation matrix (Fig. 8) to explore relationships between variables and a scatter plot (Fig. 9) depicting the relationship between area and price colored

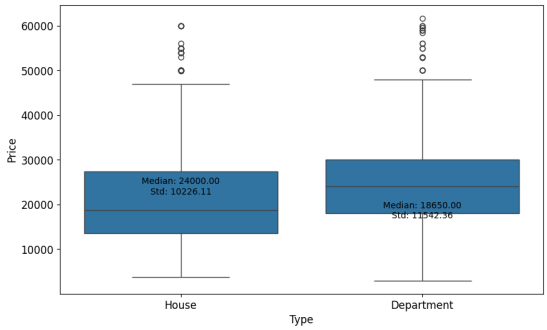

Fig. 6. Visualization of the distribution of price by type of residency

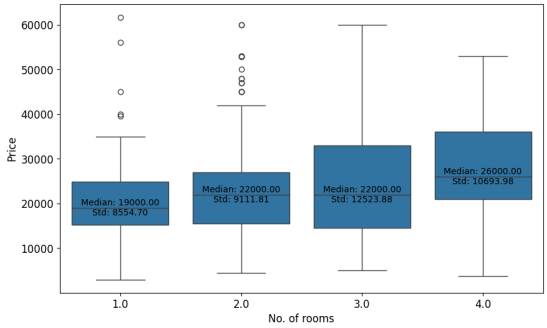

Fig. 7. Distribution of price by No. of rooms visualization after interpolation on missing no. of rooms value

by Municipality.

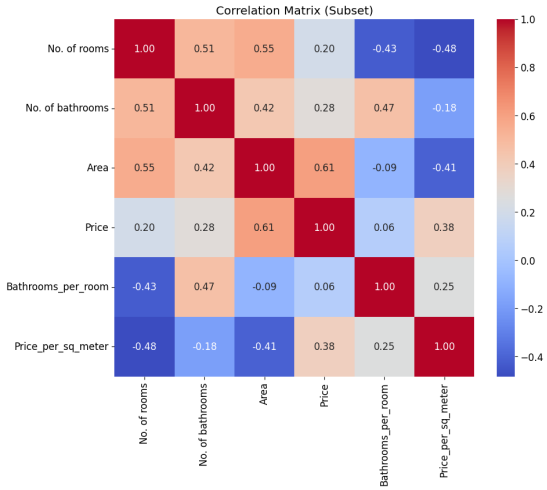

Fig. 8. Correlation Matrix

C. Statistical analysis

We begin by examining the distribution of our data to ascertain its normality, a crucial step in understanding the underlying characteristics of our dataset. This assessment is pivotal because it informs the choice of appropriate statistical methods and models, as well as the reliability of subsequent insights derived from the analysis. In our analysis, we focus

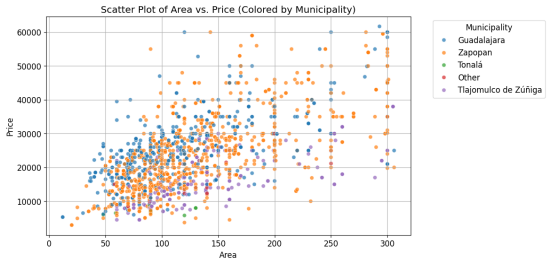

Fig. 9. Scatter Plot of Area vs. Price (Colored by Municipality)

particularly on the distribution of price data, recognizing its pivotal role as the primary variable we aim to predict through subsequent analytical methods.

To evaluate the normality of the price data distribution, we employ the Shapiro-Wilk Test, a widely utilized statistical test that assesses whether a given sample comes from a normally distributed population [10]. The results of the Shapiro-Wilk Test for our dataset yield a test statistic of 0.9525908827781677and a corresponding p-value of $p-value : 8.145374820270653 * 10^{-}22$, With a conventional significance level α of 0.05, given that the calculated p-value is significantly smaller than α, we reject the null hypothesis, which posits that the data follows a normal distribution, this implies that there is insufficient evidence to support the assumption of normality in the distribution of our price data and as a result, we acknowledge the non-normal nature of the data distribution of the target variable price.

We then perform a 1-way ANOVA (Analysis of variance) test, a statistical technique used to determine whether there are statistically significant differences in the means of a continuous variable, depending on the different values of a categorical feature [9]. We perform this test to determine whether the variables "No. of rooms", "No. of bathrooms", "Municipality" and "Type" have a significant impact on the distribution of values of the target variable "Price". This test proposes a null hypothesis that states that there is no difference in the means of the target variable, depending on the different possible values of a single factor. If the p-value is less than the significance level α of 0.05, we can reject the null hypothesis and state with 0.95 confidence that the means of the "Price" variable are different depending on the different values of the feature being considered. If we conclude that the means are different after performing this test, we can conclude that the considered feature has a statistically significant impact on the target variable (difference values of this feature are associated with significant differences in the target variable). Therefore, it would be a good idea to consider these features as good candidates for the prediction of house prices. We can visualize the value distribution of the "Price" variable, depending on the values of the variables "No. of rooms", "No. of bathrooms", "Municipality" and "Type". We can observe this distribution density in the figures 10,11,12 and 13

Once the distribution of our data is visualized, we proceed to analyze the variation in price means across the mentioned

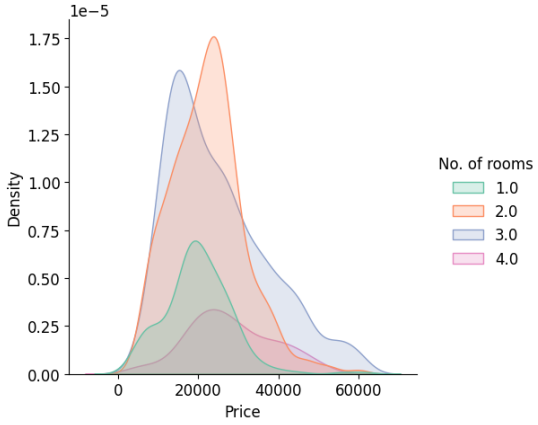

Fig. 10. Mean price by no. of rooms

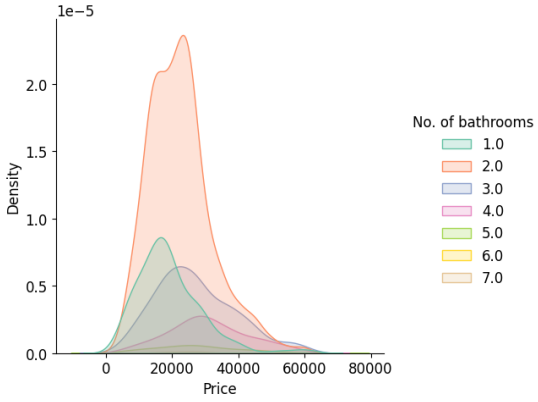

Fig. 11. Mean price by no. of bathrooms

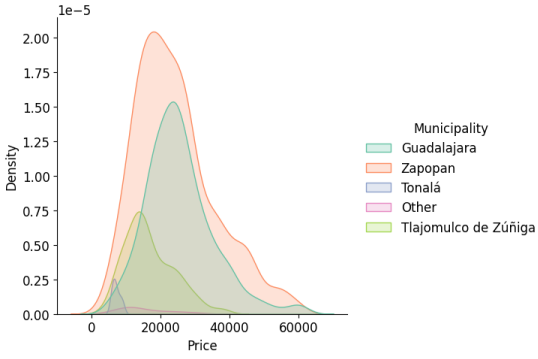

Fig. 12. Mean price by muunicipality

factors using the one-way ANOVA test. The results of the one-way ANOVA test are presented in Table I.

TABLE I
RESULTS FROM THE 1-WAY ANOVA TEST

	Rooms	**Bathrooms**	**Municipality**	**Type**
Statistic	21.18	25.35	33.45	33.25
P-Value	1.90e-13	8.14e-29	8.84e-27	9.77e-09

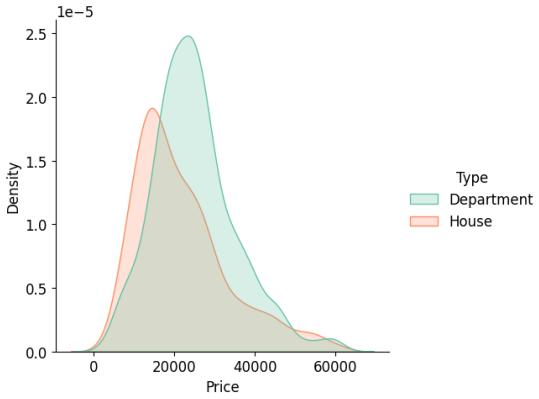

Fig. 13. Mean price by type of property

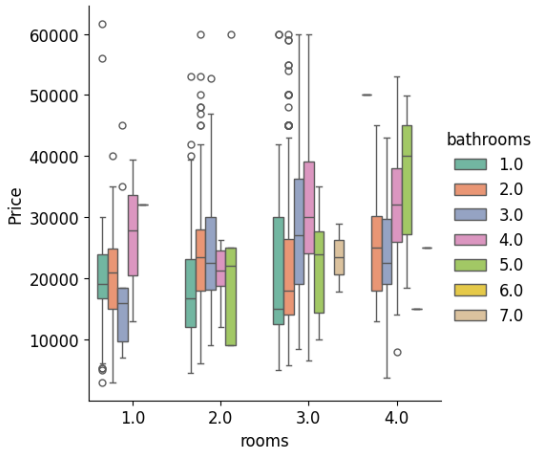

Fig. 14. Mean price by number of rooms and number of bathrooms

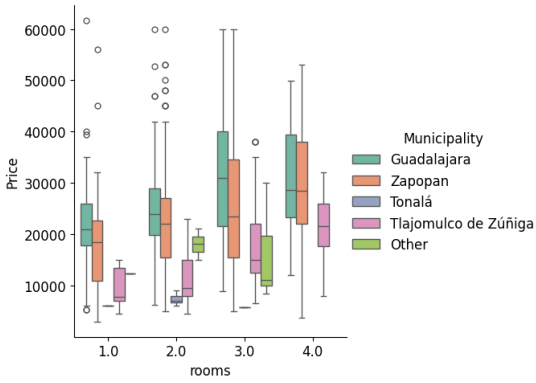

Fig. 15. Mean price by number of rooms and Municipality

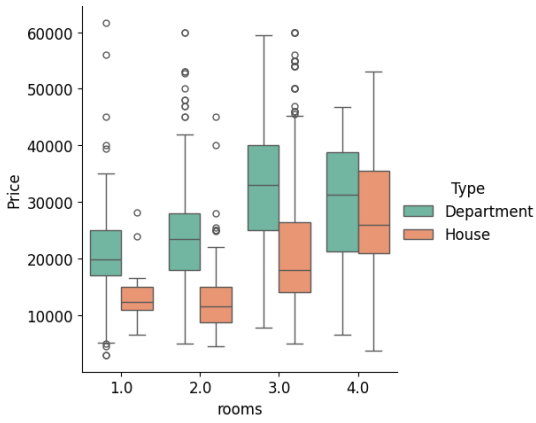

Fig. 16. Mean price by the number of rooms and property type

The statistic column represents the F-statistic calculated for each factor, while the P-value column indicates the probability of observing the given F-statistic if the null hypothesis were true. In this context, the null hypothesis assumes that there is no difference in price means across the categories being compared.

Based on the results from the one-way ANOVA test, we observe extremely low p-values for all factors (Rooms, Bathrooms, Municipality, and Type), indicating strong evidence against the null hypothesis. Consequently, we reject the null hypothesis and conclude that there are significant differences in price means across different values of each factor. In other words, at least one category within each factor has a different average price compared to the others. Therefore, we can conclude that each of these variables (number of rooms, number of bathrooms, municipality, and property type) has a statistical impact on the value of the target variable (price), and thus would be good candidates to be used as features for a prediction model to estimate the pricing of houses.

For further analysis we proceed to do the 2-way ANOVA test, for which we select 2 variables to use with the price to run the test, we end up making 6 tests with the data groups, the 1st one being mean price by the number of rooms and the number of bathrooms (fig. 14), 2nd one being mean price by number of rooms and Municipality (fig. 15), 3rd one mean of price by number of rooms and property type (fig. 16), 4th one mean price by number of bathrooms and municipality (fig. 17), 5th one being mean of price by number of bathrooms and property type (fig. 17) and lastly, the 6th one being mean price by municipality and property type (fig. 19). The distribution of the values of the price variable can be observed in these figures, depending on the different values of the two features considered for each box plot.

The 2-way ANOVA test is performed to determine the significance of the impact of two features on the target variable, as well as whether the combined effect of the two variables (their interaction) has a significant impact on the target variable, beyond the individual effect of each feature [7]. The two-way ANOVA obtains the significance of each variable by determining whether the mean of the target variable changes depending on the values of each feature, similar to the one-way ANOVA. However, it also determines whether there is a significant combined effect of both features on the target variable. Since we already performed one-way ANOVA for the variables of interest, we will focus on the interaction

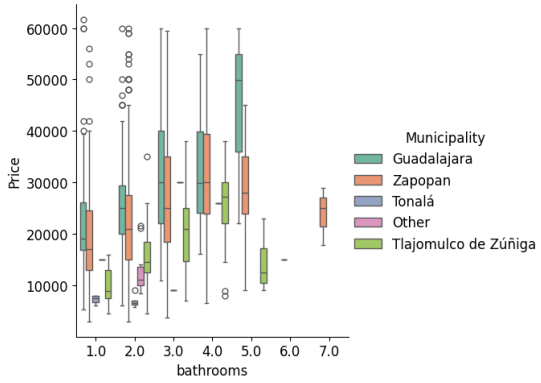

Fig. 17. Mean price by number of bathrooms and municipality

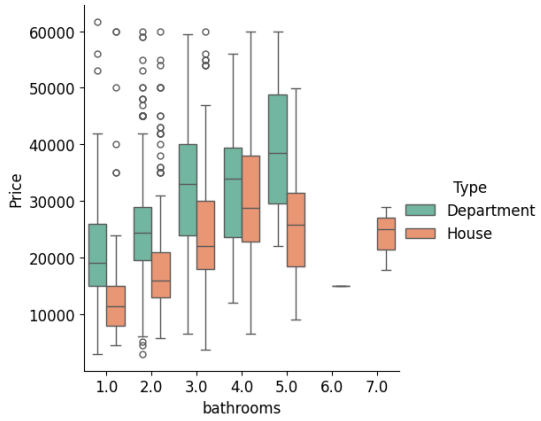

Fig. 18. Mean price by number of bathrooms and property type

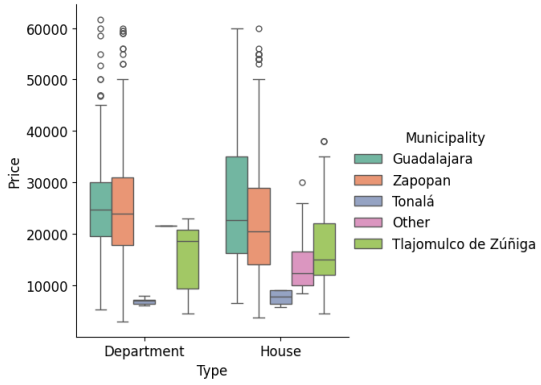

Fig. 19. Mean price by Municipality and property type

effect of two variables at a time. The null hypothesis for the interaction effect in two-way ANOVA is that there is no significant interaction effect between the two features on the target variable. That is, the impact of one feature on the target variable does not depend on the values of the other feature. If the p-value obtained by this test is lower than the significance level of 0.05, we can reject the null hypothesis and determine that there is a significant combined effect of both features on the target variable beyond their individual effects. We can observe the results of the 2-way ANOVA performed for each pair of features in table II.

TABLE II
RESULTS FROM THE 2-WAY ANOVA TEST

Features	P-Value
Rooms and Bathrooms	5.92e-10
Rooms and Municipality	4.88e-01
Rooms and Type	3.18e-02
Bathrooms and Municipality	1.02e-68
Bathrooms and Type	1.02e-05
Municipality and type	4.59e-01

The p-value (5.92e-10) for the combination of "Rooms and Bathrooms" is very low, indicating a statistically significant effect. This suggests that the interaction between the number of rooms and the number of bathrooms has a significant impact on the target variable. The p-value (4.88e-01) for the combination of "Rooms and Municipality" is above the significance level of 0.05 which suggests that there is no significant interaction effect between the number of rooms and the municipality on the target variable. The p-value (3.18e-02) for the combination of "Rooms and Type" is below 0.05, indicating a statistically significant effect. The p-value (1.02e-68) for the combination of "Bathrooms and Municipality" indicates a significant effect of the interaction of these features on the target variable. The p-value (1.02e-05) for the combination of "Bathrooms and Type" below 0.05 suggests that the interaction between the number of bathrooms and the type of property has a significant impact on the price. Finally, The p-value (4.59e-01) for the combination of "Municipality and Type" indicates no significant interaction effect between the municipality and the type of property on the target variable. Considering this analysis, we can conclude that the pairs of the features Rooms and Municipality and the features Municipality and Type do not have a significant interaction effect on the value of the house prices, so maybe it is not necessary to consider them together in the same model as predictive features to estimate the target variable.

D. Method

We will implement and evaluate three regression models to predict property prices in Zona Metropolitana de Guadalajara. The models selected for this analysis are multiple linear regression, random forest regression, support vector machine regression and Elastic net linear regression. Each model will be trained using a combination the following features: the number of rooms, number of bathrooms, area, neighborhood, municipality, type of property, price per square meter, and bathrooms per room. Following training, we will perform a comparative analysis to assess the predictive performance of these models. Evaluation metrics including mean squared error (MSE), R-squared, and mean absolute error (MAE) will be used to compare the accuracy and effectiveness of each model in predicting property prices.

*1) **Multiple Linear Regression**:* Multiple linear regression is a simple and effective regression model that can be used

for predicting housing prices [6]. In this model, we consider a linear relationship between the target variable (price) and the multiple independent variables mentioned earlier. The model's training involves fitting a linear equation to the training data using gradient descent, with weights representing the impact of each feature on the property price. The goal of training is to minimize a cost function, in this case, the mean squared error (MSE). We will evaluate the model's performance using the metrics mean squared error (MSE), R-squared, and mean absolute error (MAE) on the testing dataset. The interpretability of the coefficients allows us to understand the relative importance of each feature in determining property prices.

*2) **Random Forest**:* Random forest regression is a robust ensemble learning technique that combines multiple decision trees to make predictions. Each decision tree in the forest is trained on a random subset of the data and features, resulting in diverse trees that collectively provide more accurate predictions [1]. In the context of predicting property prices, random forest regression can capture non-linear relationships and interactions between features effectively. During training, the algorithm constructs an ensemble of decision trees, where each tree is grown to minimize the variance of predictions. The final prediction is obtained by averaging the predictions of all trees in the forest. Evaluation of the random forest regression model involves assessing its performance on the testing dataset using metrics such as mean squared error (MSE), R-squared, and mean absolute error (MAE). The advantage of random forest regression lies in its ability to handle high-dimensional datasets and noisy data while providing reliable predictions. Additionally, it offers feature importance scores, allowing us to identify the most influential features in determining property prices.

*3) **Support Vector Machine** :* Support Vector Machine (SVM) regression is a powerful supervised learning algorithm that can be used for regression tasks [4]. SVM regression aims to find the hyperplane that best fits the data points while maximizing the margin between the hyperplane and the points. For predicting property prices, SVM regression seeks to find the hyperplane that separates the feature space into different regions corresponding to different price levels. During training, the algorithm adjusts the parameters to minimize the error between the predicted and actual property prices. SVM regression is particularly effective in high-dimensional spaces and when there is a clear margin of separation between different price levels. Evaluation of the SVM regression model involves assessing its performance on the testing dataset using metrics such as mean squared error (MSE), R-squared, and mean absolute error (MAE). SVM regression offers flexibility in choosing different kernel functions to capture complex relationships between features. Additionally, it can handle both linear and non-linear relationships between the input features and the target variable, making it suitable for predicting property prices in Zona Metropolitana de Guadalajara.

*4) **Elastic Net Linear Regression** :* Elastic net linear regression is a regularized regression method that combines the penalties of both Lasso (L1) and Ridge (L2) regression techniques [11]. This hybrid approach helps to manage the trade-off between simplicity and overfitting by encouraging sparsity while also maintaining the benefits of regularization. In the context of predicting property prices, Elastic net regression can effectively handle multicollinearity among the features and select relevant variables, which is particularly useful when dealing with high-dimensional datasets.

E. Data Grouping

Before training our models, we made several adjustments to the data to improve the training and validation processes. These modifications aimed to enhance the model's performance and make it more cost-effective by reducing the parameter combinations to test, ultimately leading to better results. The following are the modifications implemented:

1) Division by Municipality: We divided the data based on municipalities to ensure that the model can capture localized patterns and variations. This segmentation allows the model to focus on specific characteristics unique to each municipality, enhancing its predictive accuracy.

2) Removal of Outliers by Municipality: Outliers can significantly skew the results of a model. Therefore, we identified and removed outliers within each municipality. This step ensures that the data used for training is representative of typical conditions, thereby improving the robustness and reliability of the model.

3) Encoding by Neighborhood: To incorporate categorical data into our models, we employed encoding techniques specific to neighborhoods. This process transforms neighborhood-related information into a format that can be effectively used by machine learning algorithms, helping the model to understand and utilize these categorical variables.

F. Applying Grid Search

To determine the optimal parameters for our models, we employed a grid search to explore the hyperparameter space systematically. This involved evaluating various combinations of estimators and their respective combinations of parameters and values, specifically, we tested:

- **Linear Regression**: Grid search is not used for the linear regression model.
- **Support Vector Regression (SVR)**: Grid search was performed to find the best values for the regularization parameter C and the margin of tolerance ϵ where no penalty is given to errors, with C values from 0.1 to 10 and ϵ values from 0.1 to 0.001.
- **Random Forest Regression**: Grid search was performed to find the best values for the number of trees in the forest $n\_estimators$ and the maximum depth of the trees max_depth, with $n\_estimators$ varying from 50 to 200 and max_depth values from None to 20.
- **Elastic Net**: Grid search was performed to find the vest values for the constant α that multiplies the penalty terms, controlling regularization strength, and the ElasticNet mixing parameter $l1\_ratio$ which controls the balance

between L1 and L2 regularization, with α values from 0.0001 to 1.0 and $l1\_ratio$ values from 0.1 to 0.9.

This methodical approach ensured a comprehensive examination of potential configurations, allowing us to identify the settings that yield the best performance. We used 5-fold cross-validation and employed Mean Absolute Error (MAE) as the scoring metric to evaluate and compare the models' performance. This rigorous process led to the development of a well-tuned model.

IV. Results

A. Applying Grid Search

After performing a grid search, the best results for the hyperparameters for each model were as follows:

- **Support Vector Regression (SVR)**: C = 10 and ϵ = 0.1 for Guadalajara, C = 10 and ϵ = 0.001 for Zapopan, C = 10 and ϵ = 0.001 for Tonala, and C = 10 and ϵ = 0.1 for Tlajomulco.
- **Random Forest Regression**: max_depth = none and $n\_estimators$ = 50 for Guadalajara, max_depth = none and $n\_estimators$ = 200 for Zapopan, max_depth = 20 and $n\_estimators$ = 50 for Tonala, and max_depth = 20 and $n\_estimators$ = 50 for Tlajomulco.
- **Elastic Net**: α = 0.0001 and $l1\_ratio$ = 0.1 for Guadalajara, α = 0.01 and $l1\_ratio$ = 0.9 for Zapopan, α = 0.01 and $l1\_ratio$ = 0.9 for Tonala, and α = 0.01 and $l1\_ratio$ = 0.9 for Tlajomulco.

B. Model Performance

The performance of the four regression models—Multiple Linear Regression, Support Vector Machine (SVM) Regression, Random Forest Regression, and Elastic Net Linear Regression—was evaluated across four municipalities in the Guadalajara Metropolitan Area: Guadalajara (GDL), Zapopan (ZPN), Tonalá (TNL), and Tlaquepaque (TLJ). The models were assessed based on their mean absolute error (MAE) and R-squared (R^2) metrics for both training and testing datasets. The results are summarized in the table below.

Finally, when analyzing the data from the rented houses and departments after a month, we got the following results displayed in table IV when looking for the average price, area in m^2 no. of rooms, and type of property for each municipality

V. Discussion

The results indicate that Random Forest Regression generally outperformed the other models across all municipalities in both the training and testing phases. Specifically, it achieved the lowest MAE and the highest R-squared values, suggesting a superior fit and predictive capability.

Guadalajara (GDL): Random Forest achieved an MAE of 1588.20 and an R-squared of 0.93 on the training data, and an MAE of 4276.60 with an R-squared of 0.67 on the testing data. This performance indicates that Random Forest is able to capture complex patterns in the data, resulting in more accurate predictions.

TABLE III
Performance metrics for regression models across different municipalities

Municipality	Models	Training		Testing	
		MAE	R-Squared	MAE	R-Squared
GDL	Linear Regression	3461.90	0.73	4377.93	0.65
	SVR	6911.56	0.04	7463.10	0.10
	Random Forest	1588.20	0.93	4276.60	0.67
	ElasticNet	3592.89	0.73	4355.10	0.67
ZPN	Linear Regression	3450.48	0.81	3954.55	0.76
	SVR	8605.52	0.04	8503.58	0.05
	Random Forest	1545.96	0.96	3930.99	0.73
	ElasticNet	3743.39	0.80	3926.61	0.78
TNL	Linear Regression	0	1	353.34	0.83
	SVR	616.42	0	1225	-0.40
	Random Forest	260.67	0.84	509	0.78
	ElasticNet	0.3	1	319.53	0.86
TLJ	Linear Regression	1577.81	0.89	2417.50	0.69
	SVR	5485.75	-0.02	4516.31	-0.01
	Random Forest	887.84	0.97	2376.81	0.65
	ElasticNet	1634.01	0.89	2390.02	0.69

Municipality	**Price (MN)**	**No. of rooms**	**Area (m^2)**	**Type**
Guadalajara	25000	2	90	Department
Zapopan	25000	3	80	House
Tlajomulco de Zuñiga	15000	3	160	House
Tonala	7000	2	90	Department
Other	10000	3	89	House

TABLE IV
Rented property data by municipality

Zapopan (ZPN): Similar trends were observed with Random Forest showing the best performance with an MAE of 1545.96 and an R-squared of 0.96 on training data, and an MAE of 3930.99 with an R-squared of 0.73 on testing data.

Tonalá (TNL): Here, Elastic Net also showed strong performance with an MAE of 0.3 and an R-squared of 1 on training data, and an MAE of 319.53 and an R-squared of 0.86 on testing data. Random Forest was also very competitive with an R-squared of 0.84 on training data and 0.78 on testing data.

Tlaquepaque (TLJ): Random Forest again led the performance with an MAE of 887.84 and an R-squared of 0.97 on training data, and an MAE of 2376.81 and an R-squared of 0.65 on testing data. Elastic Net also performed well with an R-squared of 0.69 on the testing data.

Support Vector Machine (SVR) Regression consistently underperformed across all municipalities, indicating it may not be suitable for this specific application of property price prediction.

Multiple Linear Regression and Elastic Net Regression provided reasonable performance, with Elastic Net showing better generalization ability due to its regularization properties.

VI. Conclusion

This study aimed to predict property prices in the Guadalajara Metropolitan Area using multiple linear regression, random forest regression, support vector machine regression, and elastic net regression models. Through comprehensive data analysis and model evaluation, several key conclusions have been drawn:

The analysis revealed that several factors significantly influence property prices. The number of rooms in a property was found to be a crucial determinant, with strong differences in price means across different room categories confirmed by the ANOVA tests. Similarly, the number of bathrooms significantly affects property prices, demonstrating substantial statistical differences. The municipality in which the property is located also plays a vital role in determining its price, highlighting the importance of geographical context in property valuation. Additionally, differences in property types, such as houses versus apartments, showed considerable impact on pricing.

Among the models evaluated, random forest regression consistently outperformed the other models in predicting property prices across different municipalities. This model achieved the lowest Mean Absolute Error (MAE) and the highest R-squared values, indicating its superior fit and predictive capability. In certain municipalities like Tonalá, the elastic net regression also showed strong performance, particularly in training data, demonstrating its effectiveness in capturing the underlying data patterns.

The findings from this study provide valuable insights for tenants, property owners, and real estate professionals. Understanding the significant factors influencing property prices can aid in strategic decision-making, optimize pricing strategies, and ultimately contribute to a more efficient market. The application of predictive models enhances transparency in the property market, offering more accurate price estimates and benefiting all market participants by improving information symmetry.

Future research could explore additional features impacting property prices, including more granular neighborhood characteristics, temporal dynamics, and other property attributes. Advanced feature engineering techniques could be employed to enhance the predictive power of the models, such as creating new features from existing data or extracting relevant information from external sources.

REFERENCES

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Rohit Singh Chauhan. *Applying System Dynamics to Simulate and Forecast Rental Real Estate Market*. PhD thesis, Massachusetts Institute of Technology, 2024.

[3] Stephen Clark and Nik Lomax. Rent/price ratio for english housing sub-markets using matched sales and rental data. *Area*, 52(1):136–147, 2020.

[4] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. 9, 1996.

[5] Lirong Hu, Shenjing He, and Shiliang Su. A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. *Computers, Environment and Urban Systems*, 94:101775, 2022.

[6] Gülden Kaya Uyanık and Neşe Güler. A study on multiple linear regression analysis. *Procedia - Social and Behavioral Sciences*, 106:234–240, 12 2013.

[7] Hae-Young Kim. Statistical notes for clinical researchers: Two-way analysis of variance (anova)-exploring possible interaction between factors. *Restorative dentistry endodontics*, 39:143–147, 05 2014.

[8] Eddie Chi man Hui and Xian Zheng. The dynamic correlation and volatility of real estate price and rental: an application of msv model. *Applied Economics*, 44(23):2985–2995, 2012.

[9] Eva Ostertagova and Oskar Ostertag. Methodology and application of one-way anova. *American Journal of Mechanical Engineering*, 1:256–261, 11 2013.

[10] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[11] Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005.