NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

Forecasting models for real estate market analysis
The case study of Portugal

Ana Lúcia Barata Coutinho

Project Work

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

FORECASTING MODELS FOR REAL ESTATE MARKET ANALYSIS

THE CASE STUDY OF PORTUGAL

By

Ana Lúcia Barata Coutinho

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervisor(s):

João Bruno Morais de Sousa Jardim

Miguel de Castro Simões Ferreira Neto

November 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Ana Lúcia Barata Coutinho]

[Lisbon, November 22nd, 2023]

ACKNOWLEDGEMENTS

This research would not have been possible without the incredible support of countless people who guided me through the process, shared their knowledge and answered all of my questions.

I would like to thank my family for believing in me and it was because of their emotional support that they kept my motivation high throughout the project. Thanks to my friends and data scientists who provided advice and feedback on the many challenges I had to overcome.

I am especially thankful to my two co-supervisors Miguel de Castro Neto and João Bruno Jardim and NOVA IMS for providing materials and support in the essential context that helped enrich my knowledge and this research.

ABSTRACT

The real estate market is known for its difficulties and volatility, particularly in determining property values, which significantly impact economic choices and societal well-being. Accurate predictions of property prices can facilitate informed decision-making and influence investment strategies and market dynamics.

This study focuses on predicting property prices in Portugal and looks at different regions with a focus on Lisbon. It attempts to use innovative methods such as Decision Tree, Random Forest, Artificial Neural Networks, Linear Regression, and K-Nearest Neighbors.

The research provides an in-depth evaluation of forecasting techniques by analyzing historical data and employing various machine learning algorithms. The models are specifically designed to understand the complex dynamics of different property types in different regions of Portugal. Overall, the Random Forest model is the best model for predicting property prices, followed by Decision Tree model. This forecast is also important because it provides us with an insight into the importance of certain features in real estate property prices.

Using advanced predictive models does not only enhance the understanding of real estate market trends but also enables stakeholders to make informed decisions.

KEYWORDS

Real Estate, Real Estate Property Prices, Machine Learning, Random Forest, Artificial Neural Network, K-Nearest Neighbors, Decision Tree.

Sustainable Development Goals (SGD): The study is aligned with Sustainable Development Goal 9: Industry, Innovation, and Infrastructure, employing innovative methods such as Decision Tree, Random Forest, Artificial Neural Network, Linear Regression, and K-Nearest Neighbors. More specifically, this thesis aims to produce knowledge to the technological advancement of the real estate field, by integrating real estate data to understand and forecast price dynamics. In such a complex and crucial domain, this research thus provides one more contribution to understand the real estate property market prices and addressing its current challenges.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
CFBP	Cascade Forward Back Propagation Neural Network
EEC	Energy Efficiency Certificate
FFBP	Feed Forward Back Propagation Neural Network
GBA	Gross Building Area
GDP	Gross Domestic Product
KNN	K-Nearest Neighbors
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
RMSE	Root Mean Square Error

1. INTRODUCTION

Real estate emerges prominently as a pivotal asset class that merits serious consideration within the realm of investment opportunities, as highlighted by recent research (Cocola-Gant & Gago, 2021). Having a home is an important part of our basic needs, just as Maslow's Hierarchy (Mcleod, 2018) proposes a five-level model of the conscious drivers of human behavior. This model is presented in the form of a pyramid, with the lowest level being the first to be satisfied. The physiological level is the level that must be satisfied first. It is located at the bottom of the pyramid, and it is at this level that shelter is included. We see that the fulfillment of the physiological level is mandatory to pass through the other levels of the hierarchy.

It is not just a matter of owning a home, but some characteristics are necessary to fulfill this need. It should be remembered that the definition of happiness is not universal, and the desired characteristics of a house do not apply to all buyers and will change over time and in life circumstances.

Real estate prices in Portugal have increased exponentially in recent years and have the potential to become a real estate bubble as prices are higher than anticipated by economic and financial authorities. Due to the wide fluctuations in real estate prices, it is difficult to predict what will happen in the real estate market. Many variables can predict the price of a particular house and indicate its characteristics based on historical sources. A lot of factors should be considered when analyzing the market, with particular emphasis on detailed location (Cellmer & Trojanek, 2020). All consumers and potential business owners must examine the model's list of variables and subjectively evaluate each variable's proposed business value (Robert N, 1995).

In 2022, as a result of high inflation (*Inflation in the Euro Area*, 2023) and to minimize the damage, banks decided to increase the interest rate, Euribor, which led to an increase in housing loans. This phenomenon directly affects household incomes and could lead many of them to sell their homes because they cannot afford the increased cost of housing loans. In the third quarter of 2022, real estate prices in the Portuguese real estate market increased by 13.1%, according to Eurostat data (*Housing Price Statistics - House Price Index*, 2023). Unaffordable costs also mean that demand for housing is low and house prices could stabilize or start to fall. This event is associated with inflation, which reduces purchasing power and increases real estate loans.

Additionally, the wider context of the real estate market in Portugal must be acknowledged. During the last decade, the Portuguese real estate landscape has undertaken major transformations, marked by increased demand, significant price increases, an increased prevalence of short-term rentals, and the emergence of new and fresh trend in real estate investment. This shift raises questions about the sustainability and stability of the market, requiring in-depth analysis and informed decision-making by potential buyers and investors (Cunha & Lobão, 2021). In light of these developments, this study attempts to shed light on the complex dynamics of the Portuguese real estate market. By studying historical data and identifying key factors affecting property prices, we aim to provide valuable insights into the future development of the market. This research not only addresses the pressing concerns of property buyers but also contributes to the broader discourse on real estate economics and investment strategies in an evolving market.

The objective of this study is to forecast real estate property prices, by comparing the effectiveness of the specific features of a dwelling and evaluating the prices of houses in Portugal according to the different characteristics. This project aims to deliver a different view of the real estate market. Distinguish the uncertainties that make the process fluctuate, contributing to greater certainty when thinking about buying or selling a house. The main purpose of this study is to examine and compare the performance of different real estate price property prediction models. It covers three different models for property price prediction: Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). The analysis is based on extensive real estate data covering several regions in Portugal, with a focus on the Lisbon area. Standard metrics such as R^2, MAE (Mean Absolute Error), and MAPE (Mean Absolute Percent Error) are used to evaluate and compare the accuracy of price prediction models. These metrics provide insight into each model's ability to predict actual property prices.

One of the questions is to know whether the inclusion of geographic information affects the accuracy of the model forecasts. The guiding research questions for this study are: “How can data eliminate bubbles in real estate markets?”, “How much will a house be worth in the future?” and “What is the next trend zone in Lisbon?”.

To respond to the questions above, developed research was made based investigating historical trends and searching the factors that contribute to the fluctuation of real estate property prices. Then, the historical data on real estate prices in Portugal was analyzed to make forecasts.

The diagram below was created to illustrate and simplify the complex sequence of processes performed in this study. This diagram is intended to illustrate the steps involved in data collection, preprocessing, model implementation, and evaluation. The purpose behind this visualization is to provide a clear and structured explanation of the methods used to predict real estate property prices.

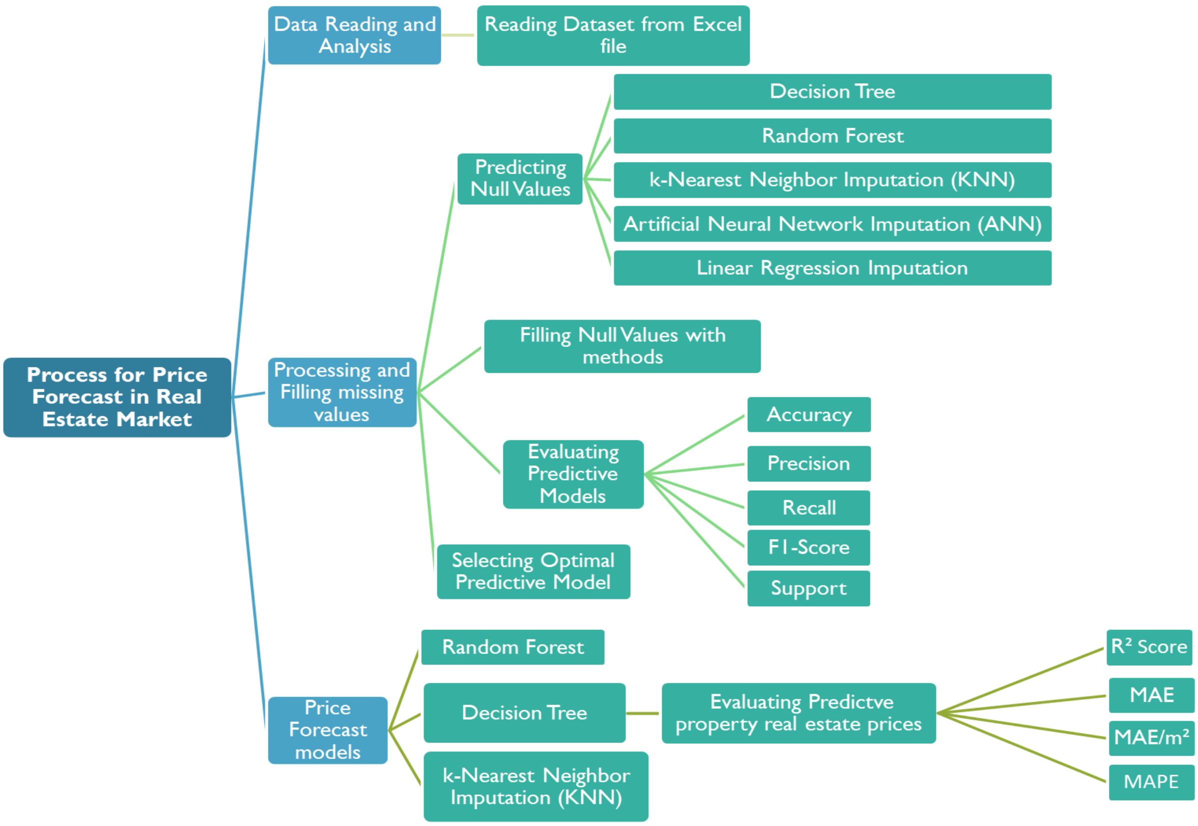

Figure 1 - Diagram that illustrates the process that was conducted in this study

2. LITERATURE REVIEW

The real estate market is experiencing a sharp rise in prices, and many studies are being conducted to predict what will happen in the future. The desire to buy a home of one's own is widespread among the population and is considered the first step into adulthood. For a buyer or seller, the ability to predict price fluctuations would improve and change the entire market.

To get an overview of the real estate market in Portugal, (Oikarinen, 2014) stated the price disparities among regions in the housing market could bring diversification benefits not only for a housing investor but also for the nation as a whole. This statement reflects the inequality of housing prices between the different regions of Portugal.

Investigators developed studies in two main directions: first, to examine the determinants of housing prices, and second, to explore the factors that link asset prices, macroeconomic developments, economic and regulatory policies, corporate governance, and the legal system.

2.1. MACROECONOMIC

In the context of the real estate market, macroeconomics refers to the global economic factors that influence real estate property prices and trends on a national or global level. These include economic variables that play an important role in the housing market and the economy as a whole. Fluctuations in real estate prices can affect a country's economy, as the total value of real estate often represents a substantial part of Gross Domestic Product (GDP) (Tripathi, 2019). In addition, the real estate market can be a source of financial risk, as housing bubbles and credit crises can affect the banking system. Access to housing is another macroeconomic concern, as property prices directly impact quality of life and wealth distribution in society (Goodhart & Hofmann, 2008).

The study conducted by (Baffoe-Bonnie, 1998) analyses the effects of four macroeconomic variables: mortgage interest rate, consumer price index, changes in employment, and money supply on house prices. It concludes that macroeconomic data are crucial for real estate forecasting and that these variables have different effects on the dynamic behavior of house prices and the number of houses sold in different periods and that these variables alone cannot explain price fluctuations.

There is a link between the real estate market and a country's macroeconomy, meaning that what affects one also affects the other. For example, during times of economic instability and crisis, real estate markets may also experience imbalances. Hence, this link between macroeconomics and the real estate market is intertwined with macroeconomic factors such as GDP, inflation, interest rate, and money supply (Dehesh & Pugh, 2000). Price volatility is influenced by mortgage rates, inflation, and employment rates, making inflation a weighing factor for buyers and thus a motivating factor for investors. (Akinsomi et al., 2018) research also concludes that the performance between real estate and macroeconomics and capital markets are correlated, they have a positive relationship. Another author who studies macroeconomics and its impact on real estate markets is (Kasparova & White, 2001), who concludes that real estate prices are driven by income growth and interest rates. In addition, the correlation between inflation and the macroeconomic values can be seen through construction costs and the money supply, these two variables are positively correlated, they move in the same direction, when one increases, the other also increases, this means an increase in the real estate prices (Alkali et al., 2018). The unemployment rate is another indicator associated to macroeconomics and since affect income will have an impact in the buying power, (Grum & Govekar, 2016) state that real estate price is drastically linked with the unemployment rate.

The factors affecting this market can be categorized as qualitative or quantitative. (Ahn et al., 2012) considered macroeconomic factors in their study (as we have already seen: unemployment rate, GDP, inflation, interest rates, etc.), real estate attributes such as current and past sales prices, construction area, floor area, and year of housing construction, categorized into quantitative elements. Qualitative information tends to refer to factors that are more difficult to measure and are subjective, such as architectural style, the way a home is landscaped, or the way a home divides up space. Therefore, qualitative information appears descriptive rather than numerical, making it difficult to convert these qualitative factors into quantitative factors to obtain metrics. Quantitative methods use time series data to determine the relationship between current and prevailing prices. As we have seen, qualitative factors are difficult to quantify and therefore are not the best way to conduct this study. Subsequently, this research is based on quantitative data from actual transaction data.

As we have noted, real estate market is linked to the entire current social and economic environment of each country and is intertwined with the global economy, so changes in the dynamics of the main factors identified affect the entire market and create vulnerabilities in prices and in the economy. Therefore, a stable macroeconomic environment will drive the expansion of the real estate market (Alkali et al., 2018).

2.1.1. Bubbles

A real estate bubble is a temporary phase characterized by high demand, low supply, and price appreciation that exceeds fundamentals. This market is characterized as a slow market, therefore never achieves the price equilibrium point, the prices are always oscillating, creating bubbles (Pilinkienė et al., 2021). These bubbles are caused by a combination of factors that include increasing economic prosperity, low interest rates, increased supply of mortgage products, and easier access to credit. The laws of supply and demand apply. Real estate prices are determined by supply and demand, and when demand increases and supply is limited, prices rise.

The best-known real estate bubbles are the Japanese real estate bubble in the 1980 and the U.S. real estate bubble in early 2000. Researchers study the many reasons why bubbles in real estate market happen and the factors that influenced prices oscillations. Economic crisis can be predicted and prices in real estate market can been protected. (Fama, 1991) argues that: is not possible to predict or avoid bubbles.

These bubbles that appear have two stages, first the boom effect and then the burst effect. The first is the period where prices increases and the bursts the reverse. (Crowe et al., 2012) found that monetary policy and macro-financial instruments have a significant direct impact on real estate market performance. These factors (such as limits on loan to value ratio) often fuel real estate cycles because most real estate activity, whether investment or acquisition, involves borrowing.

The bubble effect has an important role in predicting real estate instabilities, (Brzezicka, 2020) saw that thanks to the bubble effect policies and regulations in one area provoke increasing house prices in other areas.

(Agnello & Schuknecht, 2011) concluded that geographers play an important role in identifying bubbles, which can help to understand whether geographers can play an important role in identifying bubbles at an early stage to prevent them from forming and affecting the economy. Since not all geographic regions experience the same price inflation, the need for housing bubbles can also influence policy interventions and help stem these bubbles and regulate housing prices. Geography is therefore of great importance to the way financial markets are structured and regulated.

2.1.2. Tourism

Another factor contributing to the volatility of real estate property prices is tourism. The (*World Travel Awards 2022 Europe Winners*, 2022) recognize and reward all major industries around the world, including tourism. In 2022, Portugal won several awards in Europe. For the study of this project, it was important to highlight two awards, firstly, Portugal's award as "Leading European Destination 2022" and secondly, the city of Lisbon's award as "Europe's Leading Urban Destination 2022". Despite all the benefits and tourist attractions that come with these awards (Biagi et al., 2015), they argue that tourism has a positive impact on real estate prices. But there is also the other side of the coin, where tourism affects the economic costs, it imposes on countries and cities that depend on tourism (Ahmad et al., 2020). (Barron et al., 2021) have found an impact on rental prices through the Airbnb Impact Study, which is influenced by tourism activity.

In the long run, tourism has a positive impact on economic growth (Alhowaish, 2016) because it improves a country's foreign exchange earnings, increases employment opportunities, promotes incremental development, enables new technologies, etc.

Since Lisbon is a metropolitan area, whose main activity is tourism, it is more likely to respond to higher real estate price growth than metropolitan areas with less tourism activity. Many researchers have studied real estate property prices to see if tourism factors influence this market (Schäfer & Hirsch, 2017) and found that real estate prices in Europe are associated with tourism factors. Does this mean that more tourism activity leads to more price volatility? (Badarinza & Ramadorai, 2018) research shows that foreign capital is the main reason for excess returns in real estate prices in some European cities. Since foreigners are investing heavily in houses all over the world and in Portugal, buying them and selling them at a higher price, making a profit, this will cause house prices to rise, foreigners have more purchasing power and allow them to make these transactions, earning extra money.

The government of each country must ensure that price increases are controlled. Portugal has taken some steps in this regard by limiting the allowance of new tourist housing (Airbnb) and increasing income tax for EU citizens buying property in Portugal. However, tourism is one of the drivers of economic growth and development in the country and contributes to social and cultural change. Therefore, tourism in Portugal is linked to the real estate market and contributes to the Portuguese economy.

2.1.3. Crime

The value of real estate is also influenced by factors in the surrounding area, with security and crime playing an important role in the search for a residence. (Foryś & Putek-Szeląg, 2017) they investigated whether there is a relationship between crime rates and real estate prices in Poland. The study led them to conclude that geographic areas with lower values appeared in areas with higher crime rates. They also examine the effects of gentrification, a process that attracts high-income households and businesses, resulting in higher incomes and education levels and lower crime rates among this population. Thus, an increase in housing prices is associated with a decrease in crime. "The decline in safety coincides with the deterioration of the living conditions of the population but also reduces trust in the police and judicial system" (Foryś & Putek-Szeląg, 2017, p. 52).

Crime reduces the demand for real estate in that geographic area, which reduces the area's popularity and prices. Buyers are willing to pay more to live in areas with lower crime rates so they can enjoy greater safety and a better quality of life (Buonanno et al., 2013).

2.2. PRICE FORECAST

There are many studies that use different forecasting methods. We will analyze some of them and their results to determine the best method. Some studies are done using micro variables, as in our case. Variables such as number of bedrooms, number of bathrooms, size of the house, whether it has a garage, a balcony, or even a swimming pool are included in this analysis.

Furthermore, real estate property prices are affected by many characteristics, some of which are related to the characteristics of a specific home. The most common variables for the model studied by (Wing & Chin, 2003) include attributes such as type, age of the building, number of bedrooms and bathrooms, facilities (e.g. swimming pool), garage, patio, gross area, and building services (e.g., air conditional, lift, etc). The application of hedonic price to real estate markets identifies a set of key variables to be included in the model, showing that the effects of one or more real estate attributes on real estate prices can be observed, while other factors being constant. (Fletcher et al., 2000) conducted a study with the aim of determining the list of variables to be included in a hedonic pricing model and found that, the number of bedrooms, bathrooms and gross area can be positively related to its price. In the research of (My-Linh, 2020) revealed that the age of the building can impact property real estate prices negatively.

To estimate property prices (Pai & Wang, 2020) also included variables such as distance to the city center and proximity to the subway in his predictions, and used many methods (Decision Tree, Back Propagation Neural Networks, General Regression Neural Networks and Least Squares Vector Regression) for predict property prices. They conclude that it is possible to predict the prices using these methods and that the method that most accurately reflects real estate property prices is least squares vector regression. (Liu, 2022) research analyses the real estate market applying Multiple Linear Regression Model, the study shows a strong correlation between the income levels with real estate prices versus, the models demonstrate a high accuracy with lower prediction error, indicating its practicality in forecasting real estate market trends.

Random Forest is a decision tree-based ensemble model that uses a technique called bootstrap aggregation (or bagging) to create multiple independent regression models from random samples of the training data set. It is a versatile method that can handle both regression and classification tasks. The main process involves building various decision trees by randomly selecting data points and features during training, which ultimately helps improve model accuracy (Ja'afar et al., 2021). In a comparative study by (Čeh et al., 2018), the Random Forest model was compared with Linear Regression to predict real estate prices. The results show that Random Forest consistently performed better than Linear Regression model on a variety of performance metrics. Furthermore, Random Forest have lower error rates in predicting real estate prices compared to other algorithms (Shinde & Gawande, 2018).

In a study conducted by (del Cacho, 2010), various models such as Naïve Neighbors, Linear Regression, Decision Trees, K-Nearest Neighbors (KNN) and Neural Networks were used to predict real estate prices. A point to highlight related to this study is the fact that Naïve Neighbors is a classifier that assumes that all variables are independent of each other. The results were evaluated based on various indicators and it was found that the Decision Tree model provided the best indicator performance in terms of real estate prices.

Artificial Neural Network (ANN) have a wide range of uses, forecasting is one of the purposes so many researchers decided to apply this technique. It is able to address problems related to nonlinearity and multicollinearity. One of the advantages of this model is its versatility, it has the ability to model relationships between dependent and independent variables. (Ghodsi et al., 2010) in their research of estimation of Iranian house prices, in order to form an Artificial Neural Networking, considered economic variables including, the housing price index, GDP, and cost of construction materials, the result was a successful estimation of the prices of the variables. (Tabales et al., 2013) did forecasting for the real estate market and show that ANN has better prevision results than econometric models. The study conducted by (Rahman et al., 2019) also using ANN to forecast housing prices, concluded that ANN has the potential to forecast house prices with high performance and also confirmed the findings made by (Tabales et al., 2013), the ANN model achieved better results when tested in a large dataset. (Bahia, 2013) used two types of ANN: Feed Forward Back Propagation Neural Network (FFBP) and Cascade Forward Back Propagation Neural Network (CFBP) to predict real estate prices. The difference between these two types is that the second method, CFBP, includes a weighted connection from the input to each layer and successive layers. The conclusion of this research is that the CFBP method shows results that are closest to real prices since it is possible to adjust the weights, the results were more accurate, this study achieved more than 96% accuracy in predicting real estate prices.

Other methods are also used by more researchers, such as Multiple Linear Regression, which we have previously validated (Grum & Govekar, 2016) to see the relationship between house prices and macroeconomic variables in different regions and cultures (Greece, Poland, etc.), and it turns out that there is a strong relationship between the variables studied, namely the unemployment rate and house prices, which accurately illustrates the macroeconomic impact on this market. Using macroeconomic variables, GDP, house price index, consumer index, (Li & Chu, 2017) and with the forecasting methods Back Propagation and ANN, it is concluded that these variables are not strong enough to predict house prices. Based on the final price of several houses over 4 years, (Singh et al., 2020) used three different methods: Random Forest, Gradient Boosting and Linear Regression, in order to predict house prices, the conclusion is that all forecasting models were able to predict prices but the model with higher forecasting accuracy was the Gradient Boosting.

One problem that most of the previous studies have is the fact that they don't take into account the economic situation over time, so, their studies need to be updated frequently otherwise they became outdated. (Yasnitsky et al., 2021) saw this as an opportunity and developed a mass valuation for real estate in Russia based on fixed factors, location, and variable factors, economic parameters. The result of the model elaborated was that houses in different geographical locations react differently to the economic changes where they are inserted, the volume of loans, and the GDP of the country. This model is very different in that it can easily adapt to the economy and the constant changes that exist.

(Hishamuddin et al., 2020) saw that not many studies combine microeconomic and macroeconomic variables, so they decided to consider both in their paper to have a vision in a short and long term for property prices, the method chosen was ANN. Using R^2 in error measurement as a basis, they conclude that the percentage of error is less when R^2 has a high value when only macroeconomic variables were considered. When only microeconomic variables were used alone the results were really weak, and when both variables were combined, the results were better than using microeconomic variables alone but worse than using macroeconomics variables alone. Therefore, it is necessary to study this combination of variables in more detail to avoid value discrepancies and to obtain a more reliable price prediction. With this study, it is possible to see that there is a gap in research and an opportunity to explore the combination of these two macro variables. We need to consider both the micro variables described above and the macroeconomic variables such as inflation, unemployment rate, interest rate, etc.

2.2.1. Comparative analysis of predictive models

The real estate market is influenced by several quantitative and qualitative variables, making price forecasting a complex task. To address this challenge, various predictive models have been used to fill in the null values that we have in the data and to estimate real estate prices. This research aims to examine and compare several predictive models, including Artificial Neural Network (ANN), Linear Regression, Decision Trees, K-Nearest Neighbor Imputation (KNN), Random Forest, and Gradient Boosting. Each of these models has unique strengths and limitations that make it suitable for different scenarios. By evaluating and comparing these models against each other, we aim to provide valuable insights into their performance and effectiveness. Table 1 lists various forecasting models, each with a brief description, strengths, and limitations. Among the authors who studied predictive models, those who are mentioned are part of the research carried out that gave rise to the table.

Predictive Model	Description and Definition	Strengths	Limitations
ANN (Hush & Horne, 1993)	It's a type of Neural Network algorithm that consists of interconnected nodes or neurons organized into layers, usually including input, hidden, and output layers. These networks take input data and predict an output based on it. During training, neural networks adjust the connections' weights to learn	• Capable of handling non-linear and complex data. • Suitable for big data applications due to their ability to model intricate relationships.	• Requires a large training dataset for optimal performance. • Prone to overfitting if not carefully regularized.

	complex patterns and make predictions.		
Linear Regression (Montgomery et al., 2021)	Is a statistical model that establishes a linear relationship between predictor variables and the target variable. It assumes a linear combination of features and finds coefficients that minimize the error between predicted and actual values.	• Simple and easy to interpret, making it a good choice for initial insights. • Wide applicability to a extend of problems.	• Method limited to linear relationships between variables, which may not capture complex patterns. • Sensitive to outliers and multicollinearity.
Decision Trees (de Ville, 2013)	Are constructed by algorithms that explore different ways of splitting the data into distinct segments. They make decisions based on feature splits, creating a tree where each node represents a decision. The tree is split based on the feature that provides the best information gain.	• Easily interpretable models, as they mimic human decision-making processes. • Can handle both numerical and categorical data.	• Tendency to overfit with deep trees, which requires pruning or ensemble methods. • Sensitive to small variations in the training data.
KNN (Weinberger & Saul, 2009)	KNN is a non-parametric algorithm that predicts a target variable based on the majority class of its k-nearest neighbors in a feature space. It determines the target value by considering the labels of these neighbors. Prediction with KNN requires define the chosen variables, the distance metric to calculate the distance between observations and the number of neighbors.	• Ability to adapt to complex data patterns, especially when the number of neighbors (k) is appropriately chosen. • Performs well on small datasets and with few features.	• Sensitive to the choice of 'k', which can significantly affect results. • Is influenced by outliers in the data, leading to inaccurate predictions.
Random Forest (Oshiro et al., 2012)	Random Forest is an ensemble method that combines several Decision Trees. This model is particularly known for its ability to effectively handle a large number of features and	• Reduces overfitting through ensemble averaging. - Handles missing data effectively	• The relationship between the response and the independent attributes is

	efficiently process extensive datasets. Also improves prediction accuracy and reduce overfitting by aggregating prediction from multiple decision trees.	without preprocessing.	complicated to understand. • Tend to have longer training times compared to other methods, so is less suitable for real-time applications or very large datasets.
Gradient Boosting (Hastie et al., 2009)	Is a method that sequentially constructs an ensemble of weak learners (usually decision trees) to create a strong predictive model. It builds trees sequentially, focusing on the mistakes made by previous trees. This method gained popularity due to the superior performance of Decision Trees and led to the development of algorithms such as XGBoost and LightGBM.	• High accuracy and effectiveness across many scenarios. • Handles missing data effectively.	• Can be sensitive to outliers in the training data. • Prone to overfitting, especially when the algorithm is allowed to create complex models.

Table 1 - Comparative analysis of predictive models

3. METHODOLOGY

The objective of this study is to build a property price forecasting model using predictive models. Property price forecasts are designed to help homebuyers understand future price ranges and plan their finances in advance. In addition, property price forecasts are also beneficial for real estate investors to understand the trend of property prices in a particular region. This study refers to Portugal Continental and was a focus on Lisbon since Lisbon is the capital of Portugal. Lisbon is where there are more data and exist a large difference in price variation and where the differences between cultures, trends and lifestyles are most pronounced, where is a lot of business activity and tourism, therefore this allows for a comprehensive analysis with abundant data.

To make accurate forecasts, we need suitable mathematical models to estimate how time series data generate house prices. In this context, CRISP-DM method was adapted to this situation, as discussed by (Shearer, 2020). Analyzing current trends and predicting future trends is becoming increasingly important today. Predictive models are gaining increasing value in organizations that face the challenge of choosing the right technology for the business and understanding how to use existing data (Nwagwu et al., 2021).

The real estate market is known for its volatility, characterized by fluctuations over time and continually changing trends. Given the complexity of the real estate market, which often exhibits non-linear relationships, it becomes necessary to leverage historical data and employ machine learning methods for analysis (Han, 2018). Machine learning offers the advantage of automatically learning patterns and insights from historical data, making it well suited for handling the difficulties of the real estate market.

For the construction of the prediction model, the programming language selected was *Python*. It is the most used language worldwide occupying the first place in October 2023[1], and is a user-friendly interface as it is easy to understand and has a wide variety of libraries available, which helps to increase of those uses this language saving programming time that would be manual.

In line to ensure that the information is accurate in this context, the data were subject to treatment and rules were defined. Before starting to develop the model, it is necessary to process data that involves importing files in CSV format, Exploration and Data Understanding, Data Pre-processing, Feature Selection, Modelling and Performance Assessment.

3.1. DATASET DESCRIPTION AND MODIFICATIONS

The study began by importing data from two Excel files into Jupyter. One of these files contained the year of construction of the properties, while the other file held a wide range of variables that constitute the dataset, such as the type of property (house or apartment), number of bedrooms, energy efficiency certificate, gross building area, property's condition, the presence of garage, swimming pool, terrace and patio, contained as well variables geographic related, such as parish number, postal code, latitude and longitude, and lastly information linked to real estate market prices, the last market appearance date and the selling price observed. To verify if a property had multiple values in the construction year variable, inconsistencies were checked, and each property ID was assigned a unique construction year.

[1] TIOBE Index - TIOBE

A merge operation was conducted between two Excel files to create a unified dataset. The data for this study consisted of properties up to the construction year of 2024 and there are historical price sales between 2017 and 2022.

Geographical information, such as parish number and postal code, allowed for the augmentation of the dataset with additional information. This data was derived from two external files[2]. The following columns were added: district, zone, municipality, and parish name, along with the rural/non-rural designation. Various transformations were applied to these new variables to make them usable. The properties have been classified into zones and rural and non-rural according to the municipalities, there are five zones covering the dataset and Portugal, the division is done according to the characteristics of the zones and municipalities, so we will have properties in the same zone more similar to each other[2].

Two specific variables that had no significant value to the dataset and only added complexity were removed, such as latitude and longitude, as their information was found to be redundant since they provided information that could be derived from the postal code variable. Latitude and longitude represent geographic coordinates and are useful when working with map or geocoding services. However, in this case, the latitude and longitude do not change for the same postal code, which is actually incorrect, so we consider removing these two variables and keeping the dataset concise.

Regarding the year of construction, the dataset compilation revealed several inconsistencies and nonsensical years, including the year 1050. As a corrective approach, all properties with a construction year prior to 1755 were removed from the dataset. This decision was influenced by historical events, notably the devastating earthquake that struck Lisbon in 1755, leading to widespread destruction by tsunami and fire. Given the severity of the damage, it was assumed that none of the houses constructed before that year remained habitable. Consequently, these properties were excluded from the analysis. The removal of houses constructed before 1755 and the handling of inconsistencies, such as the year 1050, ensured data quality and reliability in the subsequent analysis.

Several transformations and modifications were made to simply the data. Start by transforming the property ID variable into numerical format to facilitate further transformations. The property condition variable was transformed into numerical values ranging from 1 to 4. The rural variable was converted into a binary format, with 0 indicating a rural location and 1 representing a non-rural location. Further analysis revealed that the EEC (Energy Efficiency Certificate) variable could take on nine distinct letters that significantly different EEC levels. To simplify this variable, it was grouped into four distinct classes: Class 1 includes certificates A and A+, Class 2 encompasses certificates B and B-, Class 3 consists of certificates C and D, and Class 4 comprises the remaining certificates E, F, and G.

The modified dataset consists of 326537 rows representing the number of properties and includes 26 variables that will facilitate house price prediction modeling. Below, in table 2, is presented all the 26 variables selected variables from the modified dataset, which will enable us to build the most accurate predictive model. Is also present the two extra variables during the forecast.

[2] https://www.gpp.pt/images/Estatisticas_e_analises/Estatisticas/associadasmedidasapoio/Territorios_Rurais.pdf
http://www.proder.pt/ResourcesUser/Documentos_Diversos/33/PDRc_Freg_ZRurais_NUTIIs_rev2_corrigido.pdf

Variable	Description	Data Type
Property ID	The unique identifier of the properties	Object
Property Type	Property type. 1 if is an apartment and 2 for a house	Integer
Nº Bedrooms	Number of bedrooms	Integer
EEC	Energy Efficiency Certificate: Number 1 includes certificates A and A+, Number 2 certificates B and B-, Number 3 consists of certificates C and D, and Number 4 comprises the remaining certificates E, F, and G.	Integer
GBA	Gross Building Area	Integer
Parish	Parish	Object
Postal Code	Postal Code	Object
Garage	Indicates whether the property has a garage or not	Bool
Terrace	Indicates whether the property has a terrace or not	Bool
Pool	Indicates whether the property has a swimming pool or not	Bool
Patio	Indicates whether the property has a patio or not	Bool
Property Condition	The state of conservation of the house. It can be classified as good, bad, new, or used. For each state was assigned a number	Integer
Conservation State Verified	Indicates if the conservation state was provided by clients or verified with photos	Bool
Last Day in Market	Date of the last market update	Datetime
Initial Offer Date	Date of the initial offer	Datetime
Observed Selling Price	The selling prices that were observed by the company	Integer
Construction Year	Year of construction	Integer

District	District	Object
Region	Indicates one of the five regions into which Portugal was divided (North, Center, Lisbon, Alentejo and Algarve)	Object
Municipality	Municipality	Object
Parish Name	Name of the parish	Object
Last Market Update Year	Year of the last market update	Integer
Initial Offer Year	Year of the initial offer	Integer
Rural	Indicates whether the property is located in a rural area or not	Bool
Decade	Decade of construction (e.g., 50's)	Object
Century	Century of construction (e.g., 20)	Integer
Predicted Selling Price	The selling prices that were predicted by the different models	Float
Square meter Price Difference	Price per square meter. Difference between observed and predicted prices	Float

Table 2 - Variables Description

Since this dataset contains six variables with a lot of missing values, further transformations were performed to impute the data, considering the correlation between the variables. Based on literature review, the following models were selected: Decision Tree Imputation, Random Forest Imputation, K-Nearest Neighbor Imputation (KNN), Artificial Neural Network Imputation (ANN), and Linear Regression Imputation. The model chosen for each variable was the one that gave the best results compared to all the models studied, in terms of performance metrics (further explained in the next section 3.2).

3.2. Model Evaluation Metrics

In this project, a variety of evaluation metrics have been employed to assess the performance of different models, both in filling missing values and predicting property prices. These metrics serve as essential tools to gauge the effectiveness of the models.

To evaluate the performance of data imputation models, we focus on the metrics presented in table 3, based on the classification results.

Metric	Definition	Interpretation of the result
Precision	Measures the proportion of correctly imputed values (true positives) out of the total imputed values.	A high precision score indicates that the method is adept at accurately filling in missing data.
Recall	Quantifies the ratio of correctly imputed values (true positives) to the total actual missing values.	A higher recall score implies that the method effectively captures missing data.
F1-Score	The F1-score strikes a balance between precision and recall, offering a single metric to assess overall performance. It combines both precision and recall into a unified score, making it valuable for assessing the overall quality of data imputation.	A higher F1-Score indicates a better balance between precision and recall, meaning a better performance predicting missing values.
Support	Represents the number of instances for each class, reflecting the distribution of missing data points across different categories.	A higher support indicates that the model's predictions are more reliable and trustworthy in impute missing values.

Table 3 - Explanation of the metrics to fulfill missing values

These metrics are fundamental for evaluating the performance of data imputation models in accurately filling missing values. Additionally, they guide us in choosing the most effective approach to ensure the reliability and completeness of our dataset, which, in turn, contributes to more accurate predictions of house prices.

For the prediction of property prices, it is important to evaluate the performance of each model developed using different methods. The emphasis lies on six metrics: Accuracy, Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and R^2 Score. Before using these five evaluation indicators, it is necessary to understand their meanings and consider their interpretation within the specific context of this project. In table 4 is represented an explanation of the metrics to evaluate each of the models.

Metric	Definition	Interpretation of the result
Accuracy	Measures the proportion of correctly classified observations relative to the total.	The higher the percentage, the better the performance of the model.
MAE	Measures the difference between two continuous variables, translated into the average of the errors of a set of forecasts.	A lower value indicates that the models is capable to predict with accuracy the values.

MSE	Represents the difference between the observed values used to develop the model and the observed values predicted by the model. It comes from a more consistent metric because it is used to evaluate how well the model fits the data.	A lower value means that the quality of the model is higher, and the model makes predictions more accurate and closer to the true value.
RMSE	Is the square root of the MSE. Plots the average error between model predictions and actual values.	The closer the value is to zero, the better the model performs.
MAPE	Measures prediction error.	The lower the value, the more accurate the model. MAPE <= 10% - Strong precision < MAPE <= 20% - Good precision 20% < MAPE <= 50% - Moderate precision MAPE > 50% - Poor precision
R^2 Score	The R^2 Score is a measure that tells us how well the dependent variable predicts the variations in independent variables.	Values close to 1 indicate that a high percentage of the variance in the dependent variable can be predicted from the independent variables, indicating a good model fit. An R^2 score close to 0 means that the independent variables do not effectively explain the variance in the dependent variable, indicating a weak model fit.

Table 4 - Explanation of the metrics selected for evaluation of each of the models

3.2.1. Data Imputation of the missing values

Starting to fulfill the missing values by implying some specific rules related to the nature of this research. These rules refer to the year of construction, the type of house (apartment or house), the area where the house is located, and some specific features (pool, garage). The complete set of applied rules is outlined in table 5.

Rules
Houses with a total area of less than $100m^2$ do not have swimming pool or garage
Swimming pools are not present until the year of 1960
Apartments located in rural area do not have swimming pool

Apartments with either zero or one bedroom and total area of less than 50m² do not feature a terrace
Apartments constructed before the year of 1960 do not have garage
Houses situated in the region of Algarve and Alentejo have both a terrace and a patio
Houses located in the Algarve region and constructed after the year of 2000 have swimming pool and garage
Apartments constructed after the year of 2000 feature garage
If the Energy Efficiency Certificate (EEC) rating is above E (E, F or G), there is no swimming pool
Apartments with an area exceeding 150m² have terrace
If the state of conservation is new, have garage
When the state of conservation is new then the Energy Efficiency Certificate (EEC) is 1 (meaning A+ or A)

Table 5 - Explanation of the rules applied

The models mentioned in chapter 3.1, were employed to address missing values in the six variables that contained missing values: Energy Efficiency Certificate (EEC), garage, swimming pool, patio, terrace, and construction year. For each variable and using the methods to impute null values, an analysis was conducted. Based on the metrics, the best model was chosen and will be applied to the final dataset.

To better understand the data, correlations between variables were examined and Spearman coefficients were considered. Therefore, garage variable is analyzed first. This variable has a strong correlation with other two variables, terrace, and patio, which also contain null values, and the null values in these two variables were filled using the mode. For all the variables in analysis, the Decision Tree Imputation presented to be the one with better results individually in each variable, so this was the method selected for all the variables.

To handling missing values the methods: Decision Tree, Random Forest, KNN, ANN, and Linear Regression were utilized. Table 6 presents the evaluation metrics: precision, recall, F1-score, and support, demonstrating the outcomes of the most effective method among them for the respective variable.

Method	Variables	Predicting	Accuracy	Precision	Recall	F1-Score	Support
Decision Tree	Garage	False	**0.81**	**0.81**	**0.96**	**0.88**	**45774**
		True		**0.83**	**0.48**	**0.61**	**19534**
	Pool	False	**0.97**	**0.98**	**0.99**	**0.99**	**63352**

		True		**0.56**	**0.22**	**0.31**	**1956**
	EEC	1	**0.75**	**0.75**	**0.83**	**0.81**	**11784**
		2		**0.72**	**0.51**	**0.60**	**5609**
		3		**0.72**	**0.91**	**0.80**	**24559**
		4		**0.88**	**0.38**	**0.53**	**8818**

Table 6 - Results of the best method to handle the missing values

The observed metric variations, such as precision and recall, stem from the dataset's imbalanced nature, where one class outweighs another significantly. In this setting, model performance, particularly in a less represented class, can show seemingly lower results. When addressing null values, the strategies might have accounted for these imbalances, impacting the overall metric performance by favoring the dominant class during prediction. A method was applied to balance the dataset, but the results were worse and in this specific case it does not make sense for the data to be balanced, there aren't the same properties with and without garage and the same happens for the variable swimming pool.

The treatment of the missing values was essential to achieve the main objectives of this study. With all variables without null values, the obtained dataset is ready to be tested and predict real estate property prices.

4. RESULTS AND DISCUSSION

This section details the obtained results and findings regarding the constructed models.

4.1. EXPLORATORY ANALYSIS

Since the dataset includes a time component, the initial point in this analysis was time trends, how variables change over time. The time variable chosen was Last Market Update Year since is the variable that gives us the last market update. Starting by analyzing the yearly evolution of real estate property prices and relations between the variables during the period under study. These insights provide valuable data-driven perspectives on housing trends in Portugal over the years, offering valuable information for decision-making and further analysis.

4.1.1. Exploring relationships among key variables

In Figure 2, six graphs represent distinct variables, showcasing their mean values over the years. The Energy Efficiency Certificate (EEC) variable operates on a descending scale. This scale ranges from 1 to 4, where a lower number indicates a higher energy efficiency rating. For instance, a certificate value of 1 represents the highest energy efficiency ratings, such as A and A+, while a rating of 2 signifies certificates like B and B-. Moving down the scale, a value of 3 encompasses certificates C and D, and a value of 4 includes the less energy-efficient certificates E, F, and G. Therefore, the closer the assigned number is to 1 on this scale, the more efficient the energy rating of the property is. Signaling an increasing societal concern for environmental issues and the pursuit of more energy-efficient housing. Consequently, new constructions tend to exhibit higher Energy Efficiency Certificate ratings, while existing properties seek to enhance their energy performance.

Turning to the property condition variable, there was a significant drop in 2018 compared to 2017, indicating a rise in the number of properties in poorer conditions during that year. Presently, it is discernible that, on average, residences in Portugal fall within the spectrum of “good” to “used” conditions.

The property commodities play a crucial role in gauging whether the properties in the dataset possess specific characteristics, thereby aligning them more closely with the general landscape of properties in Portugal. Regarding the garage variable, the dataset reveals a reduction in the number of properties with garages, suggesting that there are more properties without this feature than with it. Similarly, the terrace, patio, and swimming pool variables exhibit analogous behavior, indicating that 2021 marked the year with the highest number of houses in the dataset featuring these characteristics. The properties tend to not have these characteristics.

In Figures 3 and 4 appear charts to complement Figure 2. In Figure 3 it’s represented the evolution of the property commodities over the years in terms of number of properties that in that specific year appear in the dataset. As show in Figure 4, it displays the representation that each variable has across the entire dataset, it can be concluded that garage variable is the property commodity that most properties have.

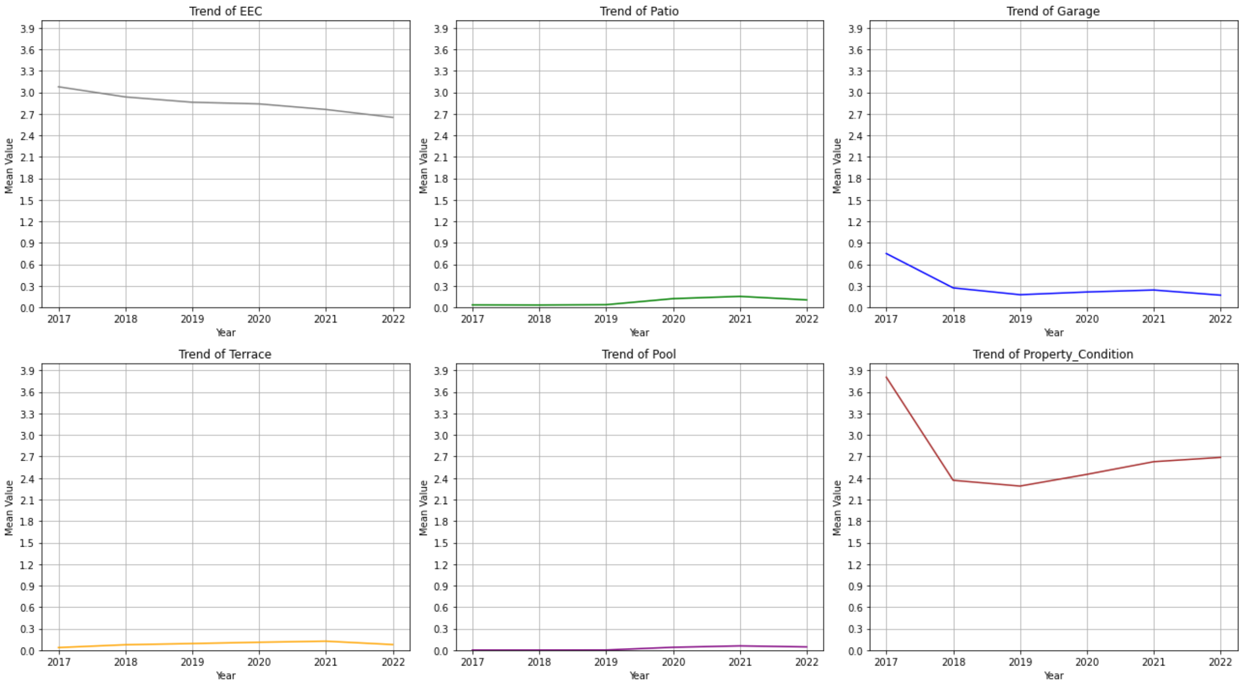

Figure 2 - Trend of variables over the years

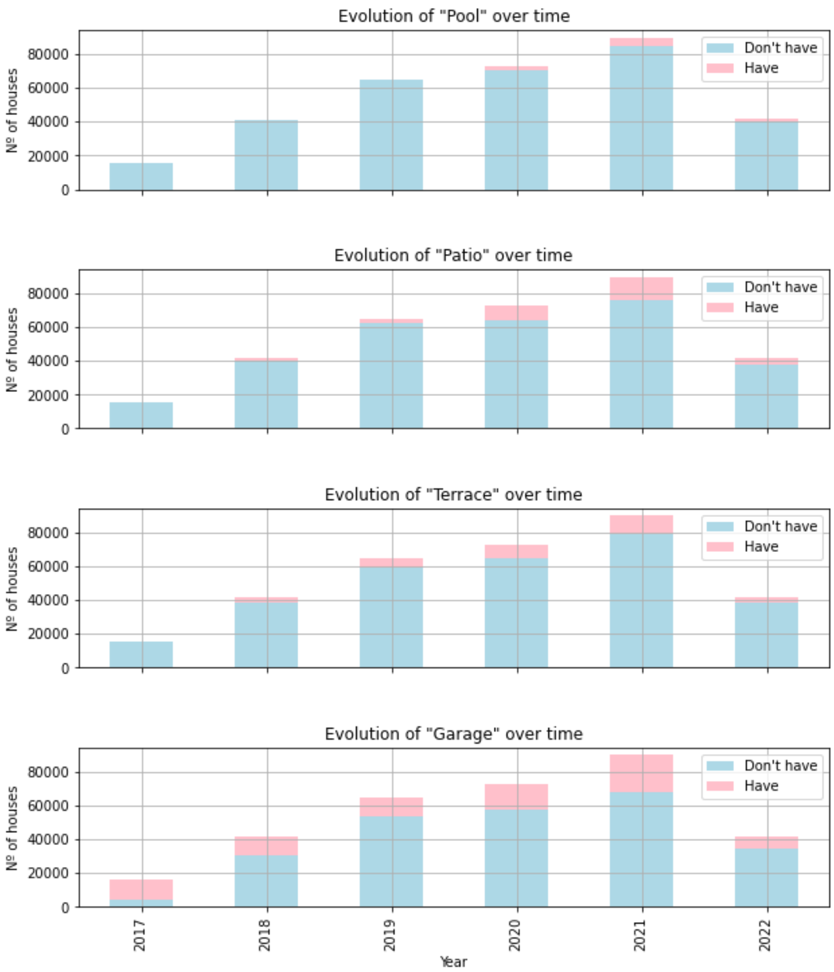

Figure 3 – Presence or absence of property commodities over time

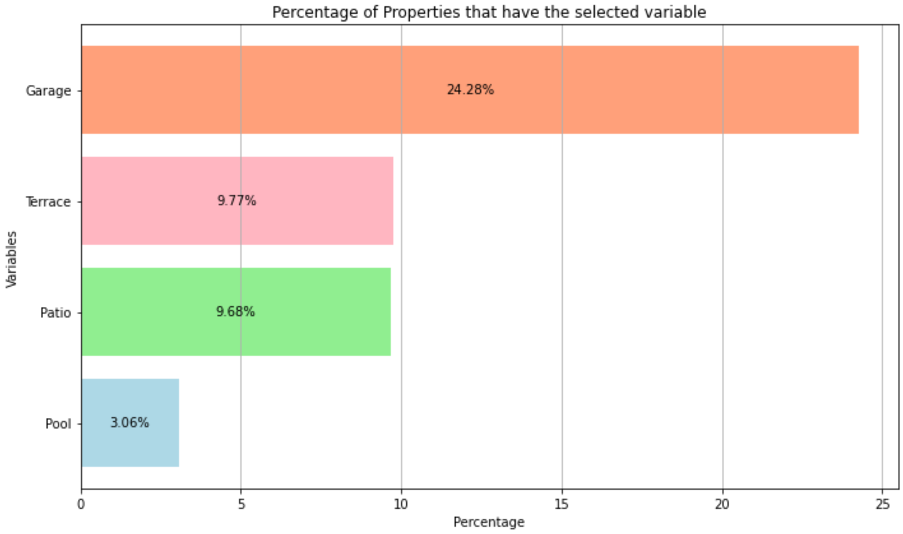

Figure 4 - Percentage of properties that have the selected variable.

Complementary Figure 2, Figure 5 gives us the number of houses and the evolution over time for the variables property condition and EEC. The number of properties in the "used" condition is increasing, indicating an increasing trend for renovations, as well as an increase in newly constructed properties. The observed trend in the EEC indicates an improvement in property conditions over time. While the majority of properties possess certificates of C or D level, considering a bad certificate, there is a notable increase in higher-grade energy certificates such as A+, A, B, and B-. This signifies a progressive shift, reflecting an emphasis and increasing valuation towards energy efficiency within the real estate sector.

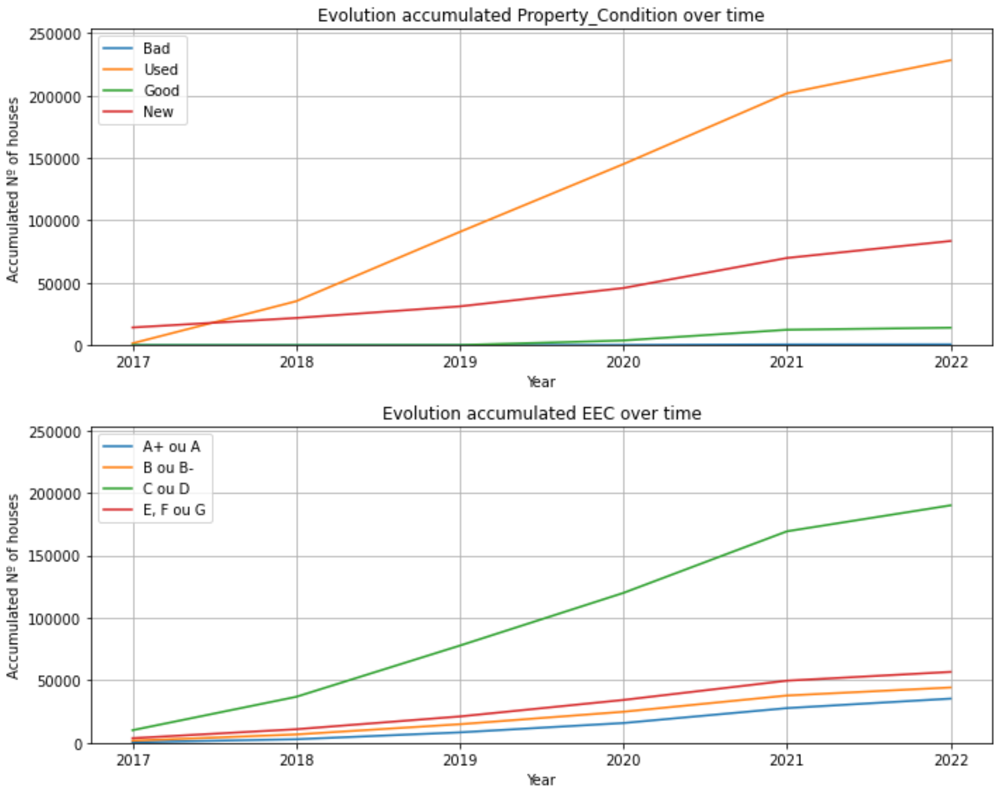

Figure 5 – Evolution of Property Condition and Energy Efficiency Certificate over time

As we observe an increase in the number of newly constructed properties, it is important to analyze the significance of property commodities in these new houses. From Figure 6, we can conclude that new properties tend to exhibit these characteristics more notably. Garage variable appear in more than half of new constructions, providing information about which builders and buyers are increasingly valuing. In the Figure 7, we can also see the difference between the number of properties being built as apartments versus the number of homes, with more apartments being constructed.

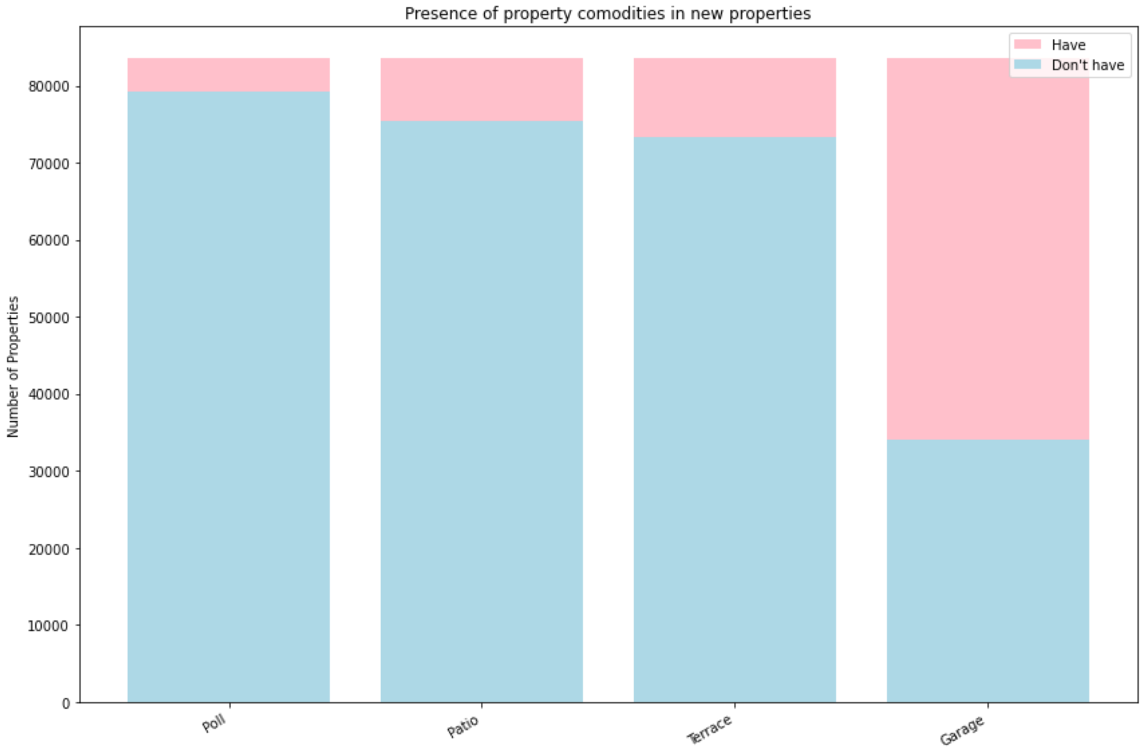

Figure 6 - Number of newly constructed properties with specific characteristics

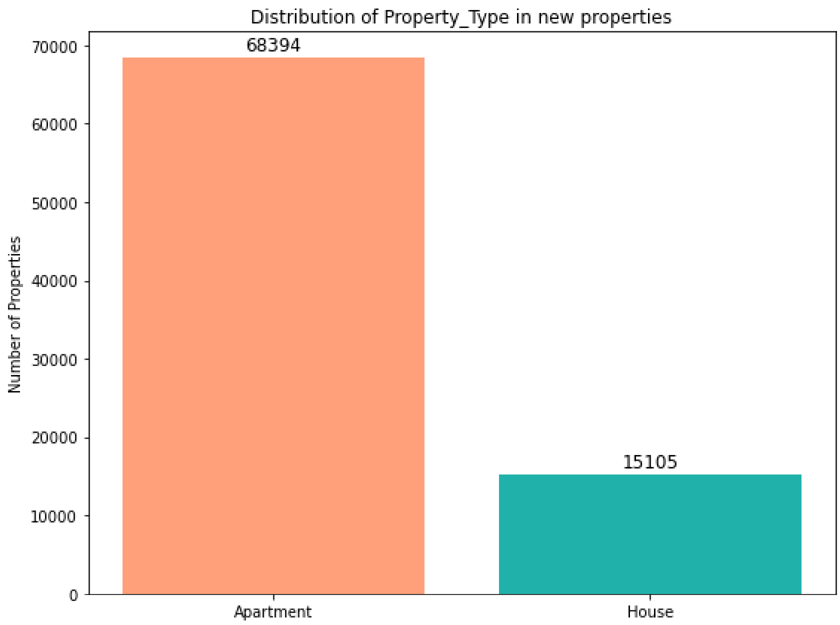

Figure 7 - Number of newly constructed properties as apartment or houses

To examine the geographic relationship between properties in rural or non-rural areas, analyzes were conducted to assess the behavior of property commodities in relation to the property types. In Figure 8, the difference between the percentage of the number of properties with and without these of property commodities are not substantial. The most notable difference can be observed in the property type, apartment, or a house. There are more apartments in non-rural areas compared to rural areas, whereas in rural areas, there is a higher prevalence of houses. The reason for this distinction can be attributed to urbanization trends, where non-rural areas often witness higher population densities, leading to a greater demand for apartments. Conversely, rural areas tend to have more space available, which encourages home construction.

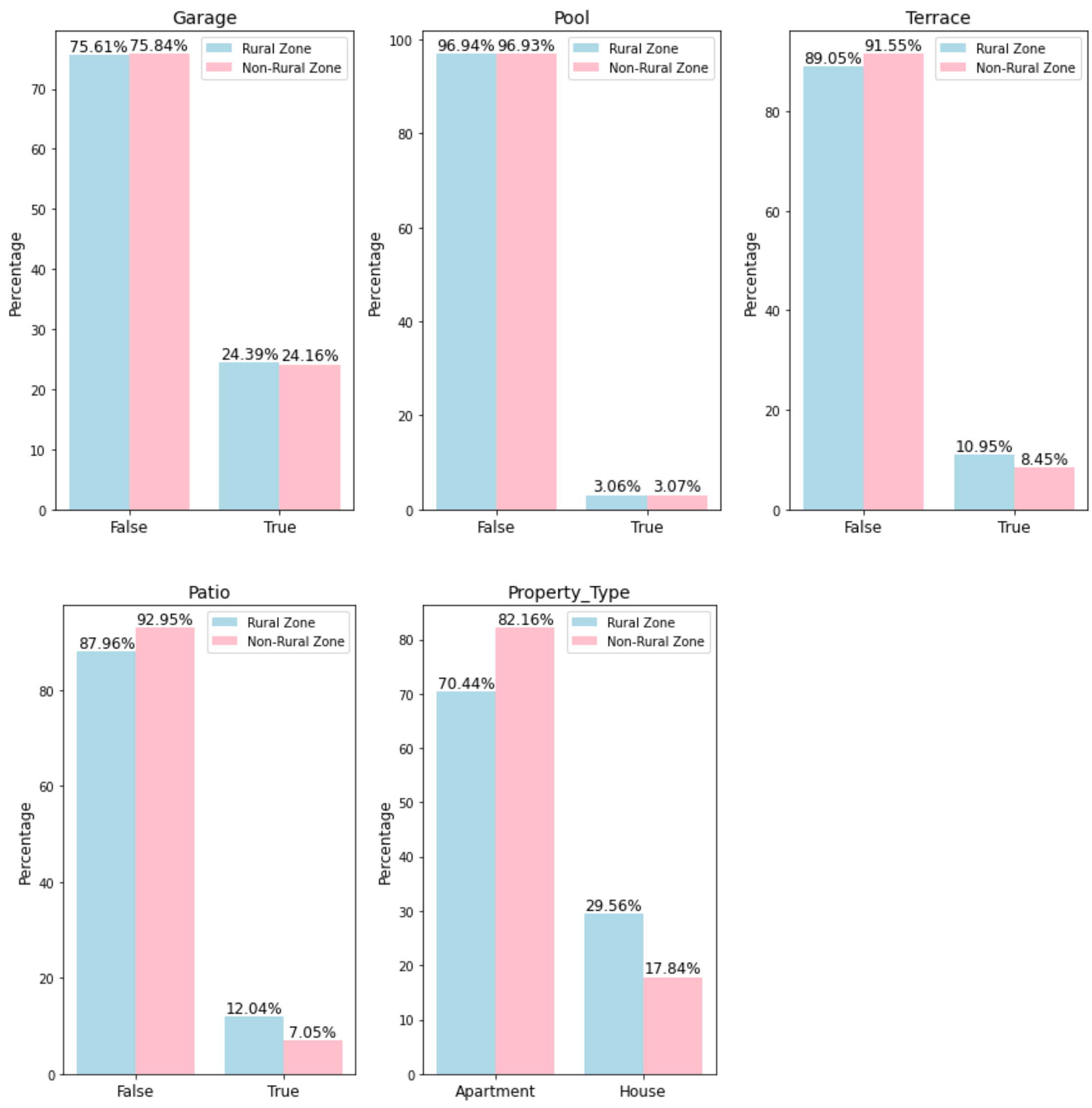

Figure 8 – Property commodities analysis by property type in Rural and Non-Rual areas

Following the geographical analysis of property characteristics, a thorough study of property commodities and property types was carried out to understand their distribution across the five zones in Portugal, as shown in Figure 9. The garage variable displayed a consistent percentage throughout the country, with the absence of garages being more common.

Remarkably, the swimming pool variable stood out in the Algarve region, where approximately 10.73% of properties featured this variable. In the southernmost regions of country, such as Alentejo and Algarve, the presence of terrace and patio was significantly more common.

The Lisbon region has a lower proportion of properties classified as Houses, while the zones Center and Alentejo have a more balanced ratio between apartments and houses. It is notable that the majority of properties in the Alentejo are considered Rural, making it the predominantly rural area in the country.

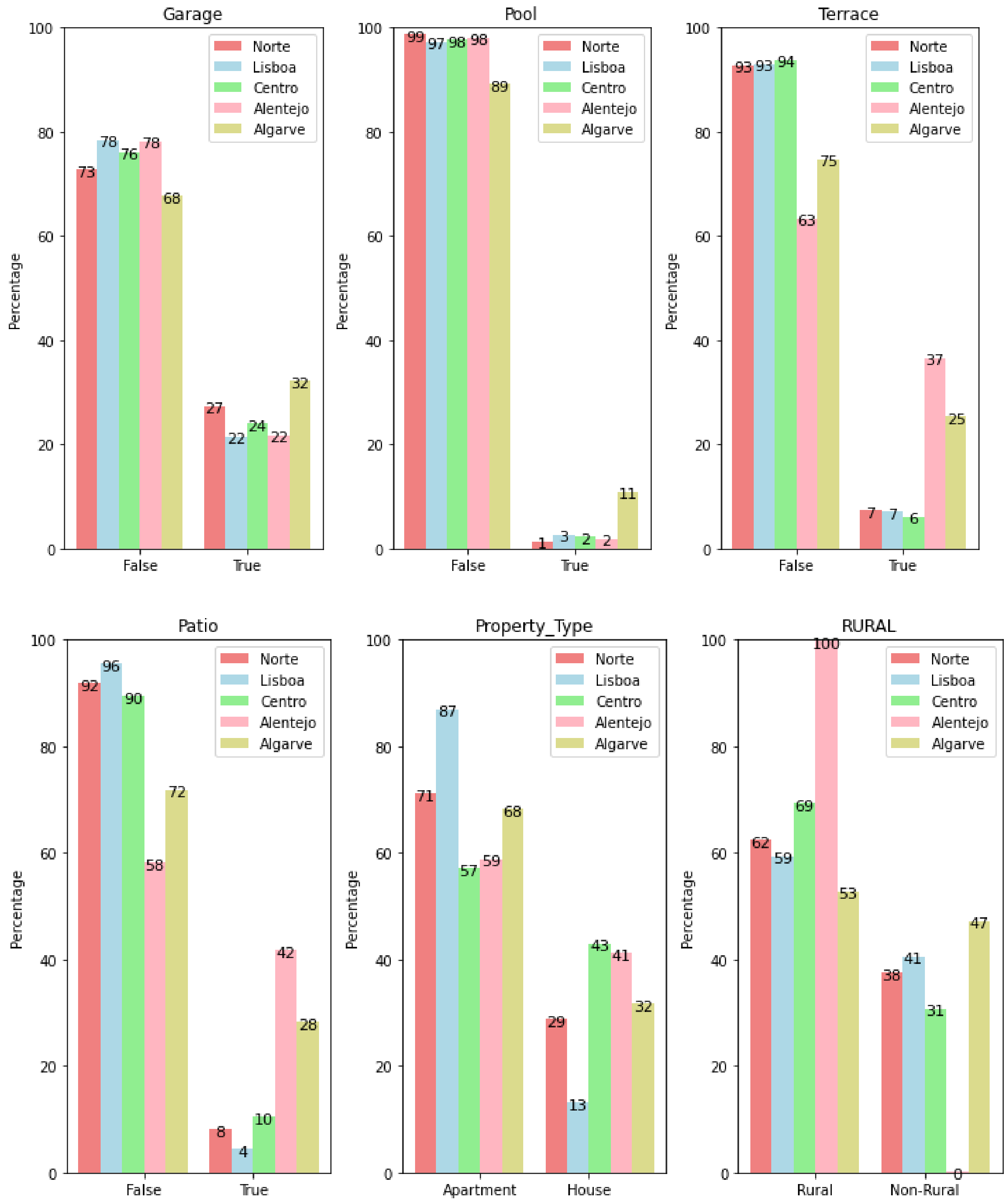

Figure 9 - Distribution of properties along the regions

Figure 10 illustrates the existence of a moderately strong positive relationship between the variables Groos Building Area and the number of bedrooms. The correlation coefficient is 0.68, which is a relatively high relationship, indicating a clear trend where an increase in the gross private area of properties corresponds to an increase in the number of rooms.

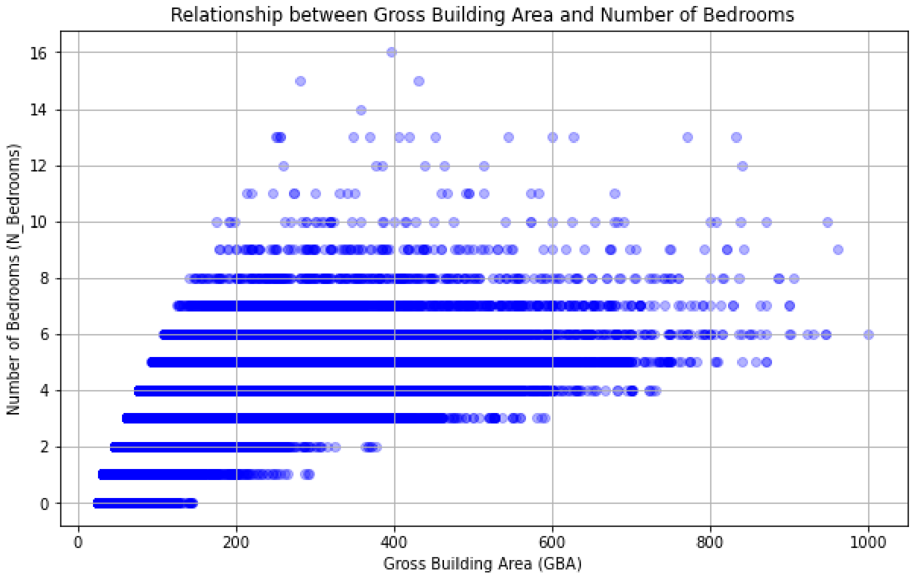

Figure 10 - Relationship between GBA and Number of bedrooms

4.1.2. Exploring price trends in geographic context

Continuing the analysis of real estate property prices and examine their development in different contexts.

Figure 11 provides an insight into the average development of property prices in the years considered. It is evident that the prices in 2018, there was a significantly increase compared with the previous year. Throughout the years 2018 to 2020, there were no substantial fluctuations, and from the year 2020 onwards, prices have consistently been on the rise.

Figure 11 - Evolution of real estate property prices

In Figure 12, the progression of property prices across the five regions and the study years is represented. The North region exhibits a gradual and less pronounced growth, while the Algarve region experiences slight price fluctuations. The Center and Alentejo zones have maintained their values. In contrast, the Lisbon region boasts the highest real estate prices prices and displays a more significant price variation, particularly marked by price increases.

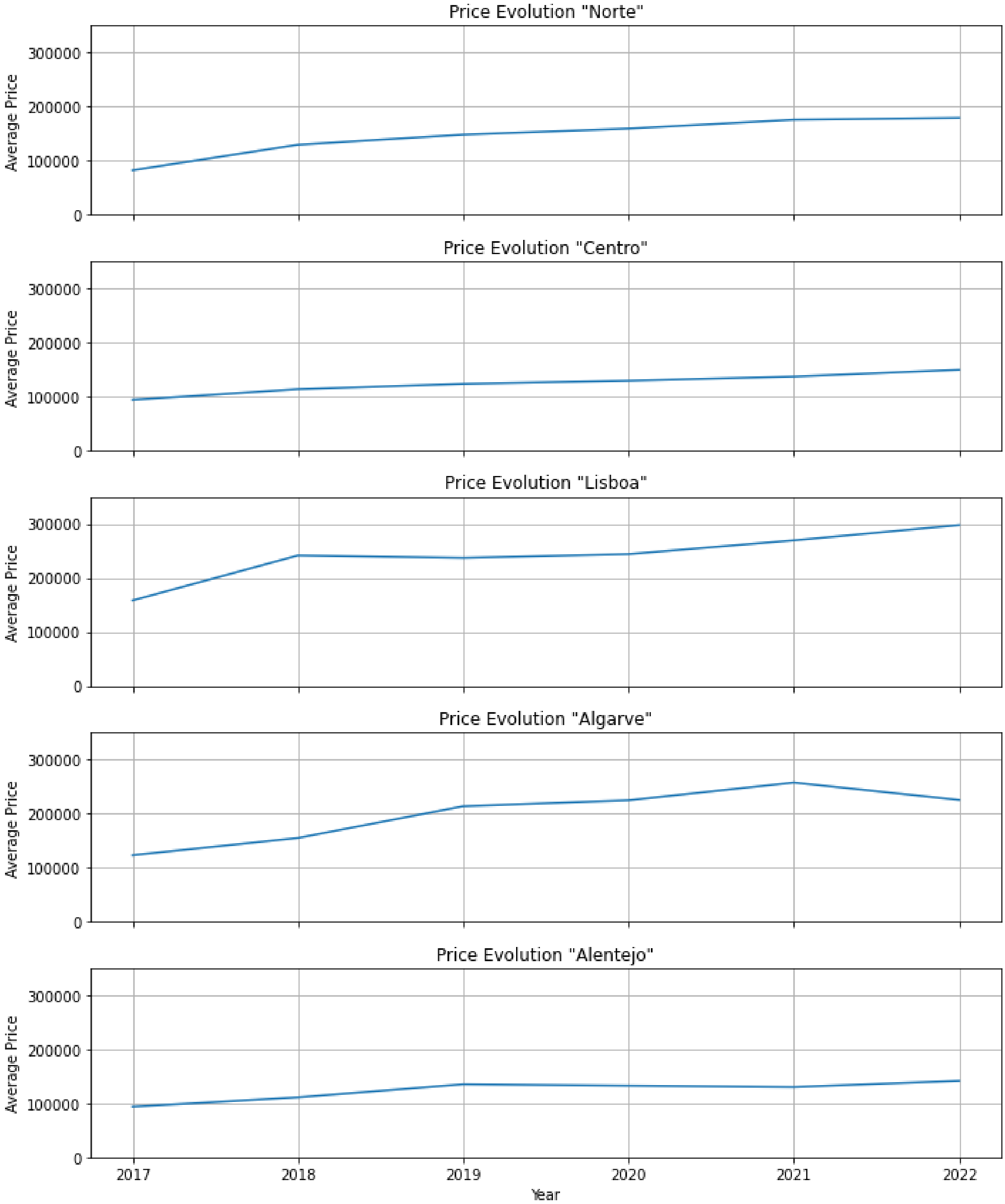

Figure 12 - Evolution of real estate property prices along the regions

In order to organize the zones from the highest average selling prices to the lowest, the Figure 13 is presented. It is evident that Lisbon region has the highest real estate prices, closely followed by the Algarve region. The Center region is characterized by more affordable property prices.

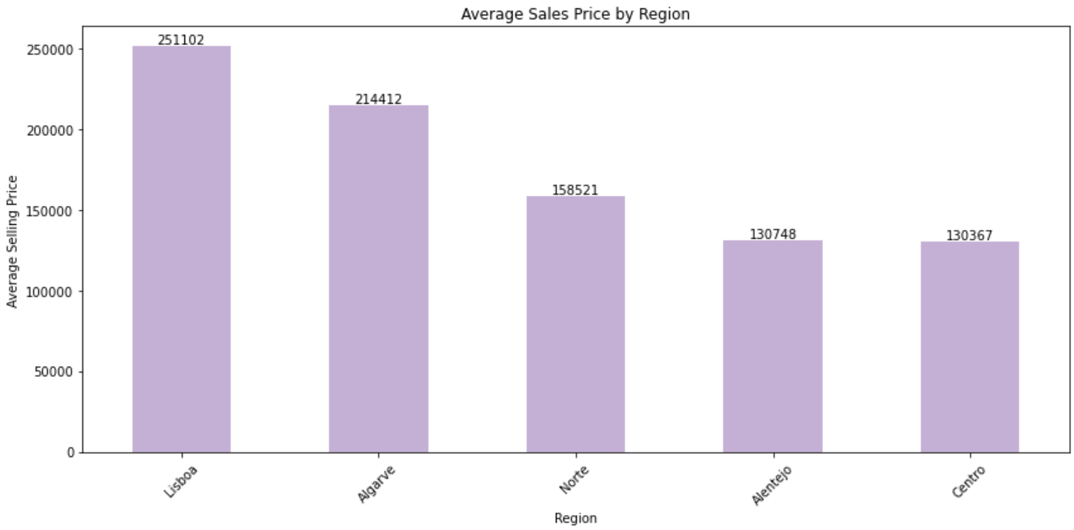

Figure 13 - Average property selling prices along the regions

It is important to have a sense of the price disparities among the districts of Portugal. Confirming the previous information, Figure 14, illustrates that the Lisbon district has the highest average property prices, closely followed by the Faro district. In contrast, the Guarda district emerges as the most budget-friendly option. Remarkably, the average housing price in Lisbon is over three times higher than that in Guarda. In terms of pricing within this market, Lisbon is 37% more expensive than the second most populous district in the country, which is Porto.

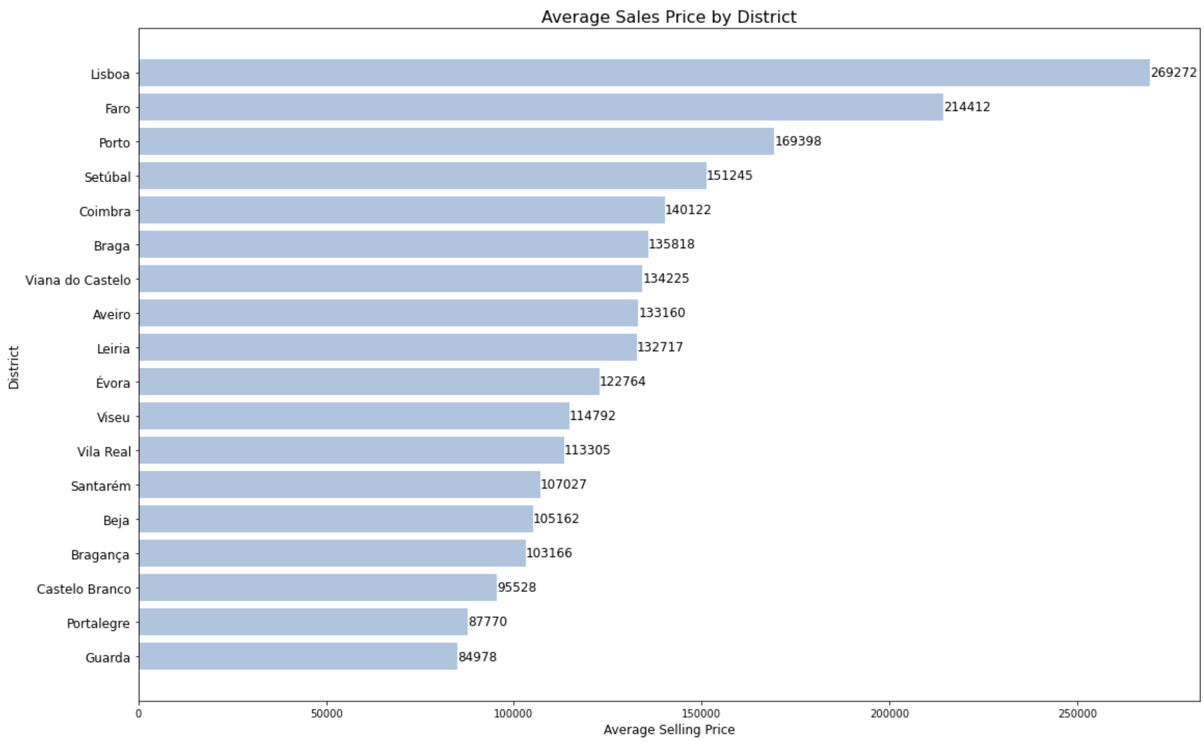

Figure 14 - Average property selling prices in the districts of Portugal

Given that Lisbon is the district with the highest real estate property prices, it becomes important to narrow down the geographical analysis. Consequently, the Lisbon municipality was examined, and Figure 15 appears. Among the municipalities within the district, Cascais stands out with the highest average housing prices, closely followed by the Lisbon municipality, with little difference between the two values. In contrast, Azambuja, a municipality in the Lisbon district, has the lowest average housing prices, with a value that is over three times cheaper compared to Cascais.

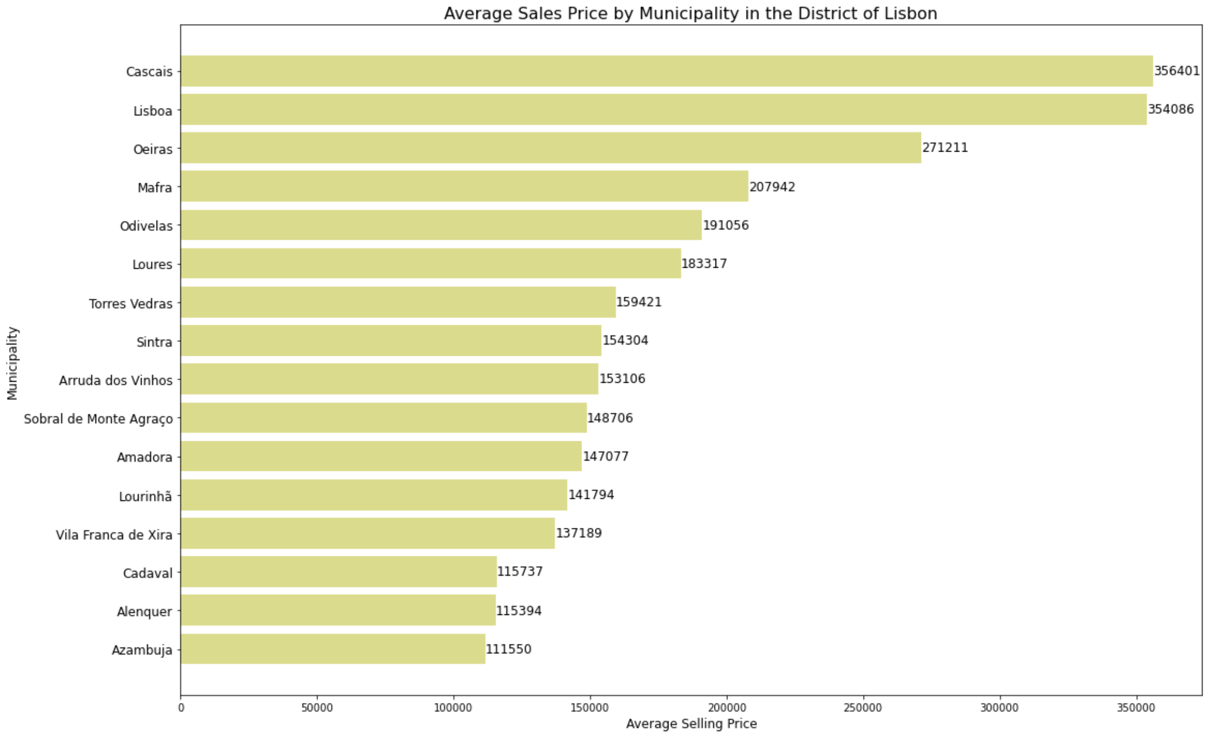

Figure 15 - Average property selling prices in municipality of Lisbon

Since Cascais and Lisbon were the municipalities with the highest real estate selling prices, Figures 16 and 17 represent the prices within the respective parishes. In Cascais, the difference between the parish with the highest and the lowest prices is approximately the double. In Lisbon, the differences are even greater, with prices varying by almost four times from the parish with the highest average prices to the one with the lowest.

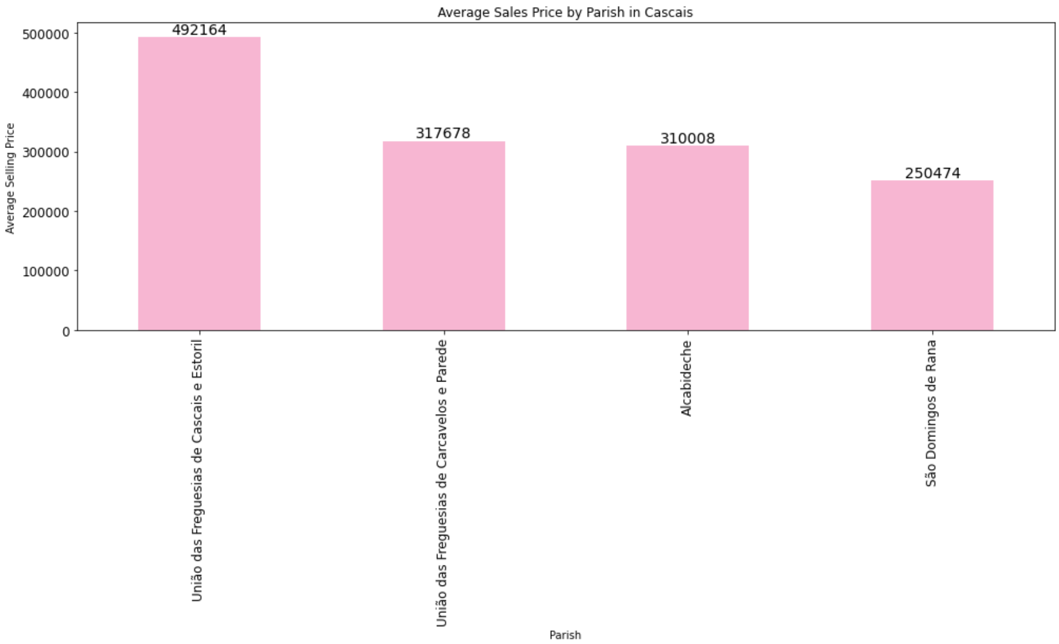

Figure 16 - Average property selling prices in the parish of Cascais

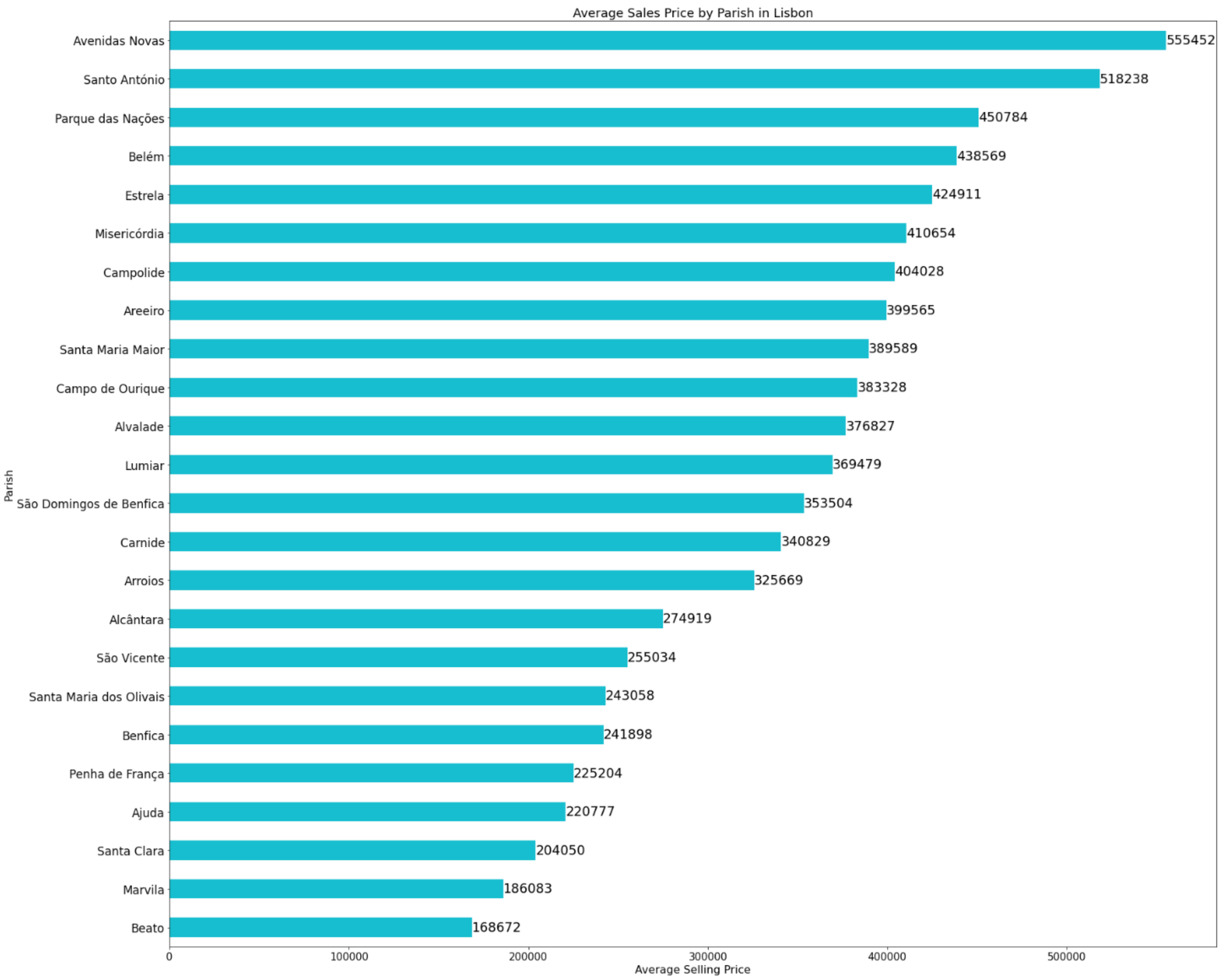

Figure 17 - Average property selling prices in the parish of Lisbon

As we can see, real estate property prices in Lisbon are higher and the differences between parishes are very noticeable. To better understand how prices develop in this parish, the as shown in Figure 18 were created. It is possible to examine the price evolution and maximum price spikes in each parish of Lisbon.

In the year 2017, properties with the highest prices were located in the Santa Maria Maior and Santo António parishes. In 2018, the Avenidas Novas parish experienced a significant price increase, becoming the parish with the highest real estate property prices. In 2019, the Parque das Nações parish doubled in value compared to 2017, now ranking among the parishes with the highest housing prices, consistently increasing in price over the years.

By 2022, we can conclude that the top five parishes with the highest property prices, in descending order, are Avenidas Novas, Parque das Nações, Santo António, Belém, and Campolide. The parishes with lower prices that have remained comparatively more affordable include Ajuda, Santa Clara, Beato, Penha da França, and Marvila.

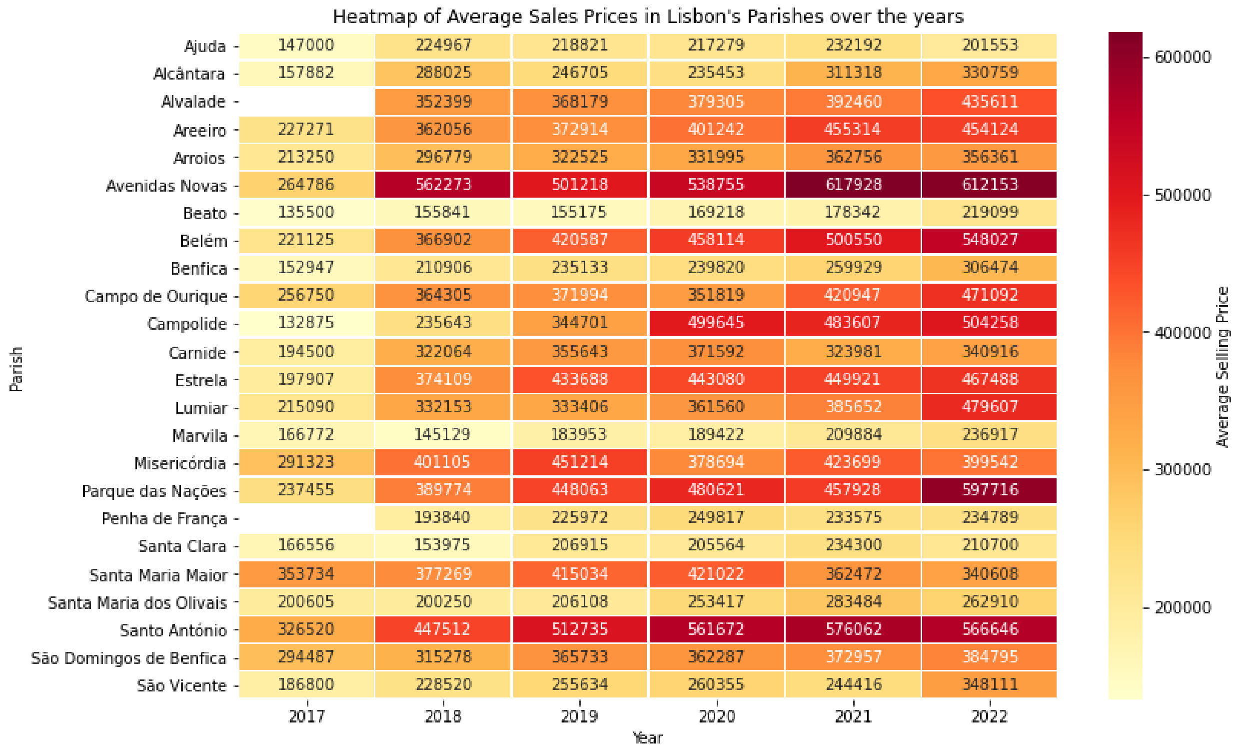

Figure 18 - Heatmap of the evolution of property selling prices in Lisbon's parishes

4.2. Models composition

The final dataset represents a culmination of the best models previously tested to fulfill missing values. With these refined variables now integrated, the dataset stands prepared for the accurate prediction of property prices. Next, we perform a rigorous comparison of models to discern their effectiveness in the real estate context. Through systematic evaluation and careful analysis based on the previously mentioned metrics, we aim to identify the model that performs best for the real estate price forecasting task and the existing differences when analyzed in different geographic context.

For all the predicting methods, the dataset was divided into two sets, the training set, and the test set. The original dataset was also split into independent and dependent variables, as this split is essential for evaluating model performance using previously unseen data. The split between training and testing dataset was explored for the different scenarios. The split used was 80% - 20%, which means 20% of the data is used for the testing dataset and the remaining 80% of the data in the training.

After training, predictions were made on the test basis. These forecasts represent estimates of the dependent variable based on the independent variable.

In all models, the analysis of feature importance was conducted, and these features were carefully chosen to create the forecasting inputs. The main focus is to include features that have a strong correlation with the target variable and show a significant impact on property selling price. During model testing and validation, the consequences of including or excluding certain features were evaluated. Attributes that significantly improved predictive performance of the models were incorporated into the final input. This deliberate selection ensures that the forecasts are optimized, as it focuses on the features that exert the most significant impact on the predictions, enhancing the overall accuracy and reliability of the models.

To determine the performance metrics of the regression models, several metrics were evaluated. After training the model with the different forecast models, add the price prediction to the database by creating a new column "Predicted Selling Price".

To optimize the performance of our machine learning models, an extensive experiment was conducted to evaluate various combinations of hyperparameters. The objective was to identify the optimal settings for each algorithm, ensuring that the models were well-tuned and highly accurate. Table 7 highlights the specific hyperparameters used in the prediction model.

Machine Learning Models	Best Hyperparameters
Random Forest	Random State: 42. N estimators: 100. Min Samples Split: 2. Min Samples Leaf: 2. Max Features: sqrt. Criterion: MSE. Bootstrap: True
Decision Trees	Splitter: best. Random State: 42. Min Samples Split: 5. Min Samples Leaf: 2. Max Features: sqrt. Criterion: MSE
KNN	Weights: uniform. N neighbors: 5. Metric: minkowski. Leaf size: 16. Algorithm: kd_tree

Table 7 - Best hyperparameters for the models

4.3. MODEL RESULTS

Predicted real estate prices start with all properties in the dataset and use three different forecast models: Random Forest, Decision Tree and KNN. The results are shown in Table 8.

Random Forest model achieved the highest R^2 score, 82,34% which indicates that approximately 82% of housing prices can be explained by the correlated variables. By comparing the price between "Observed Selling Price" and the model predicted price, and evaluating this price comparison using the MAE metric, it is found that the MAE between the estimated price and the predicted price is 21.969€.

This value indicates that the average absolute difference between model-predicted and observed prices for all houses in the Portugal dataset is 21.969€. Furthermore, MAPE were calculated, the MAPE obtained was 11,61%, indicating that, on average, the model's forecasts have an 11,61% difference compared to the observed prices. Based on these results, it can be concluded that the Random Forest model demonstrated a good ability to predict house prices with a relatively low average error. The performance of Decision Tree and KNN's lagged behind Random Forest, showing significantly lower R^2 Score values and higher error rates in the case of KNN. Hence, based on these results, the Random Forest model prevails as the more effective choice. When minimizing the Mean Absolute Error (MAE) is the primary priority, Decision Tree becomes an attractive option. In the context of the real estate market and the specific requirements of this study, where the main objective is to strike a balance between explaining the variability of real estate prices (as shown by R^2 Score) and minimizing the MAE, the Random Forest model emerged as the most suitable option.

As we have seen in our exploratory analysis, Lisbon has become an important region for a comprehensive study of real estate property prices. This encourages us to go deeper with the analysis by using Random Forest and Decision Tree models to forecast real estate property prices. When comparing the performance of these two models, the Random Forest model stands out with a significantly higher R^2 Score (81,46%). This highlights its superior fit to the training data, indicating its ability to explain a greater portion of the variability in house prices. Its lower absolute MAE, the Random Forest model's performance remains strong, considering the lower average property values in this region, as noted in exploratory analysis. Impressively Decision Tree, it delivers a lower MAE, 32€/m², compared to the Random Forest model, indicating more accurate predictions in terms of price per square meter. Moreover, MAPE is lower for the Random Forest model, reflecting a reduced percentage error rate in its predictions. Given these metrics and our priority of achieving accuracy in property price predictions, the Random Forest model is the model of choice.

Considering Random Forest's superior performance, it was not only chosen for predicting property prices in Lisbon region but also extended to forecasting property prices in various regions. A conducted forecasts for the five distinct regions that Portugal is divided into, aiming to assess how these prediction models perform across diverse geographical landscapes. This comprehensive approach allows us to gain valuable insights into property price predictions across the entire country.

These diverse predictions by region showcase the distinct property market characteristics across Portugal. The Random Forest model, while demonstrating strong predictive capabilities in Lisbon, faces varied challenges in different regions. So, Decision Tree model was also considered in prediction by regions.

The results indicated varying degrees of accuracy and error across different models and property categories. Notably, the Random Forest model displayed remarkable prediction capability across different regions and property types. Using the Random Forest model, the North and Alentejo regions display relatively higher R^2 Score, 75,08% and 77,17%, respectively, with corresponding MAE of 142€/m² and 135€/m². In contrast, the Center and Algarve regions exhibit lower R^2 Score, 64,11% and 61,80% and marginally higher MAE of 119€/m² and 223€/m², respectively. It achieved substantial accuracy in predicting property prices in the Lisbon region, reflecting an R^2 Score of 81.46% with a corresponding MAE of 25.176€ and 242€ per square meter.

These variations underscore the significance of region-specific models for property price predictions. The diverse landscape of property prices across Portugal's regions underscores the significance of tailoring prediction models to each area. The variations in house prices result from local market dynamics, economic conditions, and housing demands. The models' abilities to adapt to these differences demonstrate their capacity to capture and leverage regional intricacies, making them valuable tools for real estate forecasting.

When using the Decision Tree model, its performance is significantly worse compared to the Random Forest model. In the Alentejo region, the Decision Tree model is more compatible than in other regions. It is difficult for the Decision Tree model to accurately predict property real estate prices. In particular, model performance, represented by R^2 score and MAE, shows that the Decision Tree model performs below average, indicating lower accuracy and weaker adaptability, which is particularly evident outside the Alentejo region. This discrepancy demonstrates the limitations of Decision Tree model in capturing the complexity of real estate price dynamics in different regions.

The decision to make separate forecasts for apartments and houses in Lisbon provides valuable insights into the local real estate market. It highlights the main differences in performance indicators and sheds light on buyer preferences and property characteristics in Lisbon. The results show that the Random Forest model performs well in predicting Lisbon apartment prices, with a high R^2 Score of 83,49%, a low MAE of 241€/m², and a MAPE of 10,29%. This suggests that the housing market is more predictable, possibly due to the homogeneity of properties and continued demand. In contrast, forecasting house prices in the same area produced more modest results. The R^2 Score is reduced by 71,04%, the MAE is significantly higher at 49.205€, with an MAE of 252€/m², and the MAPE is higher at 15,44%. This means the real estate market is more complex, with different property characteristics affecting prices and making them less predictable. Overall, the segmented forecast highlights the need to understand market trends in Lisbon. Property type (apartments vs. houses) plays a crucial role in forecast accuracy and highlights the importance of customized models that consider the specifics of regional real estate market.

Machine Learning Models	Experimentation for Property Price Prediction	R^2 Score	MAE	MAE / m²	MAPE
Random Forest	**Predicting Real Estate Prices**	82,34%	21.969€	201€	11,61%
Decision Tree		69,01%	57.889€	27€	30,15%
KNN		38,52%	69.822€	716€	37,42%
Random Forest	**Predicting Real Estate Prices in Lisbon Region**	81,46%	25.176€	242€	10,72%
Decision Tree		66,58%	65.898€	32€	27,62%
Random Forest	**Predicting Real Estate Prices in North Region**	75,08%	16.323€	142€	11,15%
Decision Tree		48,20%	43.202€	18€	29,10%

Random Forest	**Predicting Real Estate Prices in Center Region**	64,11%	15.132€	119€	13,48%
Decision Tree		23,95%	42.605€	16€	35,94%
Random Forest	**Predicting Real Estate Prices in Alentejo Region**	77,17%	17.026€	135€	15,45%
Decision Tree		62,27%	46.100€	20€	38,17%
Random Forest	**Predicting Real Estate Prices in Algarve Region**	61,80%	27.069€	223€	12,86%
Decision Tree		18,19%	67.440€	30€	32,18%
Random Forest	**Predicting Real Estate Prices in Lisbon Apartments**	83,49%	22.822€	241€	10,29%
	Predicting Real Estate Prices in Lisbon Houses	71,04%	49.205€	252€	15,44%

Table 8 - Machine Learning Experimentation for Property Price Prediction

This forecast involves predicting property prices using three different models (Random Forest, Decision Tree and KNN) for the entire dataset, Portugal, focusing on the Lisbon region and extending to other regions of Portugal. The analysis has been performed and the metrics used to analyze the models performance, facilitating comparison with other models. This comprehensive study allowed to identify Random Forest as the model of choice due to its excellent performance in predicting real estate property prices, especially the Lisbon apartment market, and highlighting the unique dynamics of the Lisbon real estate market. These diverse forecasts highlight the importance of considering specific property types and regional peculiarities to accurately predict real estate property prices.

4.3.1. Error analysis of machine learning models

This error analysis provides valuable insights into the model's predictive capabilities and highlights varying accuracy and errors across regions and property types. Understanding these differences is critical to refining models to produce more accurate real estate price predictions.

The effectiveness of machine learning models in predicting property prices was meticulously analyzed using a suite of error metrics. The R^2 Score served as a primary indicator of the models' predictive performance. Furthermore, the Mean Absolute Error (MAE), MAE per square meter, and the Mean Absolute Percentage Error (MAPE) were pivotal in assessing the precision of these forecasts, as it is show in table 8.

4.3.1.1. Model performance overview

The visualizations presented below aim to provide insights into the predictive performance of the model developed for the predictions of real estate property price. The scatter plot illustrates the relationship between observed and predicted property prices, allowing us to visually examine the correspondence between observed and predicted values. Meanwhile, the histogram of residuals indicates the distribution and frequency of the differences between the observed and predicted values. Analyzing residuals helps to assess the model's accuracy by examining how the errors are distributed, enabling a deeper understanding of the model's predictive capabilities.

In the scatter plot, if the model predicted perfectly, all points would align along a diagonal line. Deviations from this line represent prediction errors. Points closer to zero imply that the model's predictions closely match real values. A uniform linear distribution also means that the model has positive predictive power.

The histogram of residuals aims to exhibit a uniform distribution around zero, indicating consistent predictions by the model. If the histogram has a prominent bar near zero, followed by smaller bars, it indicates that the predictions are roughly consistent with the actual values. The residuals are expressed in the same units as the predicted selling prices, such as euros. They represent the differences between the predicted and observed property prices.

This error analysis is performed on all models created to provide a comprehensive understanding of each model's error representation.

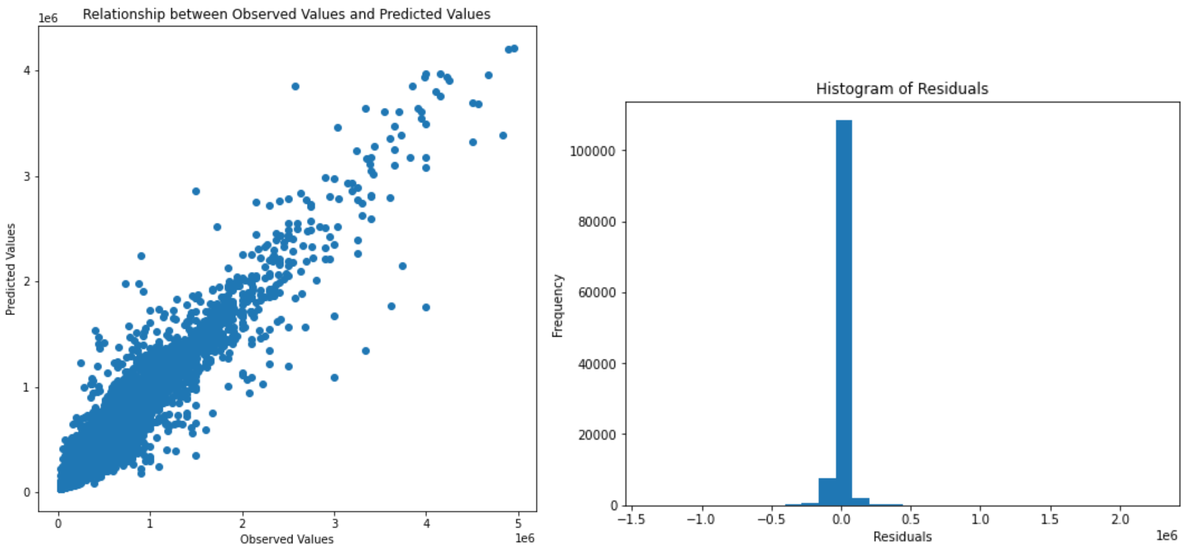

Figure 19 - Scatter plot and Histogram of residuals for predicting real estate prices using Random Forest

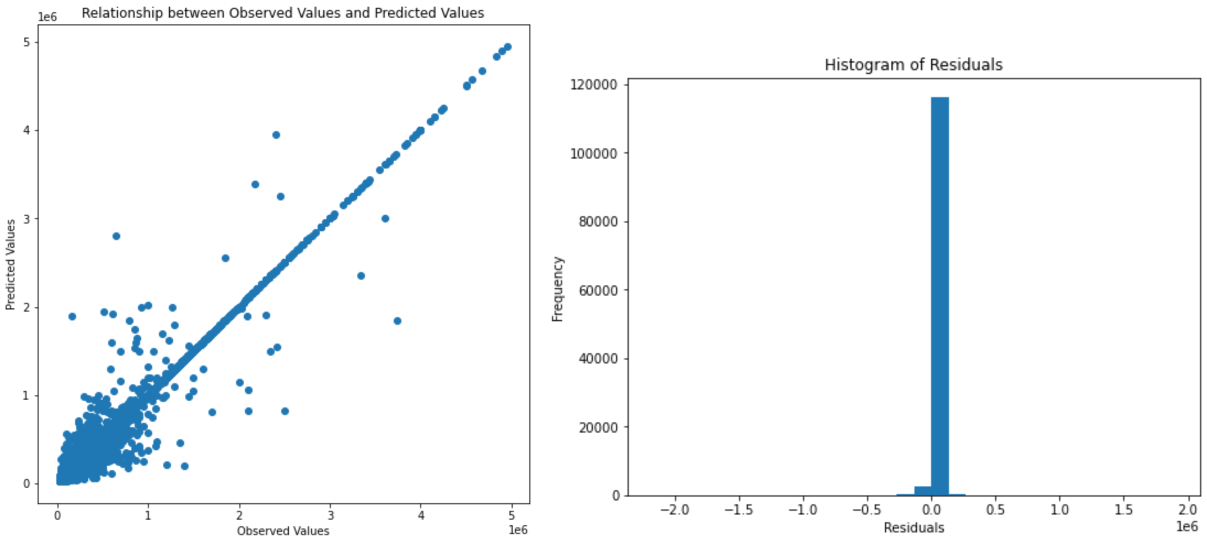

Figure 20 - Scatter plot and Histogram of residuals for predicting real estate prices using Decision Tree

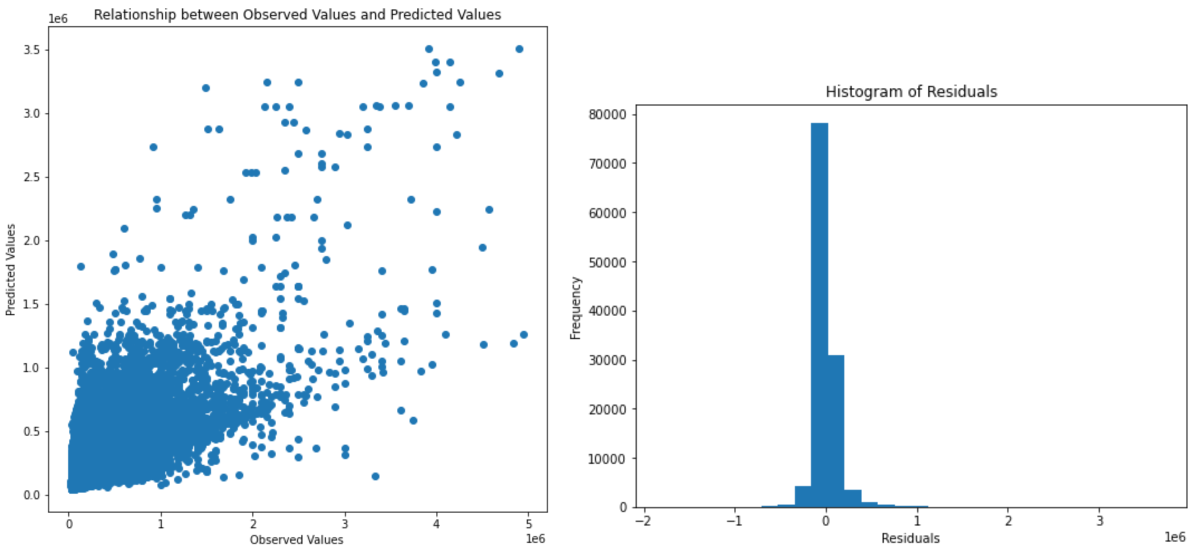

Figure 21 - Scatter plot and Histogram of residuals for predicting real estate prices using KNN

As concluded with the metrics presented in table 8, and right now with the use of the error analysis presented in Figures 19, 20 and 21, we can conclude, the same, that the Random Forest model demonstrates values closer to zero, signifying a smaller margin of error. Similarly, the Decision Tree model exhibits a linear alignment, also reflecting a minor error. Both models depict a relatively smaller deviation from the actual values. In contrast, the K-Nearest Neighbors (KNN) model showcases highly dispersed values, indicating a comparatively larger prediction error. This suggests that the KNN model has a higher level of inaccuracy in predicting the values.

4.3.1.2. Model performance overview in different regions

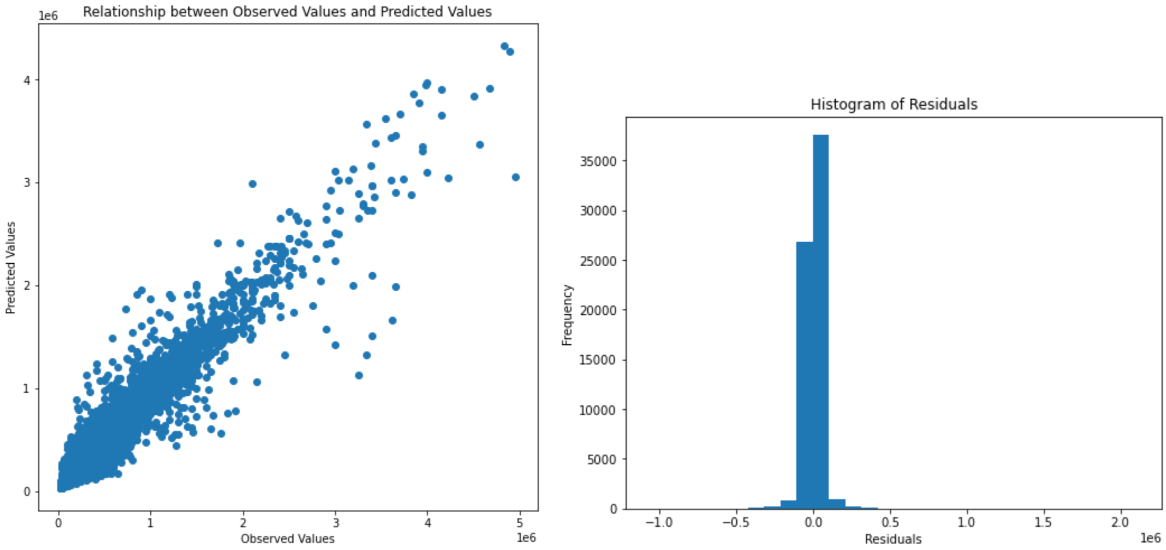

Figure 22 - Scatter plot and Histogram of residuals for predicting real estate prices in Lisbon using Random Forest

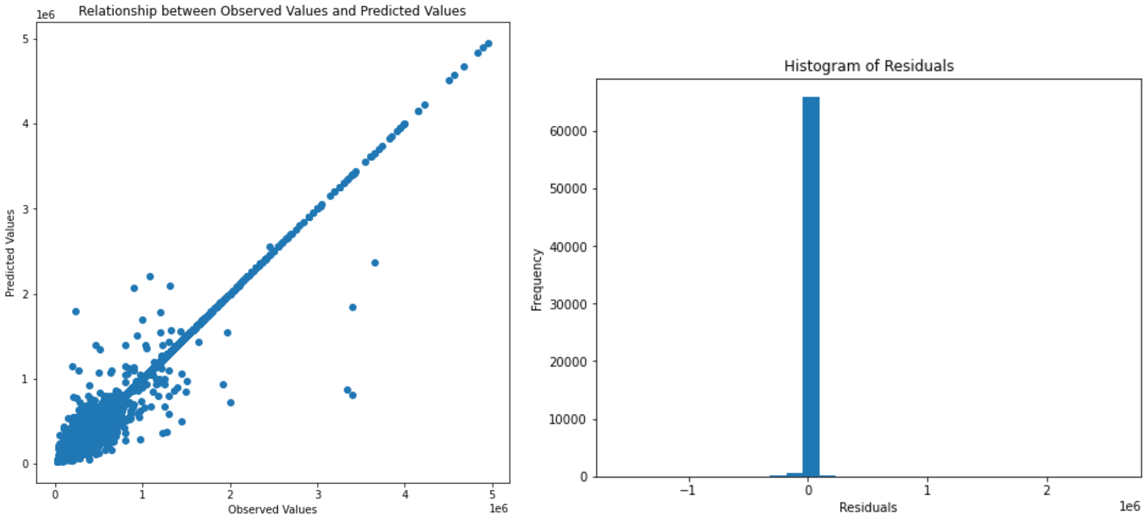

Figure 23 - Scatter plot and Histogram of residuals for predicting real estate prices in Lisbon using Decision Tree

For the Decision Tree model, the scatter plot, Figure 23, reveals a linear pattern, suggesting a strong alignment between predicted and observed values. The histogram, Figure 23, predominantly centered around zero with minimal deviation, signifies that the majority of residuals are close to zero. This uniformity implies a high level of model adjustment in predicting property prices. Conversely, in the Random Forest model, the scatter plot in Figure 22, displays a broader spread of values, deviating from a linear pattern. The histogram in Figure 22, exhibits two prominent columns near zero, indicating a higher frequency of predictions closer to the actual values, although with a certain level of variability. Overall, the Decision Tree model appears to demonstrate a slightly wider distribution of prediction errors compared to the Random Forest model.

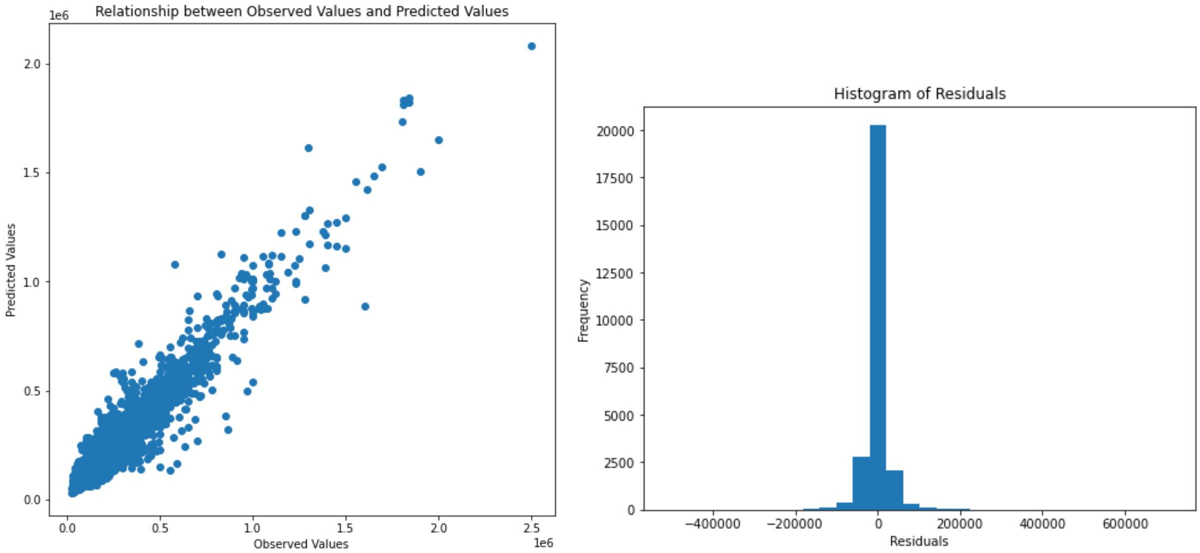

Figure 24 - Scatter plot and Histogram of residuals for predicting real estate prices in North using Random Forest

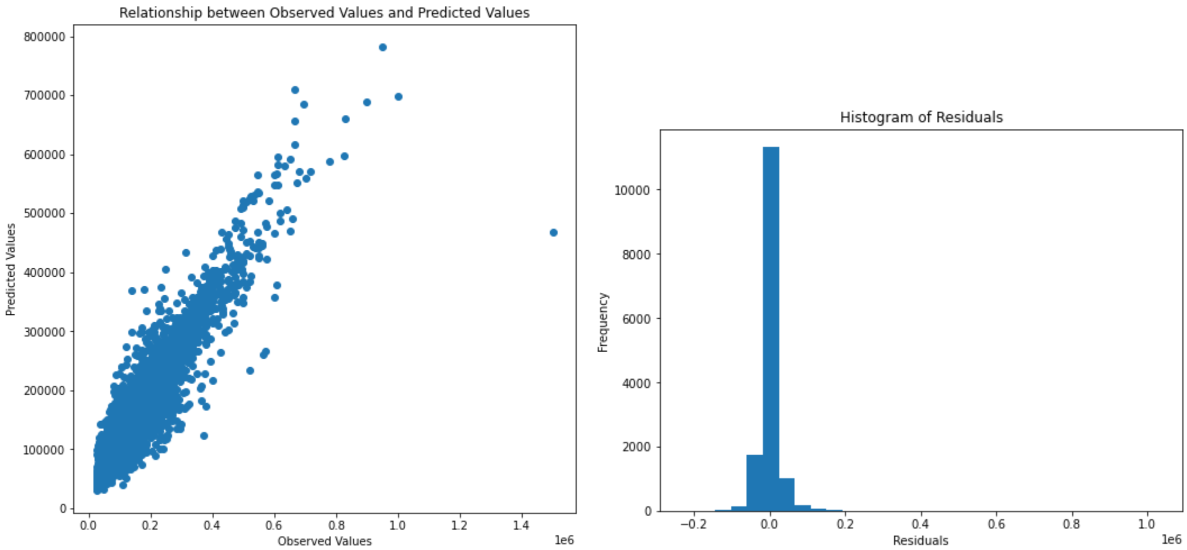

Figure 25 - Scatter plot and Histogram of residuals for predicting real estate prices in Center using Random Forest

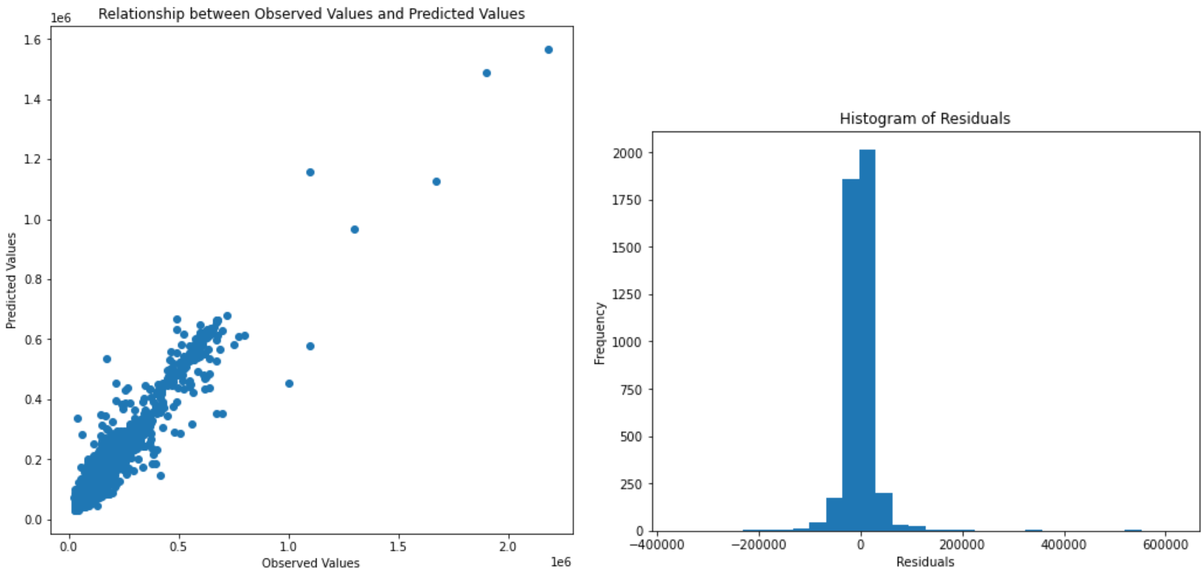

Figure 26 - Scatter plot and Histogram of residuals for predicting real estate prices in Alentejo using Random Forest

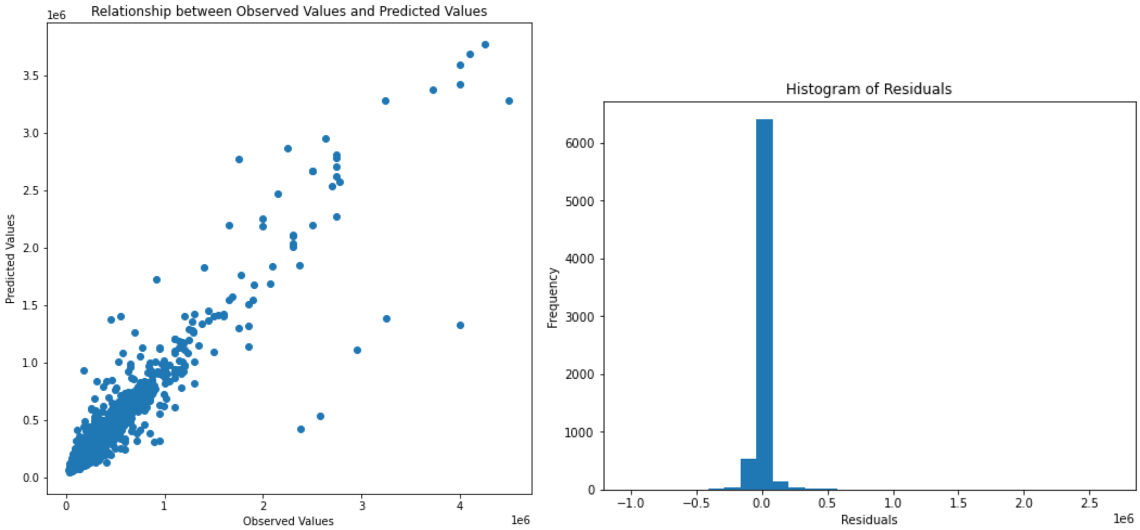

Figure 27 - Scatter plot and Histogram of residuals for predicting real estate prices in Algarve using Random Forest

The scatter plot in Figures 22 to 27 demonstrates uniformity along the regions, exhibiting similar patterns for each region, with the Algarve region displaying a higher dispersion of points, particularly those deviating from the diagonal line. Histograms in Figures 22 to 27, for different regions illustrate varying distributions. The Alentejo region presents higher density with multiple columns concentrated near zero, whereas the other regions predominantly exhibit a primary column near zero, complemented by smaller bars around this central point. These observations suggest that the model potentially showcases higher accuracy and consistency in predicting property prices within the Alentejo region compared to other regions.

4.3.1.3. Model performance overview in the property types

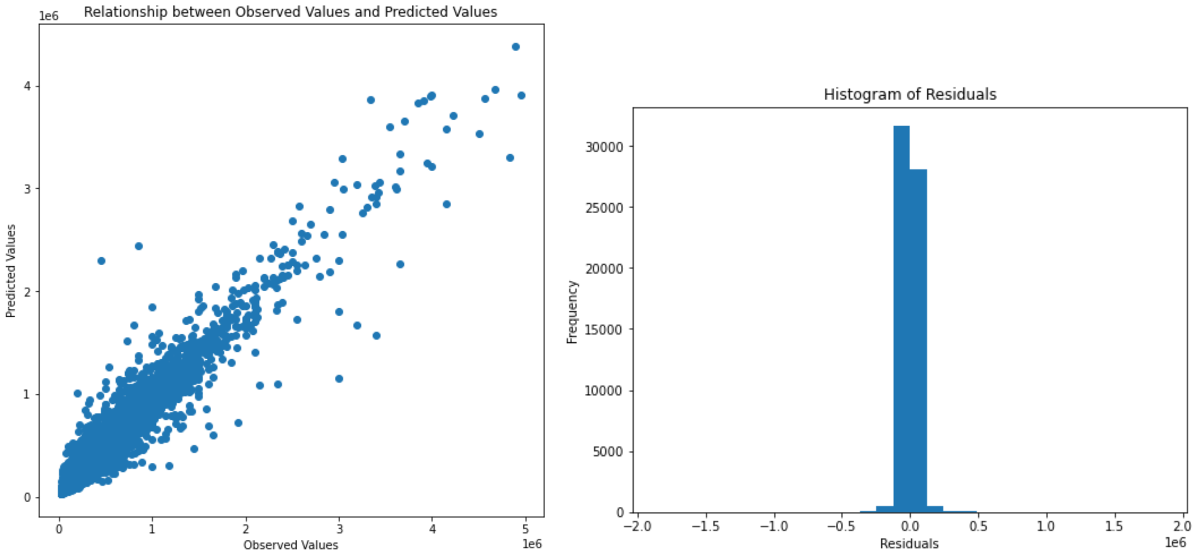

Figure 28 - Scatter plot and Histogram of residuals for predicting real estate prices in Lisbon Apartments using Random Forest

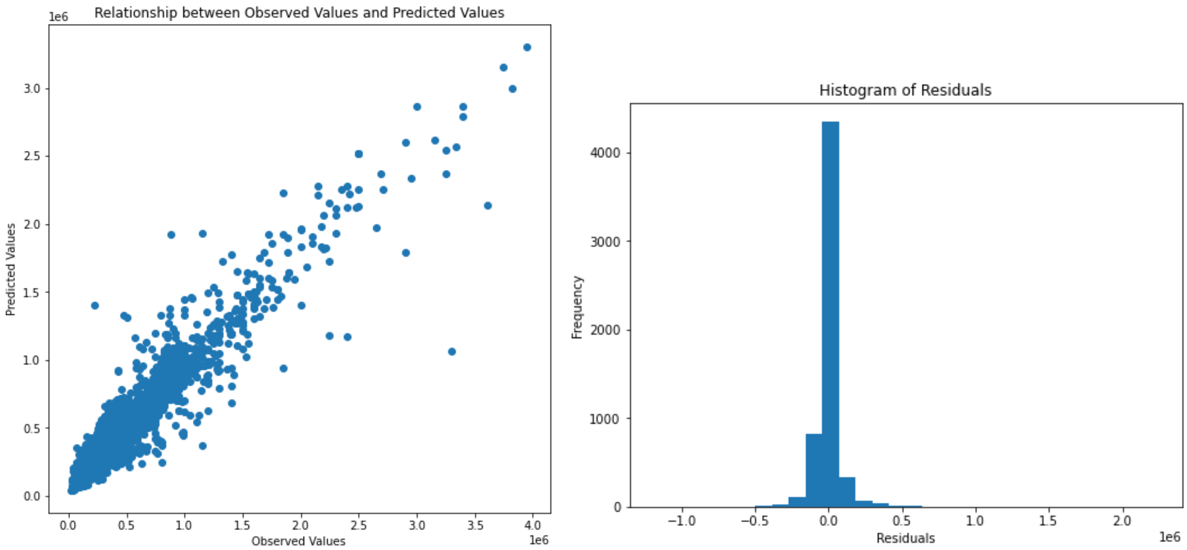

Figure 29 - Scatter plot and Histogram of residuals for predicting real estate prices in Lisbon Houses using Random Forest

From the scatter plot in Figures 28 and 29 of apartments and houses, it is apparent that both exhibit similar characteristics with a concentration of points around zero and an alignment along the diagonal line. The histogram in Figures 28 and 29 for apartments reveals two significantly tall columns concentrated around zero. On the other hand, in the case of houses, a prominent column centered at zero is followed by smaller columns in the vicinity of zero. This implies that predictions for apartments and houses are mostly close to the actual values, showcasing higher accuracy in predictions for both property types. The more prominent columns around zero signify a higher frequency of predictions closer to the actual prices, demonstrating more consistent performance in predicting property prices, in both models, with a significantly lower error.

This study shows that among the three real estate property price prediction models analyzed (Random Forest, Decision Tree and KNN), the Random Forest model becomes the first choice for real estate price prediction. The results of the scatter plot and histogram analysis show that the prediction performance for the two property types, apartments, and houses, is consistent, with deviations close to zero. Alignment along the diagonal shows that the predicted and actual values of the two models are in good agreement. These graphical observations are consistent with the statistical results presented in table 8, where the Random Forest model has the lowest error margin, followed by the Decision Tree model, while the KNN model has relatively high prediction errors.

Correlating these graphical observations with the statistical metrics presented earlier, in table 8, it is clear that Random Forest models consistently outperform Decision Tree and KNN models. The points around the zero point in the scatterplot and the higher columns in the histogram are significantly concentrated, especially in the case of Random Forest, confirming the high prediction accuracy. This is consistent with the results of the performance table and supports the conclusion that the Random Forest model has a lower error rate compared to other models, making it the most reliable model for predicting property real estate prices.

5. CONCLUSIONS AND FUTURE WORKS

The purpose of this study was to predict real estate property prices and gain a deeper understanding of the Portuguese property market. To achieve this goal, a thorough investigation was conducted into the most appropriate methods to handling the data at hand, considering the nature and historical context of the data. The literature review plays a crucial role in selecting the different methods that best suit the research objectives.

Data preprocessing includes data cleaning, exploratory data analysis, and examination of trends and correlations over time. This allowed us to calculate property prices across Portugal and analyze differences between regions, highlighting the impact of different variables and geographical features on price disparities.

The results obtained from the predictive models demonstrated the effectiveness of the chosen approach. These models show extremely high accuracy, with few, relatively minor errors, which may have limited impact within the context of the real estate market.

This main advantage of this study is the ability to predict future developments and provide customers with valuable insights to make informed decisions. This outcome is consistent with the main objectives of the project.

Despite its important findings, this work has certain limitations that must be acknowledged. Some have to do with the dataset itself. In particular, the lack of certain characteristics, such as the number of bathrooms, the presence of an elevator or the floors of the property, creates challenges in determining the price of a property. Due to the omission of these variables and the lack of clear correlations in the data set, these features were not added.

Furthermore, the dataset contains a large number of missing values, making data imputation a complex and less reliable process. While potentially valuable, excluding latitude and longitude data can provide precise geographic coordinates that allow you to determine the exact location of a property. In many cases, this information can be derived from postal codes already included in the data set. However, it is important to note that postal codes cover larger areas and different properties within the same postal code may have different coordinates. For example, two houses with the same zip code but located on opposite sides of a street or in different buildings will have different latitude and longitude coordinates. Therefore, the accuracy of latitude and longitude in this dataset is inconsistent. Future work should focus on using coordinates to improve accuracy.

Economic, political, and social factors also have a significant impact on the real estate market over time. Models based on historical data may not accurately predict such changes. Furthermore, the skewed price distribution of the dataset affects forecast accuracy, especially in extreme cases.

This study marks an important step in fully understanding the Portuguese dynamic real estate market. By using the power of predictive models and geospatial data, is possible to provide valuable insights to customers and stakeholders to shape a more well-known real estate market.

In view of these limitations, it is important to realize that this study represents a snapshot of the dynamic real estate market and provides the foundation for future work and improvement.

Realizing this work brings it closer to the ultimate goal: to help individuals, companies and investors make data-oriented decisions in dynamic and changing real estate landscape. I hope this research can support and encourage further research in this field.

BIBLIOGRAPHICAL REFERENCES

Agnello, L., & Schuknecht, L. (2011). *Booms and busts in housing markets: Determinants and implications*.

Ahmad, N., Menegaki, A. N., & Al-Muharrami, S. (2020). Systematic Literature Review of Tourism Growth Nexus: An Overview of the Literature and a Content Analysis of 100 Most Influential Papers. *Journal of Economic Surveys*, *34*(5), 1068–1110. https://doi.org/10.1111/joes.12386

Ahn, J. J., Byun, H. W., Oh, K. J., & Kim, T. Y. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, *39*(9), 8369–8379. https://doi.org/10.1016/j.eswa.2012.01.183

Akinsomi, O. K., Mkhabela, N., & Taderera, M. (2018). The role of macro-economic indicators in explaining direct commercial real estate returns: Evidence from South Africa. *Journal of Property Research*, *35*, 1 to 28. https://doi.org/10.1080/09599916.2017.1402071

Alhowaish, A. K. (2016). Is Tourism Development a Sustainable Economic Growth Strategy in the Long Run? Evidence from GCC Countries. *Sustainability*, *8*(7), Article 7. https://doi.org/10.3390/su8070605

Alkali, M., Sipan, I., & Razali, M. (2018). An Overview of Macro-Economic Determinants of Real Estate Price in Nigeria. *International Journal of Engineering and Technology(UAE)*, *7*, 484–488. https://doi.org/10.14419/ijet.v7i3.30.18416

Badarinza, C., & Ramadorai, T. (2018). Home away from home? Foreign demand and London house prices. *Journal of Financial Economics*, *130*(3), 532–555. https://doi.org/10.1016/j.jfineco.2018.07.010

Baffoe-Bonnie, J. (1998). The Dynamic Impact of Macroeconomic Aggregates on Housing Prices and Stock of Houses: A National and Regional Analysis. *The Journal of Real*

Estate Finance and Economics, *17*(2), 179–197. https://doi.org/10.1023/A:1007753421236

Bahia, I. S. H. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. *International Journal of Intelligence Science*, *3*(4), Article 4. https://doi.org/10.4236/ijis.2013.34017

Barron, K., Kung, E., & Proserpio, D. (2021). The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb. *Marketing Science*, *40*(1), 23–47. https://doi.org/10.1287/mksc.2020.1227

Biagi, B., Brandano, M. G., & Lambiri, D. (2015). Does Tourism Affect House Prices? Evidence from Italy. *Growth and Change*, *46*(3), 501–528. https://doi.org/10.1111/grow.12094

Brzezicka, J. (2020). Towards a Typology of Housing Price Bubbles: A Literature Review. *Housing, Theory and Society*, *38*, 1–23. https://doi.org/10.1080/14036096.2020.1758204

Buonanno, P., Montolio, D., & Raya-Vílchez, J. M. (2013). Housing prices and crime perception. *Empirical Economics*, *45*(1), 305–321. https://doi.org/10.1007/s00181-012-0624-y

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS International Journal of Geo-Information*, *7*(5), Article 5. https://doi.org/10.3390/ijgi7050168

Cellmer, R., & Trojanek, R. (2020). Towards Increasing Residential Market Transparency: Mapping Local Housing Prices and Dynamics. *ISPRS International Journal of Geo-Information*, *9*(1), Article 1. https://doi.org/10.3390/ijgi9010002

Cheng, I.-H., & Xiong, W. (2014). Financialization of Commodity Markets. *Annual Review of Financial Economics*, *6*(1), 419–441. https://doi.org/10.1146/annurev-financial-110613-034432

Cocola-Gant, A., & Gago, A. (2021). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. *Environment and Planning A: Economy and Space*, *53*(7), 1671–1688. https://doi.org/10.1177/0308518X19869012

Crowe, C., Dell'Ariccia, G., Igan, D., & Rabanal. (2012). *Policies for Macrofinancial Stability: Managing Real Estate Booms and Busts*.

Cunha, A. M., & Lobão, J. (2021). The effects of tourism on housing prices: Applying a difference-in-differences methodology to the Portuguese market. *International Journal of Housing Markets and Analysis*, *15*(4), 762–779. https://doi.org/10.1108/IJHMA-04-2021-0047

de Ville, B. (2013). Decision trees: Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455. https://doi.org/10.1002/wics.1278

Dehesh, A., & Pugh, C. (2000). Property Cycles in a Global Economy. *Urban Studies*, *37*(13), 2581–2602. https://doi.org/10.1080/00420980020080701

del Cacho, C. (2010, December 11). *A comparison of data mining methods for mass real estate appraisal* [MPRA Paper]. https://mpra.ub.uni-muenchen.de/27378/

Fama, E. F. (1991). Efficient Capital Markets: II. *The Journal of Finance*, *46*(5), 1575–1617. https://doi.org/10.1111/j.1540-6261.1991.tb04636.x

Fletcher, M., Gallimore, P., & Mangan, J. (2000). Heteroscedasticity in hedonic house price models. *Journal of Property Research*, *17*(2), 93–108. https://doi.org/10.1080/095999100367930

Foryś, I., & Putek-Szeląg, E. (2017). THE IMPACT OF CRIME ON RESIDENTIAL PROPERTY VALUE - ON THE EXAMPLE OF SZCZECIN. *Real Estate Management and Valuation*, *25*(3), 51–61. https://doi.org/10.1515/remav-2017-0022

Ghodsi, R., Boostani, A., & Faghihi, F. (2010). Estimation of Housing Prices by Fuzzy Regression and Artificial Neural Network. *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, 81–86. https://doi.org/10.1109/AMS.2010.29

Goodhart, C., & Hofmann, B. (2008). *House Prices, Money, Credit and the Macroeconomy.*

Grum, B., & Govekar, D. K. (2016). Influence of Macroeconomic Factors on Prices of Real Estate in Various Cultural Environments: Case of Slovenia, Greece, France, Poland and Norway. *Procedia Economics and Finance*, *39*, 597–604. https://doi.org/10.1016/S2212-5671(16)30304-5

Han, J. H. (2018). Comparing Models for Time Series Analysis. *Wharton Research Scholars*. https://repository.upenn.edu/wharton_research_scholars/162

Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5

Hishamuddin, M., Ali, H., Achu, K., Binti, R., Abdul Jalil, R., Folake, A., & Yakub, A. (2020). Journal of Critical Reviews THE EFFECT OF ADOPTING MICRO AND MACRO-ECONOMIC VARIABLES ON REAL ESTATE PRICE PREDICTION MODELS USING ANN: A SYSTEMATIC LITERATURE REVIEW. *Journal of Critical Reviews*, *7*, 2020. https://doi.org/10.31838/jcr.07.11.88

Housing price statistics—House price index. (2023). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Housing_price_statistics_-_house_price_index

Hush, D. R., & Horne, B. G. (1993). Progress in supervised neural networks. *IEEE Signal Processing Magazine*, *10*(1), 8–39. https://doi.org/10.1109/79.180705

Huy, D. T. N., Nhan, V. K., Bich, N. T. N., Hong, N. T. P., Chung, N. T., & Huy, P. Q. (2021). Impacts of Internal and External Macroeconomic Factors on Firm Stock Price in an Expansion Econometric model—A Case in Vietnam Real Estate Industry. In N. Ngoc Thach, V. Kreinovich, & N. D. Trung (Eds.), *Data Science for Financial Econometrics* (Vol. 898, pp. 189–205). Springer International Publishing. https://doi.org/10.1007/978-3-030-48853-6_14

Inflation in the euro area. (2023). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Inflation_in_the_euro_area

Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW. *PLANNING MALAYSIA*, *19*. https://doi.org/10.21837/pm.v19i17.1018

Kasparova, D., & White, M. (2001). The Responsiveness of House Prices to Macroeconomic Forces: A Cross-Countr Y Comparison. *European Journal of Housing Policy*, *1*(3), 385–416. https://doi.org/10.1080/14616710110091561

Li, L., & Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters. *2017 International Conference on Applied System Innovation (ICASI)*, 87–90. https://doi.org/10.1109/ICASI.2017.7988353

Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. *Scientific Programming*, *2022*, e5750354. https://doi.org/10.1155/2022/5750354

Mcleod, S. (2018). *Maslow's Hierarchy of Needs*.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

My-Linh. (2020). The Hedonic Pricing Model Applied to the Housing Market. *International Journal of Economics and Business Administration*, *VIII*(Issue 3), 416–428. https://doi.org/10.35808/ijeba/526

Nwagwu, U., Ayinde, A., & Olasoji, Y. (2021). Application of Instance Learning Algorithms to Analyze Logistics Data. *International Journal of Engineering Research*, *10*(07).

Oikarinen, E. (2014). *Studies on housing price dynamics*. https://www.utupub.fi/bitstream/handle/10024/99387/Ae9_2007Oikarinen.pdf?seque

Oshiro, T., Perez, P., & Baranauskas, J. (2012). How Many Trees in a Random Forest? In *Lecture notes in computer science* (Vol. 7376). https://doi.org/10.1007/978-3-642-31537-4_13

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

Pilinkienė, V., Stundziene, A., Stankevičius, E., & Grybauskas, A. (2021). Impact of the Economic Stimulus Measures on Lithuanian Real Estate Market under the Conditions of the COVID-19 Pandemic. *Engineering Economics*, *32*(5), Article 5. https://doi.org/10.5755/j01.ee.32.5.28057

Postal, P. C. (2018, January 25). *Códigos de Freguesia e Concelho por distrito*. Portal Código Postal. https://codigopostal.ciberforma.pt/ferramentas/codigos-de-freguesia-e-concelho-por-distrito

Proder. (2020a). *Freguesias Rurais*. http://www.proder.pt/ResourcesUser/Documentos_Diversos/33/PDRc_Freg_ZRurais_NUTIIs_rev2_corrigido.pdf

Proder. (2020). *Zonas_Rurais*. https://www.gpp.pt/images/Estatisticas_e_analises/Estatisticas/associadasmedidasapoio/Territorios_Rurais.pdf

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*. https://doi.org/10.21837/pm.v17i9.581

Robert N, L. (1995). A nonfinancial business success versus failure prediction model for young firms. *Journal of Small Business Management 33(1):8-20.*

Schäfer, P., & Hirsch, J. (2017). Do urban tourism hotspots affect Berlin housing rents? *International Journal of Housing Markets and Analysis*, *10*(2), 231–255. https://doi.org/10.1108/IJHMA-05-2016-0031

Shearer, C. (2020). *The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, Volume 5, Number 4, Fall 2000, page. 13-22.*

Shinde, N., & Gawande, K. (2018). *Survey on predicting property price*. 1–7. https://doi.org/10.1109/ICACE.2018.8687080

Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*(2), 208–219. https://doi.org/10.1007/s13198-020-00946-3

Tabales, N., María, J., & Carmona, R. (2013). *Artificial Neural Networks for Predicting Real Estate Prices*.

TIOBE Index. (2023). TIOBE. https://www.tiobe.com/tiobe-index/

Tripathi, S. (2019). *Macroeconomic Determinants of Housing Prices: A Cross Country Level Analysis*.

Weinberger, K. Q., & Saul, L. K. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research, 10(2).*

Wing, C. K., & Chin, T. (2003). A Critical Review of Literature on the Hedonic Price Model. *International Journal for Housing Science and Its Applications*, *27*, 145–165.

World Travel Awards 2022 Europe Winners. (2022). World Travel Awards. https://www.worldtravelawards.com/winners/2022/europe

Yasnitsky, L. N., Yasnitsky, V. L., & Alekseev, A. O. (2021). The Complex Neural Network Model for Mass Appraisal and Scenario Forecasting of the Urban Real Estate Market Value That Adapts Itself to Space and Time. *Complexity*, *2021*, e5392170. https://doi.org/10.1155/2021/5392170

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa