TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Machine and Deep Learning Models for House Price Prediction in United States of America and Portugal

Catarina de Freitas Sanchez de la Fuente

Master in Computer Engineering and Business Management

Supervisor:
PhD, Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD, Rúben Filipe de Sousa Pereira, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

November, 2022

iscte

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Machine and Deep Learning Models for House Price Prediction in United States of America and Portugal

Catarina de Freitas Sanchez de la Fuente

Master in Computer Engineering and Business Management

Supervisor:
PhD, Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD, Rúben Filipe de Sousa Pereira, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

November, 2022

To all my friends and family

Acknowledgments

In the first place, I would like to express my sincere gratitude to ISCTE for giving the students conditions to continue their studies during the pandemic.

Secondly, I have to thank my Supervisor, Ricardo Daniel Santos Faro Marques Ribeiro, Ph.D., and my Co-supervisor, Rúben Filipe de Sousa Pereira, Ph.D., for sharing their knowledge and for their dedicated help during the past year. Most of all, for being two people that I consider as two models to follow both academically and professionally.

I wish to extend my special thanks to João Antão for his support about Real Estate subjects and for being always available to help me.

I have a debt of gratitude to my class, united and ready to help, with a special praise for Tiago Nóbrega.

Finally, many thanks to my family and friends, because without them it would have been much more difficult.

A big thank you to my mother for her hard work and excellent performance. She was always ready to help and in a good mood.

To my grandmother who is always worrying about me, but also always willing to help me.

To my uncles for the motivation and to my cousin Mário for his advice and positive thinking.

I would like to thank Ricardo Belo for his constant support, always available even when he could not be present.

Thank you to Bruno Santos for his tireless work and inspiration about new technologies.

To my friend Mariana Filipe, dedicated teacher and my favourite cello player, who never gave up on me.

I am also grateful to Beatriz Amaral for her friendship during 26 years and as always a source of strength and motivation.

I would like to recognize the support and special friendship of Ricardo Cordeiro, a source of inspiration and someone that always believed in me.

Resumo

A presente estudo utilizou a medotologia CRISP-DM para a caraterização e descrição do processo de desenvolvimento de um sistema para estimação dos preço das casas em Portugal. Duas fases importantes no processo foram: a extração de dados e a comparação entre vários algoritmos. A extração de dados foi realizada através de técnicas de Web Scraping a partir do *site* Mais Consultores [1]. Utilizaram-se métodos de Text Mining - Rule-based Matching e Similarity – para estruturar e retirar significado da informação que se extraiu do site. De seguida, realizámos a comparação entre a aplicação de algoritmos de Machine Learning e Deep Learning: Support Vector Machines (SVM), Decision Tree Regressor (DTR), Random Forest, K-Nearest Neighbour (KNN), Artificial Neural Networks ANN, Convolutional Neural Networks CNN, Recurrent Neural Networks RNN, Multi-layer Perceptron MLP and Long Short-term Memory LSTM Network.

Encontrar esta solução constituiu a principal motivação da presente tese. Os resultados obtidos pelos algoritmos utilizados, tanto os de Machine Learning como os de Deep Learning, demonstram que os algoritmos precisavam de mais dados para treino. Adicionalmente, os algoritmos com melhores resultados, i.e., com menor Mean Absolute Error (MAE), Mean Square Error (MSE) e Root Mean Square Error (RSME) e maior score foram os algorimos de Deep Learning.

Palavras chave: *Previsão Preços das Casas*, *Machine Learning*, *Deep Learning*, *Text Mining*

Abstract

The present study describes the development process of a system to predict the houses' prices in Portugal. Two main phases of this process were the data extraction and the comparison among several algorithms. Data Extraction was made through Web Scraping techniques applied the Mais Consultores site [1]. This study used Text Mining methods - Rule-based Matching and Similarity – in order to structure and obtain meaning from the information extracted. Afterwards, this thesis made a comparison among the application of Machine Learning and Deep Learning algorithms: Support Vector Machines (SVM), Decision Tree Regressor (DTR), Random Forest, K-Nearest Neighbour (KNN), Artificial Neural Networks ANN, Convolutional Neural Networks CNN, Recurrent Neural Networks RNN, Multi-layer Perceptron MLP and Long Short-term Memory LSTM Network.

Finding this solution was the prime motivation of the present thesis.

The results obtained by the used algorithms, both Machine Learning and Deep Learning, demonstrated that the algorithms needed more data for the training set. Additionally, the algorithms with the best results, i.e., with the lesser value of Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RSME) and the better score were the Deep Learning algorithms.

Keywords: *House Price Prediction*, *Machine Learning*, *Deep Learning*, *Text Mining*

Contents

List of Figures

List of Tables

Acronyms

AE: Attribute Extraction.
AI: Artificial intelligence.
ANN: Artificial Neural Networks.
API: Application Programming Interface.
AV: Attribute Value.

CNN: Convolutional Neural Network.
CRISP-DM: Cross Industry Standard Process for Data Mining.

DL: Deep Learning.
DM: Data Mining.
DSS: Decision Support System.
DT: Decision Tree.
DTR: Decision Tree Regressor.

HTML: HyperText Markup Language.

ID3: Iterative Dichotomiser 3.

KNN: K-Nearest Neighbors.

LSTM: Long Short-Term Memory Network.

MAE: Mean Absolute Error.
MAPE: Mean Absolute Percentage Error.
ML: Machine Learning.
MLP: Multi-layer Perceptron Regressor.
MPL: Multi-layer Perceptron.
MSE: Mean Squared Error.

NER: Named Entity Recognition.
NLP: Natural Language Processing.

RBN: Radial Basis Network.
RF: Random Forest.
RMSE: Root Mean Squared Error.
RNN: Recurrent Neural Networks.

RQ: Research Questions.
RSME: Root Mean Square Error.

SJR: Scimago Journal & Country Rank.
SLR: Systematic Literature Review.
SVM: Support Vector Machine.

TM: Text Mining.

USA: United State Of America.

CHAPTER 1

Introduction

Currently, a large quantity of data is available on the Internet and in organizations' databases. The power of data is becoming increasingly obvious. The transformation of data into knowledge translates to the faster growth of skills that, without working with data, would take much longer and allocate much more resources.

Data is not the same as knowledge. Data must be cleaned and processed to have meaning. Through Artificial Intelligence (AI) and Computer Science together, we find this meaning and create a specific knowledge. AI and Computer Science combined solve not only problems related to IT but also in other areas. In short, this combination solves real-life problems [2], [3].

The AI is present in many businesses use case, such as Microsoft which has worked hard on investigating and developing Enterprise Knowledge tools. It has improved the development of an automatic extraction system that contains business knowledge and transfers it to a platform. This platform was implemented with a specific structure for each organization with distinct entities and its relationship [4].

According to Ethan Cohen and Afraz Jaffri from Oracle, AI investment strategy should align closely with the company's strategy. AI operating models should enable quick changes without needlessly altering the organizational structure or business structure [5].

It is not only the Technologies sector that can improve through the application of AI. Another sector that has been the target of researchers is Real Estate.

For example, the Zillow - Real Estate Agency - in the USA AI predicted the price of sale of houses from its databases [6], [7]. Besides the site available for estimation of the price of current houses, Zillow organized a competition five years ago. It made available more than a million data, between 2016 and 2017, to the competition participants for them to try to predict the logerror - the difference between the algorithms of the price of sale and the actual price.

In Portugal, Alfredo, a startup built their web page based in AI, intending to suggest houses with the characteristics and price requested by the customer [8].

Searching for the ideal house is a complex process. In addition to the price variant, other factors can influence the decision to buy. For example, the number of bedrooms and bathrooms, the areas of the rooms, the condition in which it is, remodelled, new or for remodelling, the location and the type of services and transports nearby, among others.

The agent or client role is a process that involves human resources when searching for a house. Looking for an ideal property with specific characteristics can be time-consuming and hard work for both the customer and the agent.

This study presents a solution. It implements a Rules-based Matching with the Spacy Library – Text Mining (TM) Library - to structure the data extracted from Mais Consultores' web page [1][9]. It also develops four Machine Learning (ML) and five Deep Learning (DL) models to predict house prices in Portugal. According to Ronal Coase, winner of the Nobel Prize for Economics Science, *If you torture the data long enough, it will confess to anything* [10].

1.1. Motivation and Objective

The purpose of the real estate sector is to provide all services necessary to satisfy the customer when acquiring a service. A service can be an acquisition, selling, leasing or trespassing of a particular house. This study does not include other types of properties such as a terrain, a building, a shop, etc. It focuses on houses.

In this dissertation, we only consider the purchase of a property as a service, ignoring transfers, rents or another type of service. The agency or the agent must understand what the customer needs and, in return, present solutions that agree with the customer's request. The search should be quickly solved to keep up with market changes and align with the client's expectations. Specifically, the real estate market in Portugal has been changing quickly and it is not easy to follow up on these shifts [11].

In this context, the primary motivation is developing a system to predict house prices in Portugal. This system has as input the possible features of a house and then calculates its price.

The first module consists of identifying software or applications with the knowledge to predict house prices at international and national levels. We found a data set with a similar purpose - Zillow data set from the *Zillow Prize: Zillow's Home Value Prediction (Zestimate)* Competition - which allowed the study of several participants' investigations. We concluded that the results of these investigations showed a good performance and a small margin of error.

The next module pretends to create a data set with Portuguese data extracted from Mais Consultores' web page. All the columns in the Portuguese data set are related to a specific house or are external factors related to it. Mais Consultores is a network of real estate consultants and a platform for clients to acquire a service. We can role-play real estate agents and clients on the web page. If choosing the second, all the houses for sale in Portugal published on Mais Consultores can be seen. The last module contains the implementation of nine models with the data extracted from Mais Consultores and the comparison of these results with other investigations [1].

1.2. Methodology

The implemented Methodology included: Problem Understanding (1), Data Extraction (2), Data Cleaning (3), Data Pre-processing (4), Data Modeling and Data Deployment (5) and Data Results (6).

Problem Understanding was the initial task: researching and analysing as much information on the topic as possible. Point (1) refers to the initial module where we searched on the Internet for tools to predict house prices with two different methods: *ad hoc* and a Systematic Literature Review SLR. The first method consisted of searching the Internet for tools with the same purpose as this thesis. The second method was the process of building a SLR.

Data Extraction was the task of extracting data from Mais Consultores' [1] page to an Excel spreadsheet. Point (2) is the procedure of data extraction made through Web Scraping a technique useful to collect information from web pages. In our case, we extracted the data from a HTML page with Selenium library [1], [12], [13].

Data Cleaning was a process of dealing with null and not normalized data present in the Portuguese data set from Mais Consultores Point (3) consists of data normalization and consequent reduction of the noise. This was composed by special characters, symbols and irrelevant information. When possible, i.e., when the data set did not lose its meaning, the decision was to remove the null data. When was not possible to remove them, they were normalized. The normalization consisted of applying Text Mining (TM) techniques with Spacy Library: Tokenization, Rule-based Entity, Token-based Matching and Similarity. Additionally, we used regular expressions for string manipulation [9], [14], [15].

Data Pre-processing was the procedure of data preparation to be apply in the models. Point (4) is composed of several processes, such as encoding categorical features in numbers. The purpose of pre-processing was to adapt the data to be processed and understood by the models in the best possible way. This included the encoding of categorical features with the objective of normalizing data, i.e., all the data must be numeric and not null.

In Data Modeling and Deployment, the intention was to create and deploy glsML and DL models. Point (5) executes the following algorithms: K-Nearest Neighbors (KNN), Decision Tree Regressor (DTR), Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Radial Basis Networks (RBN), Multi-layer Perceptron Regressor (MLP), Recurrent Neural Networks (RNN) and Long Short-Term Memory Networks (LSTM).

Data Results was composed by the interpretation of the results of Point (5). Additionally, Point (6) includes the adjustment of the models' parameters to improve the results. Finally, this investigation presents the conclusions and improvements for the future.

1.3. Research Questions

The Research Questions allowed for decreasing the complexity of the project and managing to focus on the subject of this work. What is the best approach for predicting the price of a house with specific characteristics? What if the data were in different languages? Would the result be the same? Which algorithms were used for solving that type of problem? Could the content of the articles be useful to identify the best approach for this thesis? With the previous points in mind, we formulated the following research questions (RQ):

- **RQ1:** Which techniques and algorithms are adequate for predicting the price of houses in the Real Estate Sector?
- **RQ2:** Are there applications/software for predicting the price of houses in the Real Estate Sector? If there are, which are they, where are they from and how do they work?
- **RQ3:** If a predicting model uses a data model and features similar to the Zillow data set, do the results continue to be positive?

1.4. Document Structure

The present Chapter, *Introduction*, provides the baselines for this study. It defines the objective, the motivation, the strategy and the structure of the document.

Chapter 2, *Background*, transmits the knowledge necessary to understand the scenario behind the project's theme. This Chapter describes the two technologies used to build the system aimed to predict the price of houses: ML and DL. Subsequently, it also describes the chosen algorithms of each technology included in the development phase: KNN, DTR, GridSearchCV, SVM, CNN, MLP, RNN and LSTM.

Chapter 3, *Related Work*, describes the SLR and the answers to the research questions. Chapter 4, *Research Methodology*, contains the development of the CRISP-DM modulation.

Chapter 4, *CRISP-DM*, contains the research methodology. This chapter includes the phases of the CRISP-DM as well as the tasks for the construction of the house price prediction system in Portugal.

Chapter 5, *Results*, contains the results of the models' deployment and the respective explanation and the discussion, This chapter presents various regression methods applied to different data sets.

The last chapter, Chapter 6, *Conclusion* is reserved for the conclusions of this study. It presents future challenges and possibilities of improvement. The final part compares the results between the present study and the articles from previously studied literature.

CHAPTER 2

Background

Sections 2.1, 2.2, and 2.3 describe the ML and DL techniques and their algorithms. This study used the present algorithms to develop a house price estimation system in Portugal. The last sections address the methods used in data extraction and data preparation.

2.1. Overview

According to the British Cambridge Dictionary, knowledge is “Understanding of or information about a subject that you get by experience or study, either known by one person or by people generally”. In the American Cambridge Dictionary, “Awareness, understanding, or information that has been obtained by experience or study, and that is either in a person’s mind or possessed by people generally” [16].

Another way to understand knowledge and how to get it is the concept of Knowledge Acquisition. Knowledge Acquisition is the discovery of interesting facts about a subject. In our case study, we wanted to discover how the features that characterize a house can influence its sale price. Knowledge Acquisition can be expressed in two forms: static or dynamic knowledge. The first one does not change over time, while the dynamic knowledge may change due to the environment and internal or external factors that may affect it. Therefore, some studies will always be valid due to the presence of static knowledge and in others the concepts will evolve through the years and their application will not be valid or will need to be updated to be true [17]–[19]. This thesis used Knowledge Acquisition - dynamic knowledge - and this knowledge is valid in the Portuguese data set. The real estate market in Portugal or in another part of the globe is not constant and has changed over time. The data set has the actual prices of houses in Portugal. For the evolution of the prices to be always true, we need to update the extraction module and the models’ input.

Knowledge Acquisition can be obtained through three methods. The first method consists of interviews. This method raises certain doubts about time management and failure at the automation level since the interviews are conducted and analysed individually by a human being [18], [19].

The second method is the implementation of a Knowledge Acquisition System (KAS). The importance of knowledge generated over the years has increased the development of studies about KAS. The system consists of a set of rules without the intervention of none human intelligence. With KAS and its automated tasks, the organization has a lower probability of human error and there is an improvement in the processes’ speed and in their reliability. On the other hand, the technical and engineering professionals have a much higher probability of making mistakes, they represent a costly resource for the company and they are slower assets concerning productivity [18]–[21].

The third method is the use of ML and DL techniques. They are some of the most used today due to the characteristics of their automation process: reduction of time processing and analysis time. This method reduces the use of human resources and the possibility of human error. This investigation implemented these methods to solve the real estate problem of predicting houses' price in Portugal [18].

ML is part of AI. It consists of mathematical models to help the machine learn from data and create helpful information for human beings. The machine plays the role of learning human beings. For this, it has to learn the data with the training set and then it is tested with the data present in the test set. In the training process the machine learns and in the testing process it applies the knowledge acquired in the training set to the testing set, in order to be able to classify or predict a specific target [2], [3].

To support this process, ML uses programming languages and libraries that implement ML algorithms to create prediction models, classification models, topic models, sentiment analysis models, text analysis models, recommendation models, etc [22], [23].

According to William Blake, poet, painter and typographer, "The true method of knowledge is experimentation." This thought translates into the development of the house price prediction system. The process has been improved through experimentation [24].

2.2. Artificial Intelligence

First, AI is composed of several subsets concerning the creation of human intelligence in machines. Two of them are ML and DL, as observed in Figure 1 [25].

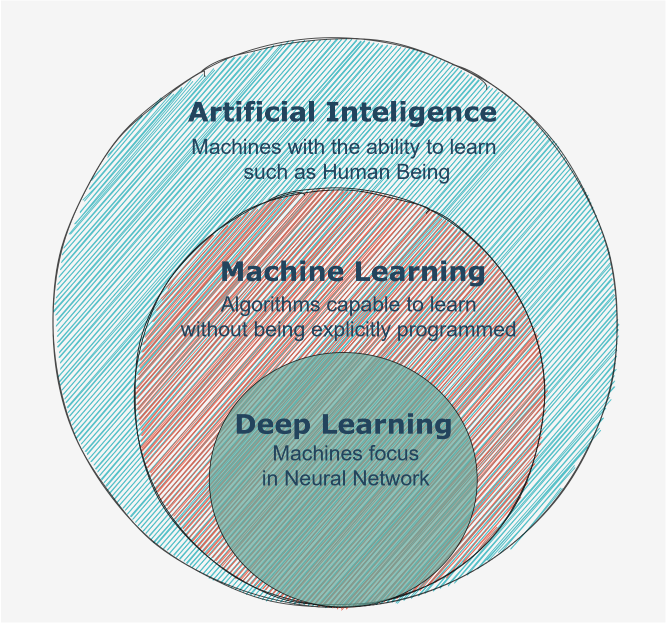

Artificial Inteligence
Machines with the ability to learn such as Human Being
Machine Learning
Algorithms capable to learn without being explicitly programmed
Deep Learning
Machines focus in Neural Network

Figure 1. Definition of Intelligence Artificial Branches: Machine Learning and Deep Learning

AI can also combine with other technologies. One of its transversal technologies is Data Mining DM. Data Mining (DM) is a process of data extraction and its conversion to knowledge [25]–[27].

The main tasks of DM are the definition of Predictive and Descriptive models. The Predictive model predicts unknown values with variables such as input. The Descriptive model finds patterns in the data. At a higher level, these models constitute a Decision Support System (DSS) [28], [29].

Over the years, organizations have realized the power of data in decision-making. A Decision Support System (DSS) can be a useful tool to increase decision-making knowledge and improve company results [28], [29].

2.2.1. Machine Learning

As observed in Figure 2, other subsets of ML techniques are represented in three types of learning: Supervised Learning, Unsupervised Learning and Reinforcement Learning [28].

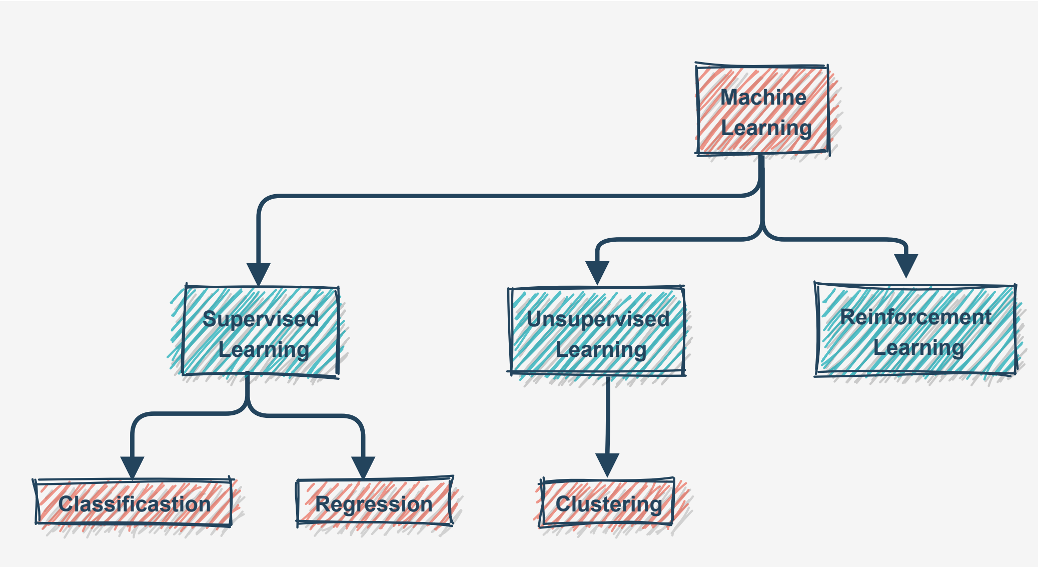


FIGURE 2. Definition of Machine Learning Types: Supervised Learning, Unsupervised Learning and Reinforcement Learning

Supervised Learning is a process divided into two groups: classification and regression algorithms. This distinction is made because there are real-life problems more easily solved with classification algorithms and others with regression algorithms. In some cases, due to the data set characteristics, the two types of algorithms can be used and provide positive results. The choice of the most adequate algorithm is influenced by the type of data and the purpose of solving the problem [23], [30].

Classification is the process of finding a function or functions to classify the data set by specific labels. The algorithm's objective is data learning and data classification of new elements [23], [28], [30].

Regression is the process of finding correlations between independent and dependent variables. This process is used as a descriptive method, i.e., it discovers patterns and makes predictions. The algorithm's objective is data learning to predict a specific target in a specific problem [23], [28], [30].

Unsupervised Learning is the process used to aggregate data in clusters without labels. Even without labels, the algorithm can discover patterns without human intervention. The algorithm can collect relevant information: similarities and differences [28], [30].

Reinforcement Learning is a process of automation of sequential tasks. This algorithm aims to optimize sequential decisions and find the best strategy to solve the problem [23], [28], [30].

The classification and regression algorithms divide the data into three sets: training, testing and, in specific studies, validation. Usually, the training set has 70% to 80% of the data, the testing set 20% to 30% and the validation set 10% [23], [28].

The purpose of the sets' division aims to allow for the algorithm to distinguish among training, testing and validation data. The data present in the training set must be different than the data present in the testing and validation sets. The objective of the algorithm is to classify and predict the target without being influenced. If the algorithm knows all the data, it manages to predict with 100 % accuracy. Not knowing the testing data, the algorithm manages to classify or predict depending on its knowledge of the training set [28].

It is from the training data that the algorithm learns. Afterwards, we test if the algorithm learned the data through the testing set. During this phase, the algorithm has to predict or classify data until then unknown [28].

The data concerning the validation set, also composed by different data from the other sets, allow to verify the algorithm accuracy and precision. If the previous sets had the same data, we could not be able to prove that the algorithm predicted or classified correctly new data [28].

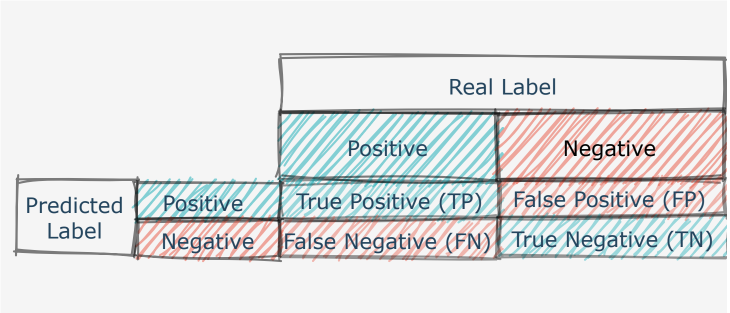
The accuracy, precision and recall of the model express the quality of the algorithm's prevision and classification. The equation 2.1 shows the accuracy formula, with the number of Correct Previsions (CP) and the total of elements present in a specific data set, represented by Total Data (TD) [28], [30].

$$Accuracy = \frac{CP}{TD} \tag{2.1}$$

Another form of calculating accuracy is the calculation of true positives (TP) and negatives (TN) and false positives (FP) and negatives (FN), as illustrated in Figure 3 [28], [30].:

TP are correctly predicted or classified by the algorithm. On the other hand, FP is the opposite of TP, representing the values that were incorrect. The more TP or TN exist, the better will be the accuracy and that means the algorithm has a good performance with that specific data set . The formula presents shown in the Equation 2.2 [28], [30]:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + TN} \tag{2.2}$$



		Real Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

FIGURE 3. Comparison between Real Label and Predicted Label with Positive and Negative Classifications

The precision takes into account the proportion of TP in the universe of positives, i.e., TP plus FP. Precision answers the question: what is the ratio between identified positives and true positives. One of the important results of the algorithm evaluation is precision. This is similar to accuracy, the following Equation 2.3 showing the differences 2.2 [28], [30]:

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

The recall is the ratio between current positives and correctly identified positives. The recall answers the question: what is the proportion of current positives in the correctly identified ones? Equation 2.4 explains the recall, i.e., the proportion of true positives:

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

Another way of thinking is to look at the error value instead of the positive and negative values. Let's look at three different types of errors: Mean Squared Error (MAE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

MAE is a model evaluation metric for regression problems. The MAE for each prediction is the error between the true value and the predicted value. The greater its value, the greater the error and the greater the distance between the predicted and the true values. Equation 2.5 explains the Mean Absolute Error MAE [23], [28], [30]:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \lambda(x_i)|}{n} \tag{2.5}$$

MSE is a model evaluation metric for regression problems. The MSE for each prediction is the square of the error between the true value and the predicted value. The greater its value, the greater the error and the greater the distance between the predicted and the true values. Equation 2.6 express the MSE [23], [28], [30]:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \lambda(x_i))^2}{n} \tag{2.6}$$

Where yi is the real value of the target value, for instance, xi, y(x) is the predicted value and n is the number of instances. RMSE is a metric of errors in problems where we are predicting a specific target. For example, RMSE is used in regression problems to measure how spread-out elements are. The metric calculates the concentration of data around a line that best fits them [30]. Equation 2.4 translate the RMSE [23], [28], [30]:

$$RMSE = \sqrt{\sum_{i=1}^{n} (\overline{y_i} - y_i)^2} \tag{2.7}$$

Where y is the predicted value and yi the number total of elements. Another model validation is R-squared - r^2 - for regression problems. This measure measures how well the data fit the model. The desired value of r^2 is 1, i.e., the closer to one, the better the data fit to the model. If the value is further from one, we have a bad fit to the model. SS_{res} is the residual sum of squares and is expressed in Equation 2.8 [31][23][22]:

$$SS_{res} = 1 - \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{2.8}$$

SS_{tot} is the total sum of squares and is translate in Equation 2.9 [31][23][22]:

$$SS_{tot} = 1 - \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{2.9}$$

r^2 is expressed by the following Equation 2.10 [31][23][22]:

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{\sum\limits_{i=1}^{n} (y_i - \hat{y})^2}{\sum\limits_{i=1}^{n} (y_i - \overline{y})^2} \tag{2.10}$$

ML algorithms are evaluated by the previous technologies. Some of them are Logistic Regression, Naive Bayes, Bayesian Network, Linear Classifier, Support Vector Machines (SVM), Decision Tree Regressor (DTR), Random Forest, k-Nearest Neighbour (KNN), Artificial Neural Network ANN, etc [22].

In the following sub-section, we demonstrate some of the mentioned algorithms. The algorithms selected are KNN, DTR, Random Forest, DTR and SVM. These were used for modeling the system to predict house prices in Portugal.

In the literature analyzed in chapter 3, the authors validated their algorithm by errors defined in this present chapter: RMSE, MSE, MAE and r^2.

In the studied literature, there were more evaluators of models with other measures. Therefore, the model measures are presented in chapter 3, *Related Work*.

2.2.2. Machine learning Algorithms

One of the most used and simplest ML algorithms is K-nearest Neighbors (KNN). This algorithm is a supervised learning algorithm for classification or regression problems. KNN was developed by the US military during the Second World War. In 1951, Evelyn Fix and Joseph Lawson Hodges Jr. developed a non-parametric pattern classification method called KNN. Their discovery was never published [32]–[34] There are two more dates worth mentioning. In 1967, Thomas Cover and Peter Hart published a paper on the multi-class KNN classification - *Nearest Neighbour Pattern Classification* - proving the upper bound error rate [32]–[34] In 1985, James Keller developed more complex pattern recognition procedures. His studies were able to prove a lower error rate [34].

The way of learning of KNN is called *Lazy Learner*. The algorithm creates a delay in the creation of the model. This algorithm memorizes all the training data and runs a model when it needs to classify or predict a label or a target [32][33][34]. The performance of KKN depends on the value of k being the best for a specific problem, depending on their neighbours as observed in Figure 4. KNN needs the definition of four components: the value of k, the number of neighbours, the distance metric and the classification model [32], [33]

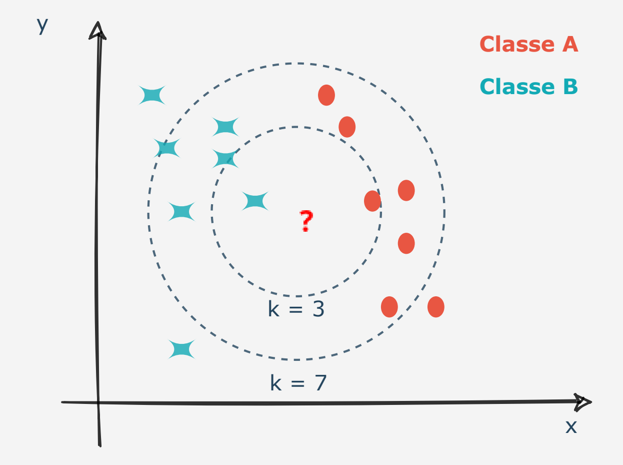


FIGURE 4. Graphic of K-nearest Neighbors (KNN) Algorithm and the Discovery of K

The choice of k is important because it can change the position of a class, i.e., the result of the classification varies dependent on k. If k is too small, the classifier can be sensitive to noise. On the other hand, if k is too big, it will be more difficult to process the data and the classes can have elements belonging to other classes. There is a method to find the best value of k. We will present it in the latest sections of this dissertation [32], [33].

When we use KNN to classify a new element present in the test set, we have to calculate the distance from all its neighbours. The closest to the element to be classified will be the one to give it a class. There are several metrics. We can consider Euclid's distance, Manhattan's distance, Minkowski's distance and Hamming's distance [32].

The Euclidean distance is for continuous variables. It is one of the most used and the least complex. Distance is a straight line in Euclidean space.

The Manhattan distance is also called taxi distance. The distance is calculated by adding the absolute differences of its Cartesian coordinates [35].

The Minkowski distance is for real-valued vectors. Calculating the distance is possible if the vector space is normalized. A vector represents space with a given length. These values cannot be negative [36].

The Hamming distance is calculated from binary data. The distance consists of comparing two binary data sets of equal size. Then, one by one, the data are similar or different. As a result, the Hamming distance gives the differences between the two data sets [37]. All the distances are for continuous variables, less the Minkowski distance. This distance is for categorical variables. The following equation explains the Euclidean Distance:

$$d = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{2.11}$$

The following equation explains the Manhattan Distance:

$$d = \sum_{i=1}^{k} |x_i - y_i| \tag{2.12}$$

The following equation explains the Minkowski Distance:

$$d = (\sum_{i=1}^{k}(|x_i - y_i|)^q)^{\frac{1}{q}} \tag{2.13}$$

The following equation explains the Hamming Distance:

$$D_H = \sum_{i=1}^{k} |x_i - y_i| \tag{2.14}$$

The Decision Tree (DT) is a famous algorithm. This algorithm is adequate for discovering knowledge, easily interpreting results and has a flexible implementation [38], [39].

This algoritm is a supervised learning algorithm with a tree structure. The algorithm is used for regression or classification problems.

DT was created in 1963 by the Department of Statistics of the University of Wisconsin - Madison. The first DT was developed based on impurity measures and splitting data recursively.

Following up, in 1966, the first publication about the application of DT was in Psychology, with the Hunt algorithm. Hunt was based on research about the modulation of the human mind learning. In 1972, the first DT for classification was developed with the knowledge of maths by Project THAID [38], [39].

In 1977, Breiman, Stone, Friedman and Oshen invented the CART algorithm. In 1984, they followed up their study and published the first article about CART.

John Ross Quinlan made a new development. He changed the structure of the trees. Until then, CART supported two answers for each question. Then, in 1986, he discovered how to have more than two answers to the same question. In the same year, Quinlan invented the Iterative Dichotomiser 3 (ID3). In 2008, Quinlan improved his previous work by creating the C4.5. This development solved some issues present in ID3, and today the algorithm is widely used in DM.

The DM algorithm is organized into four types of nodes: root node, decision node and terminal node. The root node contains all the data and is the first in the hierarchy. The decision node is internal and represents the question to the data. Thus, the node separates the data into two options. It is possible to have more than two choices in a decision node - multiple answers. The objective of the internal node is to make a decision. The terminal node or leaf is where a particular branch of the tree ends. The leaf aims to answer the questions from the decision node [38], [39]

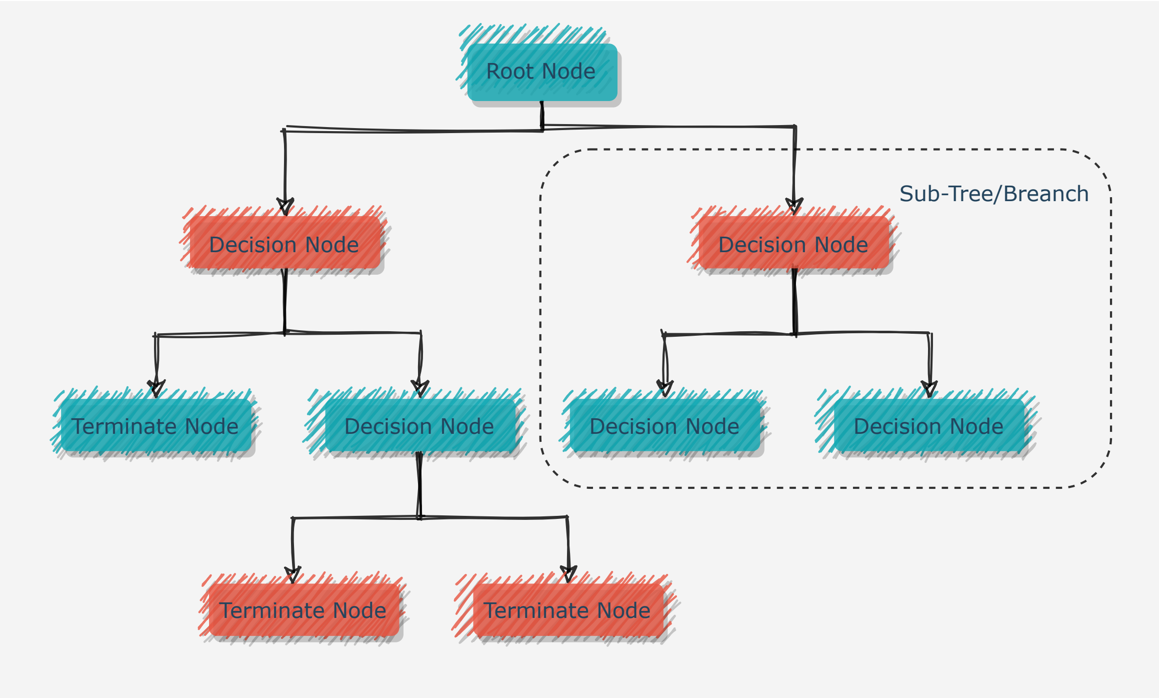


FIGURE 5. Definition of Decision Tree Architecture: Root Node, Decision Node and Terminate Node

One particularity of the algorithm is the flexibility to run both continuous and categorical features. Although Scikit-learn, Library of Python for ML and DL does not support categorical features, other libraries can.

One of the more common disadvantages is the possibility of doing an overfitting. This phenomenon happens when the algorithm training is excessive. The following equation explains the entropy:

$$I = -\sum_{i=1}^{k} |x_i - y_i| \tag{2.15}$$

The following equation explains the information:

$$Info(D) = \sum_{i=1}^{k} |x_i - y_i| \tag{2.16}$$

The Random Forest (RF) is a ML algorithm created in 2006 by Leo Breiman and Adele Cutler. RF can solve regression or classification problems [40][41][30].

This algorithm offers a solution for the possibility of overfitting in the DT. This algorithm takes as input the output of several decision trees. In this algorithm, the training data adjustment does not happen as in decision trees due to the high number of decision trees processed [30], [40], [41].

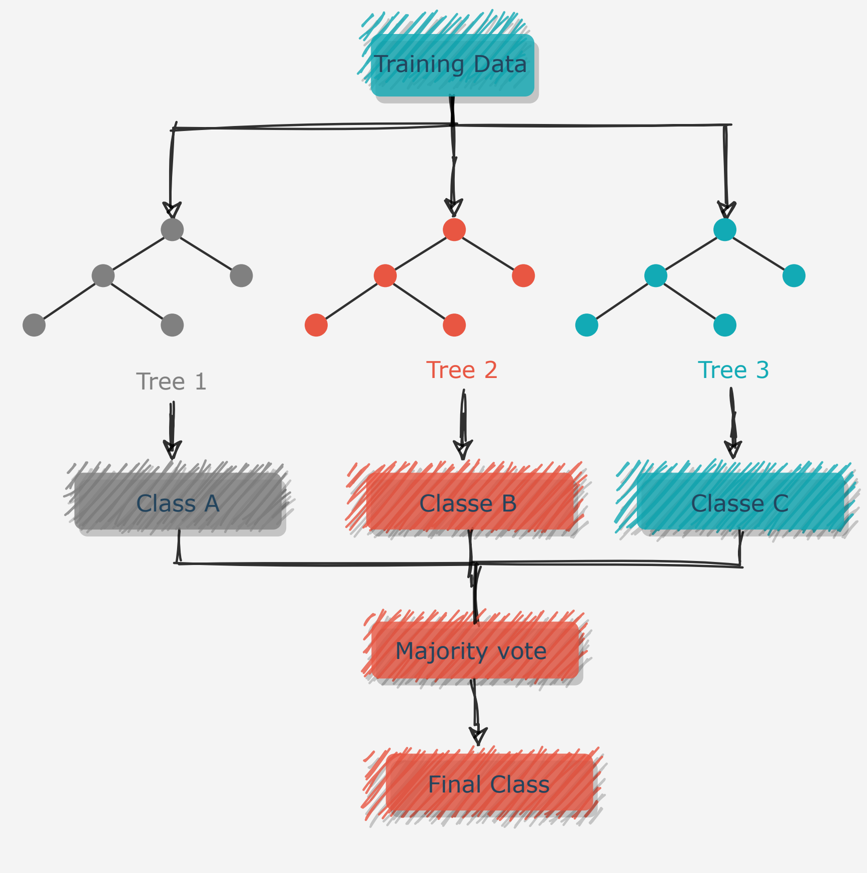


FIGURE 6. Definition of Random Forest Architecture with Input of Training Data and Output of Final Class

Support Vector Machine SVM is a ML algorithm to solve regression problems, classification and for outliers detection. In 1963 Vapnik and Lerner proposed for the first time that the optimal hyperplane separates the training elements with the most significant margin. SVM creates a line separating two classes. This separation is called a decision boundary or hyperplane. To generate the hyperplane, the SVM have support vectors that influence the position of the hyperplane [42], [43].

The algorithm understands the data as if it were a scatter plot. SVM aims to find the most optimized hyperplane possible.

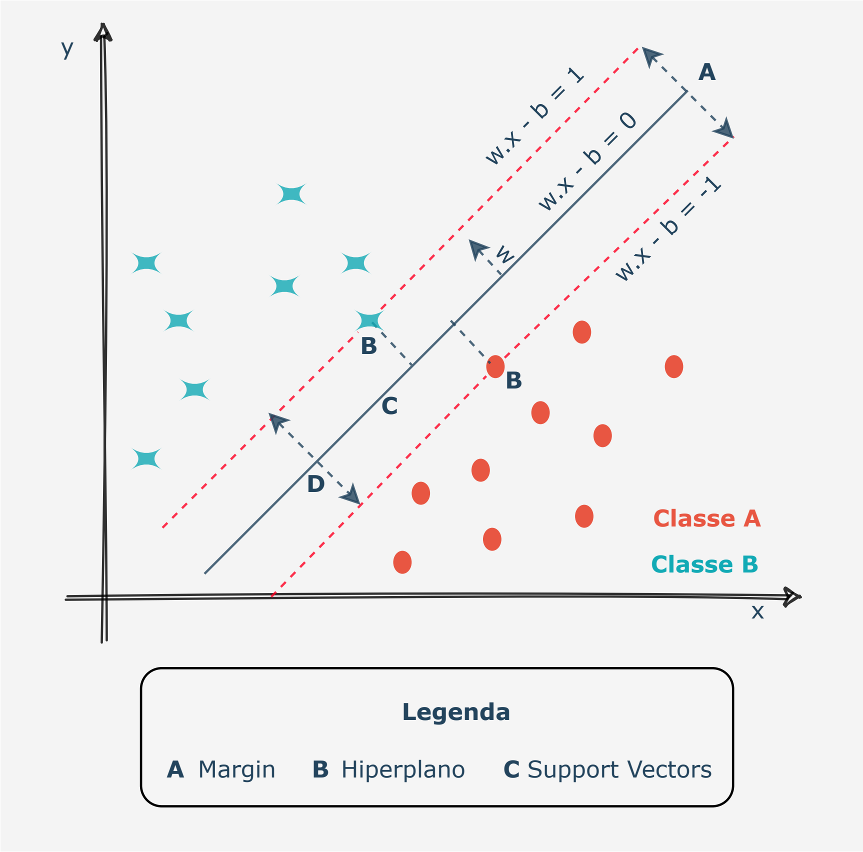


FIGURE 7. Definition of Support Vector Machines Algorithm: Hyperplane, Margin and Support Vectors

The support vectors are points close to the hyperplane. The hyperplane and its position depending on the support vectors. These are used for hyperplane optimization.

Linear kernel translates to a linear equation for predicting or classifying new data. This equation is the dot product between the input vectors, x_i, and the support vectors, x_j.

The C parameter aims to optimize the algorithm. When C has large values, the hyperplane will contain a small margin. The hyperplane will contain a large margin if C has a small value. The C and the hyperplane are inversely proportional.

The Gamma parameter influences the points that are or are not considered for a specific class. If the Gamma has a high value, the elements close to the hyperplane are considered for the class. Otherwise, the elements are not considered.

2.2.3. Deep Learning Algorithms

The Deep Learning (DL) is an AI branch, specifically a subset of ML, and uses many applications in the AI Area that provide nonlinear functions to model a dependency between inputs and features . The three areas, AI, ML and DL, are represented in Figure 1 [22], [25].

In general, the DL process contains a machine that learns alone such as Neural Network, with the imitation of the human brain represented in a computational network, with the representation of the neurons present in a human being. In 2015, LeCun, Bengio and Hinton created the following DL Equation 2.17 [22], [23], [26], [44]:

$$f(x,\theta) = W_{L\sigma L} + (W_{F-1}....\sigma_2(W_{2\sigma_2}(W_{1x})))|\theta = W_1, W_2, ..., W_L, with\theta = W_1, W_2, ..., W_n. \tag{2.17}$$

The DL has the same target as the ML, but has other specific needs and other characteristics. ML uses linear models and DL models with nonlinear functions: different types of learning.

DL is an Unsupervised Learning with less human action, it does not need to label the inputs but detects patterns in the data set. In contrast, ML is Supervised Learning that has contained the creation of labels to predict or categorize with human intervention. The DL data set can have a large quantity of data such as millions of samples. The computing power and the performance can be a problem, as a large data set needs a huge computation power to handle the data [27], [44], [45].

2.2.4. Deep Learning Algorithms

Neural Networks are algorithms that simulate the human mind behaviour. These networks try to imitate brain function. Like the human brain, a neural network contains neurons. A Neural Network comprises several layers with a minimum of three: input layer, hidden layer and output layer [26], [44]. The network has an input and an output layer. Inside, each has a specific number of neurons. This number depends on what the problem requests. The more complex the problem, the greater the need for more neurons. In our case, we want to predict the houses' price based on features. The features are the input for the first layer and the target - price property - is the output from the output layer [26][27][45][44]. The output layer can have more than one hidden layer. Neural networks are fully connected. In other words, the neurons between the layers are connected and the neurons in the same layer do not connect. The hidden layer has an activate function. This function has synapses as an input. Another critical concept when talking about the neural networks are the synapses. Synapses in neural networks have the same objective as synapses in the human brain. Their aim is the transmission of information between neurons [26][27][44].

Neural networks have many types of networks. We will explain the different types used to build the system to predict the houses' price: Artificial Neural Networks, Convolution Neural Networks, Radial Basis Neural Networks, Multi-layer Perceptron Regressor, Recurrent Neural Networks and Long-Term Memory Networks [26][27][45].

Many people think that Artificial Neural Networks are the same as Neural Networks. ANNs are a type of Neural Network. An ANN is a feed-forward network as it sends information layer by layer without touching any node more than once, so there is a continuous improvement, very similar to the human brain 2.17 [46]–[48].

This type of neural network, compared to the other types, is the simplest to implement and understand. The network is composed of only one neuron or perceptron. ANN comprises three layers: input, hidden and output. ANN has the advantage of being able to handle non-linear functions. The network learns the activation function but not the relationship between the inputs. The network deals with several problems, but it is more commonly used in text data, tables and images 2.17 [46]–[48].

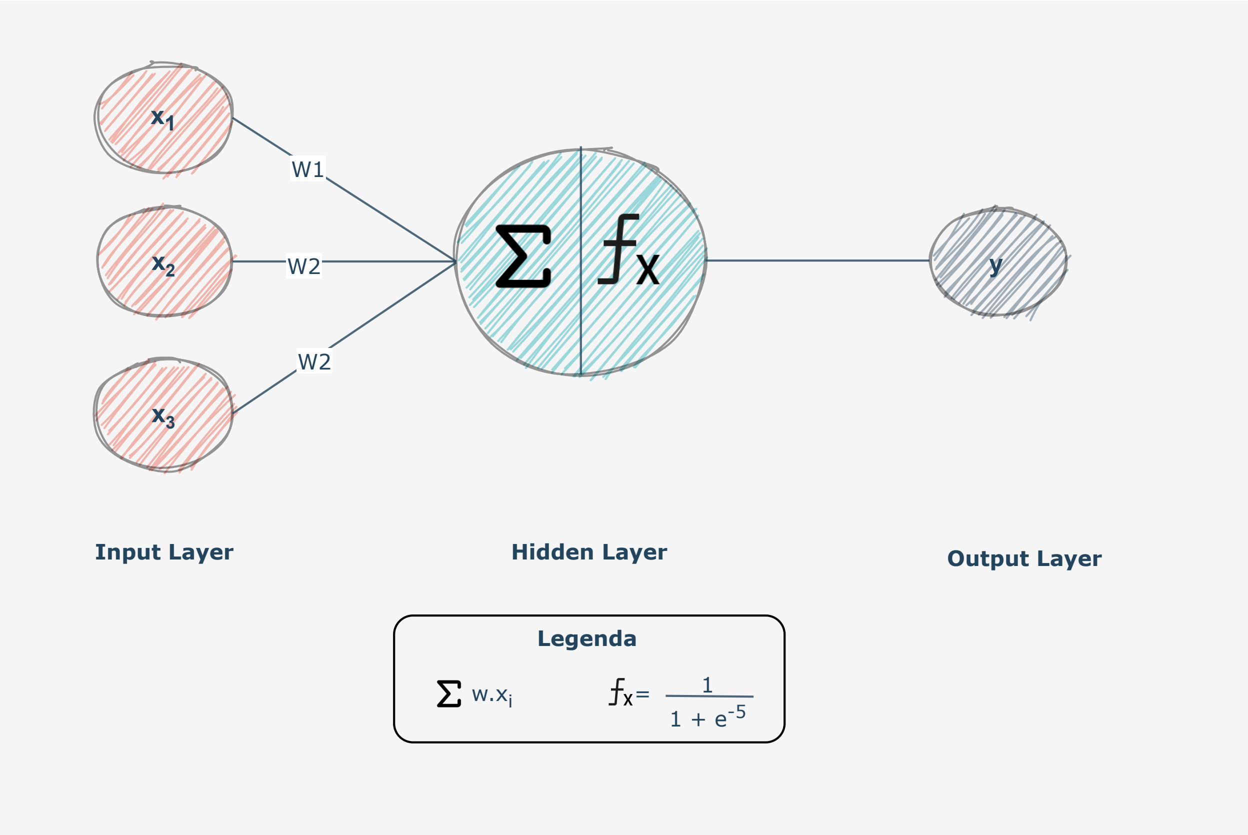


FIGURE 8. Definition of Artificial Neural Network Architecture: Input Layer, Hidden Layer and Output Layer

The following equation explains the entropy:

$$I = -w.x - b = 1 \tag{2.18}$$

Unlike ANN, which are multi-perceptron networks, Convolutional Neural Networks (CNN) are Multi-layer Perceptron Networks (MPL) and can have more than one hidden layer. CCN has the same logic as ANN: they are composed of neurons and their respective weights [49].

Each neuron receives input and calculates the scalar product between input and weight. The activate function is activated when the sum of the scalar products with a weight attains a certain value. In case the sum does not attain a specific value, the activate function is not activated. The weights are calculated again as well as the scalar product. This process is repeated until the activation of the function. The CNN has the input and output layers and another four layers: Convolution Layer, Pooling Layer, Fully Connected Layer and Non-Linearity Layer [49].

CNN is a network with a ConvNet architecture regarding neuron connectivity inspired by the Visual Cortex. Unlike ANN networks, CNN can capture spatial and temporal dependencies. CNN has the advantages of having a high accuracy value regarding image recognition, automatic detection of features not detected by human beings and weight sharing [49].The disadvantages are that networks loose data during the training phase and they cannot encode positive orientation. The following equation explains the entropy:

$$I = -w.x - b = 1 \tag{2.19}$$

The Recurrent Neural Network (RNN) is a Deep Learning algorithm. RNN uses recursion. The network saves the output value of a layer and makes it the input again[50].

RNN is composed of an input layer, one or more eye layers and an output layer. RNN has the advantage of temporal memory, which is relevant for time series forecasting. RNN has the disadvantages of implementation complexity and the inability to execute sequences with considerable sizes 2.17 [51], [52].

The Radial Basis Function (RBF) Neural Network is a Deep Learning algorithm for classification or regression problems. The network has a different architecture from other neural networks. While many of the neural networks are composed of several layers and introduce non-linearity in the activation function, RBF networks have only three layers: input, hidden and output 2.17 [50]–[53].

The input layer does not do any calculations. It just receives the input and sends it to the next layer - the hidden layer. This layer only exists to feed data to the hidden layer. So being, the number of inputs must equal the number of neurons.

The hidden layer takes the input data and transforms it into a linearly separable space (Cover's Theorem on the separability of patterns). The layer performs various vector comparisons.

Each hidden layer neuron has a prototype vector and a bandwidth denoted by μ, and δ, respectively [51][53] [52]. Each neuron calculates the similarity between the input vector and its vector.

The output layer uses the linear activate function. This layer works like the other neural networks. It calculates the linear combination between the input vector and the weight vector [51][52].

The Multi-layer Perceptron (MLP) is a Deep Learning algorithm for classification or regression problems. The network has a hidden multi-layer architecture. Each hidden layer can have the same activation function or have a different function than the other layers [53].

MPL has three layers: input, hidden and output. The network is two-dimensional. In the network, neurons in one layer are connected with neurons in the next layer [54] [55].

MLP has the advantage of being applied to complex nonlinear problems. MLP has the disadvantages of complexity and slow processing [54] [55].

The Long Short-Term Memory Networks (LSTMN) are a Deep Learning algorithm and a subtype of RNN. LSTMN came to solve the problem with the short-term memory of RNN. In the RNN, the current neuron has a short memory concerning the information passed by the last neuron [56][57].

In the case of LSTMN, it can retain information for a longer time. The short-term memory of the algorithm is stored in the cell state or hidden layer [56][57].

LSTMN is composed of four distinct gates: Forget Gate (f), Input Gate (i), Input Modulation Gate (g) and Output Gate (o) [56][57].

(f) is the parameter that determines whether to store the data in the memory or discard it. (i) is the element that defines the extent of information to be written in the cell state. (g) has the function of modeling the information given by the input to the cell state. This parameter adds non-linearity to the data and equates the average of the data to zero. Finally, (o) is the output generated from the current state of the cell state [56][57].

2.3. Data Extraction

This sub-section contains the concepts to be performed before model implementation in the Data Extraction and Data Preparation phase. To execute these procedures, we had to understand the importance of the attribute value. After that, how to extract attributes from a web page. The last section contains the concepts about Text Mining used to build the Portuguese data set.

2.3.1. Attribute Value

The Attribute Value (AV) can have several formats: discrete or real number, string, symbolic values or a set of previous formats. There must be an analysis of the attributes present in the data set to know whether or not they are relevant to the ultimate goal. If they are not, they have to be removed [58][59][60].

AV is divided into three categories: linear order, without order and partial order. The first is a set where the order of the elements matters, the second is where the order is irrelevant and the last one is where the partial order has implemented a specific order [58]–[60].

Typically, all types of attributes appear in the same data set, resulting in increased complexity. Another of the most common problems in Value Extraction is the missing values or the null values. For that, we need to adopt a strategy that implements techniques so these values are no longer null or have an associated value [58]–[60].

Being able to remove the lines in some cases is successful but, in others, we may lose relevant information. Another form is to choose a value that does not affect the result of the model and replaces all null values with the chosen one. It is also possible to create a forecast model to calculate the missing attributes. Another of these methods is addressed in later chapters.

There are several approaches and algorithms in TM. We analysed some of these methods as shown in the next subsection. All of the present explanation was referred to in the reading literature. The main objective of this section is to explore the possible approaches to use in the development of the Real Estate system, more specifically in the processing phases presented in the next chapters.

As discussed in Chapter 2, the increased growth of sharing all types of data and the easy access to them created a need to categorize data. An enormous quantity of data in the Internet is useless if it cannot be transformed into knowledge. Topic Modelling proposes to categorize and classify data with topics based on specific situations. The algorithm tries to find themes behind the unstructured information [17].

Named Entity is an object with a real meaning in the real world with the purpose to identify something such as a person, a country, an organization or another specific object. The identification and classification of entities is made with Named Entity Recognition (NER). NER is a technique of the NLP that can find entities in unstructured text [61], [62].

2.3.2. Automation Searching with Selenium

Before introducing Selenium, we have to define the concept of web scraping It is the process of accessing structured data on a web page. In our case, this method consisted in extracting HTML code and understanding its elements. The objective of this extraction was to collect possible attributes to build the data set in order to predict the houses' price in Portugal [12], [13].

At a lower level, the extraction was a simulation of the user behaviour in the Mais Consultores web page. An example of this can be a user that wants to buy a house and for this purpose the user has to execute several actions. An action could be clicking a button to download a file or going back to the previous page, etc. The extraction can help to collect knowledge in text formats, such as the pages' title or photos presented on a certain website. In this study, we just extracted text [1].

The process of searching and extracting attributes from Mais Consultores web page included several tasks. These tasks were previously programmed and automatic. This improvement allowed a less prolonged processing phase and also there was no need for human resources. Even so, as the site presents 2196 houses, our system takes 12h to be executed [1].

This study uses Selenium for automating two tasks: searching articles names or ISSN to extract Quatiles Ranking in the SJR; extracting features of the houses that Mais Consultores web page and Imovritual web page is promoting for acquisition to build the data set [1], [63].

2.3.3. Attribute Extraction

Attribute Extraction (AE) consists in extracting data with value. This value depends on the type of data set and the purpose of the analysis. A certain value of one attribute can have an important weight for a certain analysis and no weight at all for another. The relevant information extracted from the document can assume several forms, such as attributes, entities, etc. These heterogeneous features can pose a problem to the guarantee of success in structuring data [58][59][60]. AE includes three tasks: process of Selection of interesting attributes, Process of Selecting a subset of attributes, Attribute Categorization task. The first, contributes to the final result: the construction of the TM model or pre-processing the data.

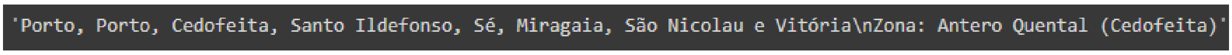
2.4. Text Mining (TM)

Regardless of the type of approach chosen, TM handles pre-processing, processing and text analysis for extraction of knowledge. There are several approaches and algorithms in TM. We analysed nine of these methods as shown in the next subsection. The main objective of this section is to explore the possible approaches to be used in the development of the Real Estate Platform, more specifically in the pre-processing and processing presented in the next chapters [17], [61], [62].

2.4.1. Text Mining Techniques

First, pre-processing is a set of tasks before processing. These tasks have the purpose of text cleaning so the text can be more easily processed at a later stage. We started with Tokenization.

Tokenization is a pre-processing task, in the sense that separates words, symbols, phrases or other elements called tokens. The aim of this method is to explore the tokens present in the sentences, depending on the type and scope of the text. The result of Tokenization. [64].

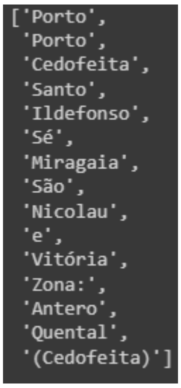


```
'Porto, Porto, Cedofeita, Santo Ildefonso, Sé, Miragaia, São Nicolau e Vitória\nZona: Antero Quental (Cedofeita)'
```

Figure 9. Element Definition without Application of Tokenization Procedure

Usually, the next step after Tokenization is the removal of Stop Words. This word type is frequently referred to in a data set, but with an irrelevant meaning, such as ’and’, ’the’, ’these’, etc. There are lists of Stop Words in several languages. The process of removing Stop Words in a specific data set also reduces the quantity of data and the complexity of the processing [64].

The second task is the Bag of Words (BOW). BOW consists in representing the text by a vector with a fixed size. Usually, BOW is a simple method, good for small data sets and used for text classification. This method gives the possibility to transform unstructured data into a vector [64].



```
['Porto',
 'Porto',
 'Cedofeita',
 'Santo',
 'Ildefonso',
 'Sé',
 'Miragaia',
 'São',
 'Nicolau',
 'e',
 'Vitória',
 'Zona:',
 'Antero',
 'Quental',
 '(Cedofeita)']
```

Figure 10. Element Definition after Application of Tokenization Procedure

This task can be extremely useful in the initial stages of building a TM model. The frequency matrix shows the data and their frequency. The frequency represents the number of times words appear in a certain sentence or document. There are several types of frequency to be taken into account. In short, BOW has three main steps: vocabulary definition, word frequency and frequency matrix.

2.4.2. Text Mining Library: Spacy

Spacy is an open-source library for NLP and not an Application Programming Interface API. Spacy does not provide software as a service or a web application. The library aims to include a large quantity of data in text format and process them [9], [14], [15].

Spacy helps extract, understand data text and pre-process text with DL. With this purpose, Spacy contains features and capabilities to process text: Tokenization, Part-of-speech Tagging, Dependency Parsing, Lemmatization, Sentence Boundary Detection, Named Entity Recognition , Entity Linking, Similarity, Text Classification, Rule-based Matching , Training, Serialization.

We will address the methods we use to build the Portuguese data set with data extracted from the Mais Consultores web page: Similarity and Rule-based Matching [1].

Spacy is a Python library for text analysis. This analysis is performed in several languages. For that, when we download the spacy object, we have to define the desired language for the text analysis. Although the library implements several languages, the English language is more versatile than the Portuguese language. English is a widely used language, so Spacy is better prepared than for text analysis in Portuguese [9].

By Spacy, Similarity() is defined as the vector comparison of words, i.e., it can be used to compare words, phrases, documents or tokens. The similarity value is between 0 and 1.0, with 0 meaning nothing similar and 1 completely similar [15].

SpaCy features a rule-matching engine, or Rule-based Matching, or just Matcher finds words in documents by applying rules. The programmer defines these rules and describes the characteristics of the attributes we want to find.

The search for specific attributes is performed by comparing them with a predefined pattern. Attributes are matched against the pattern and may or may not match the defined pattern or patterns. This thesis implemented the attributes of the following token: TEXT, LOWER, IS_DIGIT, and IS_SPACE. TEXT represents any word, while LOWER is only for lowercase words. IS_SPACE is a boolean token, or has the value of True or False, meaning a blank space exists. IS_DIGIT, like IS_SPACE is a boolean token and is intended to identify a number and not a word.

The tokens can be combined with regular expressions. For example, in our project, we used IS_DIGIT combined with a regular expression to extract the areas of the rooms of each house present in the Portuguese data set.

CHAPTER 3

Related Work

This chapter contains the previous work developed by several authors from 2013 until 2022 about the predicted price of houses Real Estate Sector. Section one retained the main objective to explain all the efforts for developing a SLR and its analysis.

Section two demonstrates the Article Analysis of the analyzed literature and compares article by article. This section explores several features such as publication year, country of origin, technologies implemented and algorithms executed.

Finally, in Section three, we analyze the results of the articles. This section presents the best approaches and solutions for predicting house prices.

3.1. Systematic Literature Review

This study chose the SLR for Review Protocol by its characteristics, being one of the methodologies that manage to gather relevant information and guaranteed quality.

This thesis performed a SLR, as proposed by Kitchenham in 2004, to identify the state of the art of predicting the price of houses. The chosen protocol review aims to elaborate a conceptual review of the definitions of existing tools for predicting house prices at the international level, identifying the techniques and algorithms most used in this practice and identifying possible gaps in the literature [65].

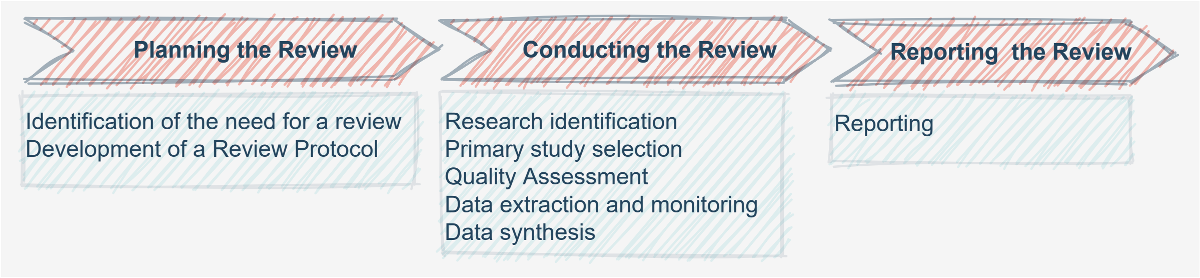


FIGURE 1. Scheme of the Systematic Literature Review SLR Process

As illustrated in Figure 1, the SLR has been divided into three phases: Planning the Review, Conducting the Review and Reporting the Review [65]. The first phase, Planing the Review, consisted of two sub-tasks: identifying the need for the review and developing a review protocol.The second phase, Conducting the Review, is composed of research identification, primary study selection, quality assessment, data extraction and monitoring and data synthesis [65]. The last phase, Reporting the Review, is not divided into tasks. This phase consists of reporting. The reporting phase is one of the most important and is addressed in this chapter in subsection three. It is important to communicate the results obtained [65].

3.2. Planning the Review

Planning the Review phase presents the two tasks in the Planning the Review phase: identifying the need for the review and developing a review protocol [65].

The objective was to plan how the process would evolve. First, this investigation defines the goals of the SLR, the restrictions and criteria and how to apply them, and how we will synthesize and report the studied literature.

The second task consisted of presenting the background, the research questions that the review was planned to answer, the strategies with the keyword definition, database definition, inclusion and exclusion criteria, and all recourse that we needed.

3.3. Conducting the Review

This subs section presents the five tasks in the Planning the Review phase: research identification, primary study selection, quality assessment, data extraction and monitoring and data synthesis [65].

A review protocol describes the procedures this thesis used to develop the SLR. The review protocol comprises the methods used in the review and provides a detailed plan for this thesis. The review protocol helped us conduct the research, answer the research questions and collect knowledge that was applied in developing the house price prediction system in Portugal. As a result of the Conduction the Review we development a SLR table as observed in table Figure 1.

Table 1. Definition of Systematic Literature Review: Keywords, Filters, Data Bases and Number of Articles

Data Base/Keyword	F1	F2	F3	F4
IEEE KW1	36	2	2	0
IEEE KW2	73	2	2	1
IEEE KW3	17	2	2	2
Web of Science KW1	641	259	131	4
Web of Science KW2	291	85	43	18
Web of Science KW3	151	84	80	3
Total	1209	434	260	28

The process included collecting theoretical and practical concepts to help build the system; searching keywords and filtering the articles [65]. In the end, we had articles with relevant information for the predicting system. This SLR had the articles with the best quality - Q1 in Journal Racking. This process was an automatic procedure.

The search was made with the Python script to collect the details of the article - authors, publish date, etc - and the value of the quartile. After collecting the Q1 articles, the script was written in the excel file. The quartite racking was extracted from the raking journal web page: Scimago Journal & Country Ranking (SJR) [63]. The last article selection consisted of reading all the abstracts and introductions and removing the unrelated articles. The following figure, Figure 2, presents the scheme of all the processes:

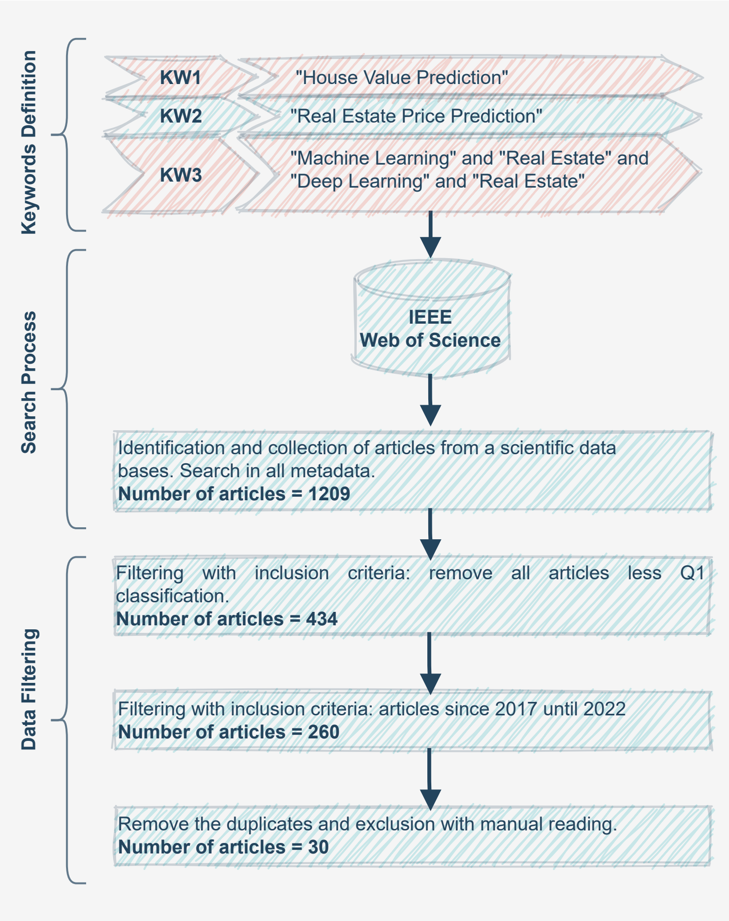

Keywords Definition
KW1
"House Value Prediction"
KW2
"Real Estate Price Prediction"
KW3
"Machine Learning" and "Real Estate" and "Deep Learning" and "Real Estate"
Search Process
IEEE
Web of Science
Identification and collection of articles from a scientific data bases. Search in all metadata.
Number of articles = 1209
Data Filtering
Filtering with inclusion criteria: remove all articles less Q1 classification.
Number of articles = 434
Filtering with inclusion criteria: articles since 2017 until 2022
Number of articles = 260
Remove the duplicates and exclusion with manual reading.
Number of articles = 30

Figure 2. Systematic Literature Review Process: Keywords Definition, Search Process and Data Filtering

The scientific databases are IEEExplore and Web of Science [66], [67]. The technique for filtering and choosing articles is based on the Quality Criteria. This classification is present in Scimago Journal & Country Rank (SJR) web page [63].

The keywords chosen were generalist, with the purpose of doing extensive research on the techniques and algorithms used, as well as a study of the tools already developed. The chosen keywords can be seen attached in the Figure 1 [65].

The research is carried out on the entire content of each article, then we classify the quality of the articles and only include the ones with the best classification - Q1. The last criterion used is the manual exclusion of articles not interesting for the study through reading [65].

The first filter, F1, was composed of the articles searched in all metadata and had a result of 1209 articles. As an inclusion criterion, this SLR filtered the Q1 classification in order to collect information with proven quality, giving a total of 434 articles. The third filter, F3, restricted articles by date: from 2013 to 2022. The last filter, F4, consisted of manual reading and selection of articles related to this thesis's theme. The filter description are presented in annexes in the Figure 2.

In the Planning the Review phase, it was predicted that we would only have the Q1 cracking restriction, but due to the high number of articles with the Q1 classification, we had to filter more as observed in Figure 3.

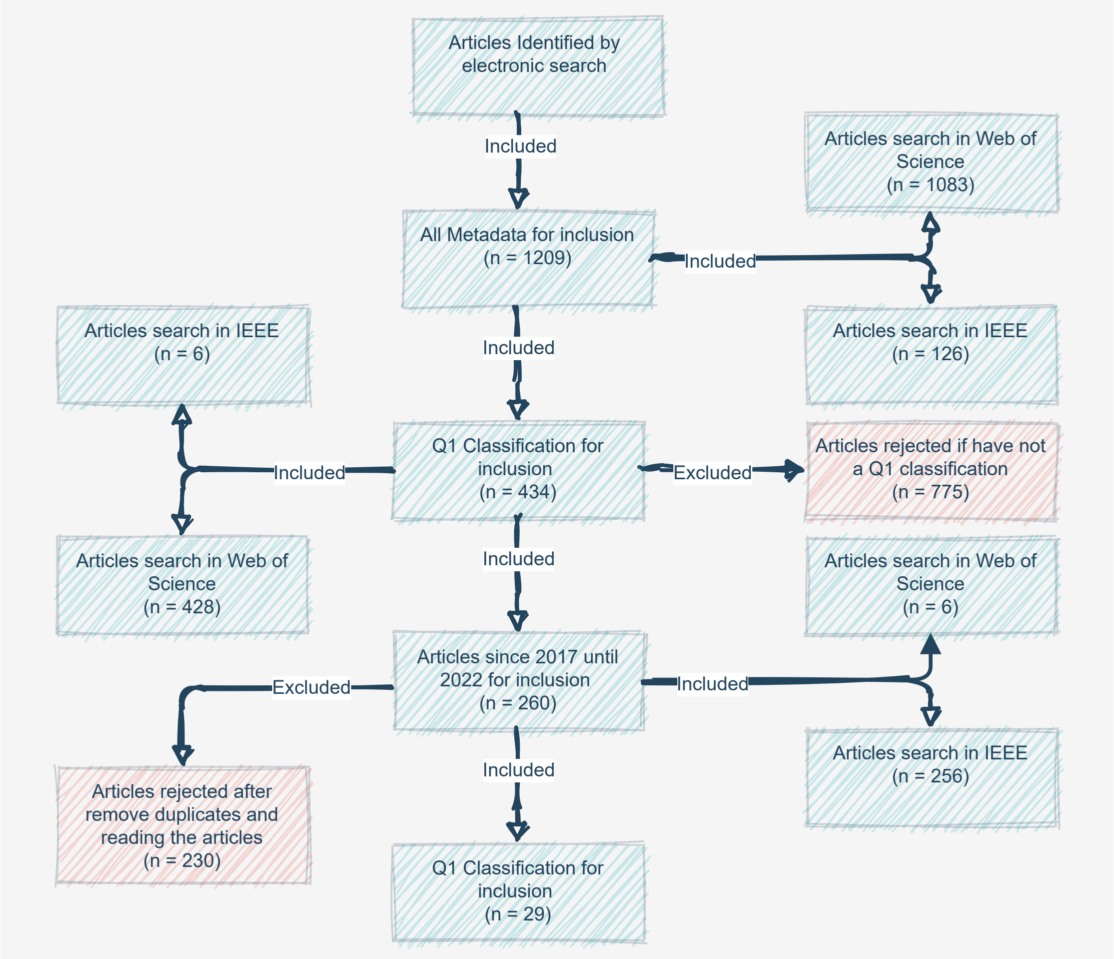

Articles Identified by electronic search
Included
All Metadata for inclusion (n = 1209)
Included
Articles search in Web of Science (n = 1083)
Articles search in IEEE (n = 126)
Included
Q1 Classification for inclusion (n = 434)
Included
Articles search in IEEE (n = 6)
Articles search in Web of Science (n = 428)
Excluded
Articles rejected if have not a Q1 classification (n = 775)
Included
Articles since 2017 until 2022 for inclusion (n = 260)
Excluded
Articles rejected after remove duplicates and reading the articles (n = 230)
Included
Articles search in Web of Science (n = 6)
Articles search in IEEE (n = 256)
Included
Q1 Classification for inclusion (n = 29)

Figure 3. Collection of Articles and the Application of Inclusion and Exclusion Criteria

This thesis has tried to have as many up-to-date articles as possible, but we have extended it to nine years due to the high number of articles. Even so, we had to read all the articles and select those interesting to our study. There is a considerable amount of information on home forecasting and different approaches. Our literature demonstrates this diversity.

The most selected articles are from the Web Science database. For example, the IEEE database had three articles; the rest was Web of Science. The first plan had more databases to reach as many authors and approaches as possible. Due to the massive amount of articles, we have reduced it to two databases.

The articles filtered by manual reading had the following selection criteria: being related to the thesis, preferring studies that demonstrate their knowledge at a practical level, if they are theoretical studies that bring some added value to the solution of this thesis and being written in Portuguese or English.

In the end, we identified that none of the analyzed studies used TM techniques for their solution. The significant similarity between the articles and this thesis was the use of ML and DL algorithms.

3.4. Reporting the Review

The reporting phase is one of the most important because this phase was where we analyzed the literature and extracted knowledge to develop the forecasting system. First, we started by analyzing the details of the articles: country, year of publication, a journal where it was published, title, keywords, technologies used in their approaches, and specific algorithms. Second, we further analyzed the implementations adopted by the literature authors extracted from SLR [65].

The results obtained in this phase focused on tools, techniques, and approaches to the problem of predicting house prices. It was one of the methods of this thesis to find if our solution had already been implemented. If some of the studies presented in the literature used text mining techniques to create a data set to be modeled for forecasting house prices in Portugal.

The list of 28 articles present in the literature is attached in Table 3 and 4. Each item has an ID to identify it. This ID will help identify each article in the following tables to be analyzed.

The article's information was used to transform data into knowledge. This method had the advantage of exploring chapter concepts and creating a general scenario for our study between 2013 and 2022 as observed in Table 5 in Annex and the Figura 4.

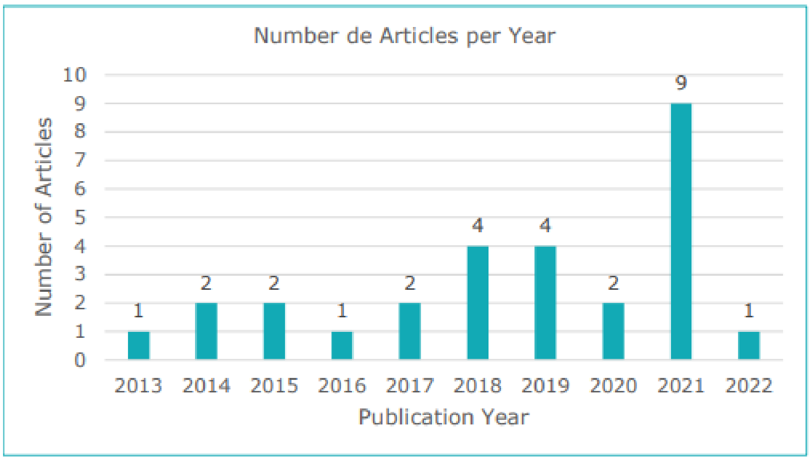


FIGURE 4. Literature Description: Number of Articles per Year

More than half of the studies were published between 2018 and 2022. The year with the highest number of articles was 2021, with nine articles published. The most recent study was published in 2022 by the USA journal Annals of Operations Research. In 2021, nine studies were published, and three articles have been published by USA in the following Journals: Expert System with Application, Journal of Property Analysis and PLoS One. Fourth of the studies were partnerships between USA and Brazil, USA and

Spain and USA and China, glsUSA, Bangladesh and Australia and with the respective journals Land Use Policy, Technological Forecasting and Social Change, Expert System with Application and Data Mining and Knowledge Discovery. Thailand, Poland and Taiwan published the rest of the studies published in 2021 in IEEE Access. For more details on the journals of each article, see attached Table 5.

United State Of America USA was the one that published more articles with a total of nine. After China with four articles published between 2018 and 2020 in the following journals: Habitat International, Technological Forecasting and Social Change, Socio-Economic Planning Sciences and Pattern Recognition. The countries analysis is observed in annex in Table 6 and in the Figure 5.

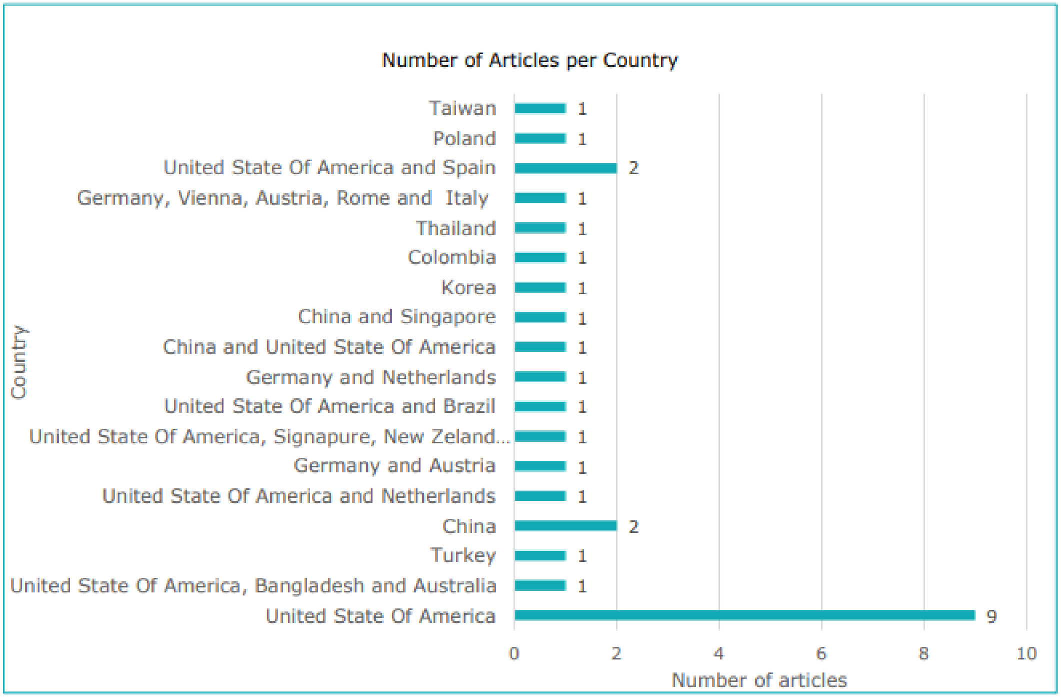


FIGURE 5. Literature Description: Number of Articles per Country

In Table 7 has the titles of the articles. The titles of the 28 articles present in the literature were presented. Eleven titles contain the words *house price* and fifteen contain the word emphpredict. Regarding technologies, two of the titles include the words DL, nine have ML and two of these contain the word *Neural Network*.

In Table 8 and Table 9 in annex has the keywords presented in the literature. The word *house* is mentioned in the articles' keywords six times, *predict* is mentioned eleven times, DL eleven times, ML twice and *Neural Networks* three times.

In terms of the technologies addressed, eleven of the studies used ML as a solution, nine of the studies used both ML and DL, and only two studies used DL. Of the remaining six studies, one used Neuro-fuzzy techniques, one performed an empirical analysis and 4 of these were theoretical studies, as illustrated in annex Table 10 and Table 6.

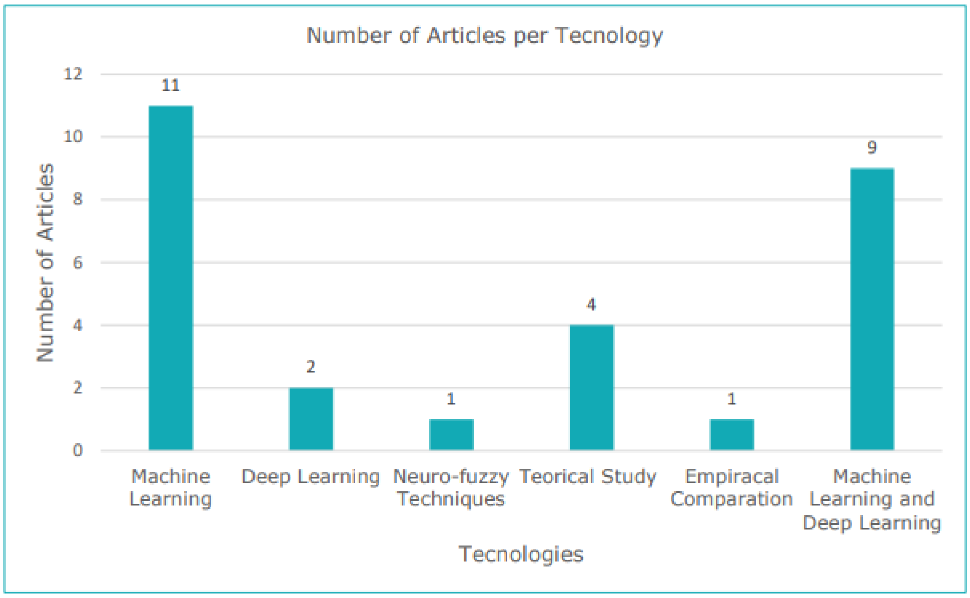


FIGURE 6. Literature Description: Number of Articles per Technology

In Table 11 in annex and Figure 7 has the algorithms used in the articles. Regression is one of the most common algorithms in forecasting cases. The case of forecasting house prices is no exception. The algorithm most used by researchers was a regression. Next, and also quite common, is the use of SVM in forecasting cases. The literature studies have four studies that implement SVM. Then we use neural networks to forecast house prices, specifically ANN. The implementation of ANN was carried out in three articles.

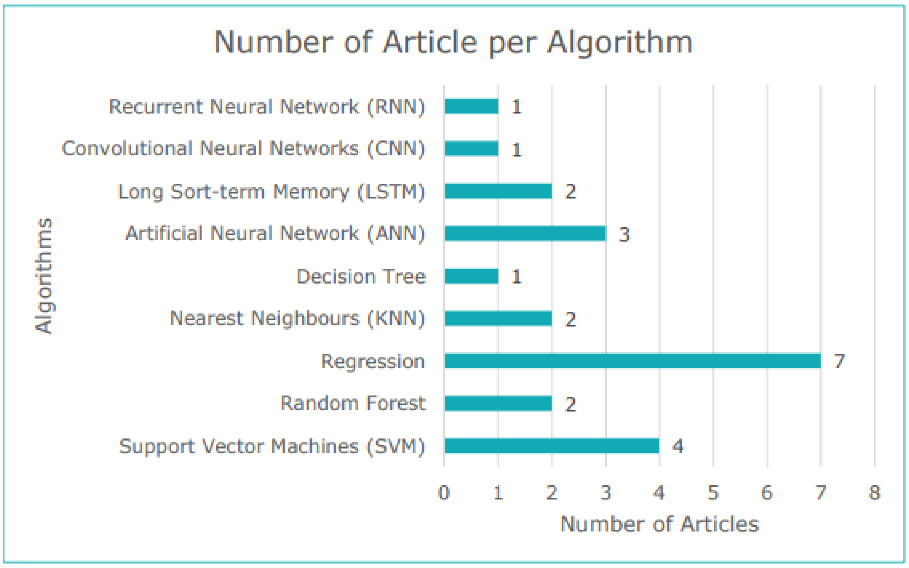


FIGURE 7. Literature Description: Number of Articles per Algorithm

In our houses' price prediction solution, we first extracted the data from a browser using web scraping procedures, Text Mining (TM) techniques to transform the data into knowledge and execution of ML and DL algorithms to predict the price of houses' price in Portugal. Finally, we analysed the results and compared them with results from other studies. This sub-section presents solutions from other authors concerning the theme of the thesis.

One of the studies published in United Kingdom, has the purposed to create an automatic model for predicting house prices in Coventry. The forecast is based on road distance and travel time. The data set of this study was created from routing, more specifically, Open Street Routing Machine (OSRM), an open-source engine for shortest paths in road networks for cars, bicycles and walking. This study used a statistical method called Geostatistical Kriging that uses non-Euclidean distance. The distance used is the approximation of road distance and travel time in a combination of linear and Minkowski distances. The predictor is validated with the Checkerboard method with spatial recognition in evaluating the results. Finally, the article compares the traditional results with the value of r^2 of 0.6901 with an error of approximately 0.18 with the result of the study with approximately r^2 0.66 and with an error of 0.21 [68].

Another study that did not follow the most common routes in the literature was [69]. This investigation implemented graphical neural networks to capture the Geospatial context of a house's neighbourhood. The authors implemented a Geospatial Network Embedding (GSNE), combined with regression techniques. The GSNE uses procedures based on the Gaussian curve. Its data set was composed by capturing the features through the geospatial neighbourhood.

The investigation of [69] proposed a system for predicting house prices implement several algorithms. The algorithms used were Lasso, Elastic Net, Kerner Ridge, Gradient Boosting, XGBoost, LGBM and averaging with KRR; GBoost and the XGB. MAE and RMSE were used as evaluation methods. We had a smaller MAE with Gradient Boosting with a value of 1.25 and a smaller RMSE with LGBM with a value of 0.180.

The study of [70] had the objective to predict of house prices houses' price in Tanzania, Uganda and Malawi, using ML techniques. The authors implemented Boosting, Bagging, Forest, Ridge, LASSO and Decision Tree. The data set they used contained data from houses in the previously mentioned countries between 2010 and 2016. As a method, they used MSE. They averaged the three countries and the algorithms with the best results were Ridge, LASSO, Bagging and Forest, with 83%.

Neuro-fuzzy techniques as a solution to house price houses' price prediction with historical data of houses for sale. More specifically, they used Adaptive Neuro-fuzzy (ANFIS), ANFIS with Grid Partition (ANFIS-GP) and ANFIS with sub-clustering (ANFIS-SC). The authors had better results with the ANFIS-SC model. As metrics, they used RMSE, MAE and r^2. The model with the best results, ANFIS-SC, obtained 2494 with RMSE 3.44 with MAE and r^2 with 0.01 [71].

One of the most curious studies present in the studied literature is about the investitor felling when buyed a house. It aims to predict future returns. In this investigation, the return is the value between buying and selling houses in Hong Kong. The authors wanted to prove that the recovery depends on the investors' sentiment. The authors had the objective of finding wanted to find an approach to the forecasting problem different from the traditional one. In the end, the result was that the sentiment is negatively related to house prices houses' price, i.e., when the investors' sentiment increases, the cost of the houses decreases [72].

An article from Austria proposed a house price houses' price prediction model in Vienna. This study implemented a new approach. The problem of predicting house prices houses' price has been addressed with models with spatial variation. The authors decided to compare several methods: Spatial Expansion Method (SEM), Moving Window Regression (MWR), Geographically Weighted Regression (GWR) and Genetic Algorithm-based Eigenvector Spatial Filtering (ESF). The authors validated the algorithms: the RSM with the best value of 0.288 and the LOOCV RSME with the value of 0.260 were used as metrics [73].

Another investigation proposed a prediction model through airborne laser scanning (ALS). They managed to decrease the error by 15% and increase the model's explanatory power by 13%. As metrics, they used RMSE, r^2 and AIC. They obtained the following results: RMSE with 0.254, r^2 0.611 and 26.933 with AIC. The authors wanted to improve the obtained results. To do so, they created an extended model to improve some parameters and obtained the following results: RMSE with 0.216, $r^2 = 0.690$, and 18.294 with AIC [74].

An investigation proposed three algorithms of ML, SVM, RF and Gradient Boosting Machine (GBM) for the price of the houses houses' price forecasting system. Their data set had 40.000 houses. The data is the history of 18 years in Hong Kong. The following metrics were used to evaluate the algorithms: MSE, RMSE, r^2, and Mean Absolute Percentage Error (MAPE). The algorithm with the best error values was the SVM with the following values: 0.01422 with MSE, 0.11925 with RMSE, 0.82715 with r^2, and 0.5467% with MAPE [75].

Another study developed an oligopoly model. This model is composed of differentiated products, for example, the suggestion that real estate agencies companies engage in price competition with their competitors. The authors support the model with empirical results [76].

The investigation of [77] proposed an investigation into the main features that influence house prices houses' price in the United States of America. Its data set comprised 13771 houses from the American Housing Survey (AHS) database. The data concerns the year 2013. In addition, the data set had 22 houses in San Francisco from the Redfin real estate brokerage database added. The authors used ML as a technology but specified a sub-set regression model. This study resulted in the essential features to increase, stabilize or decrease house prices. houses' price These features were the characteristics of the place where the house was located and its surroundings.

The authors of the article [78] created a method for building indices indexes of Real Estate prices in Singapura. Their study encompassed price information from single and repeated sales transactions. Its data set consists of data between comprises the years from 1995 to 2014 1995 and 2014. The study concluded that the best index for its data was based on standard hedonic methods. With this index, they could predict certain Real Estate Market behaviours, such as an exponential rise or decline in prices or a decline.

The study of [79] investigated whether the decision of a place to buy a house and market analysis in the Real Estate Sector. The authors' data set consisted of street view images recorded from 20.000 homes in Boston. It got 0.74 of r^2 as a result. The study concludes that geographic data can be combined with ML and attain good results.

Another study proposed a mapping of prices. The authors aimed to carry out an empirical comparison of the accuracy of univariate kriging variants, namely detrended kriging (DK) and universal kriging (UK), and multivariate extensions, including Detrended Cokriging (DCK) and Universal Cokriging (UCK). The study used the metrics of Cor1, Cor2, MSPE and RMNSE. As a result, the authors obtained the following values: 0.465, 0.047 and 2.579 in DK, 0.468, -0.046 and 2.579 in UK 0.465, 0.047 and 2.579 in DCK, and 0.465, 0.047 and 2.579 in UCK [80].

The study [81] combined a Spatial Neural Network (SNN) model called Property Appraisal 4.0. with a CNN. The authors created a neural network to automate real estate evaluation. This network can predict property or house values in an unknown neighbourhood from satellite images. The authors of [82] proposed solving the problem of predicting house prices houses' price using a modified Holt's exponential smoothing (MHES) and a model optimization algorithm - Whale Optimization Algorithm (WOA). The data set contained houses from four cities in China: Kunming, Changchun, Xuzhou and Handan. The authors concluded that the WOA-MHES method, i.e., MHES with optimization, has better results than MHES.

The authors of the article [83] used ML and DL to predict the prices. The study proposes a solution based on cost-sensitive deep forest and discrimination methods to better classify price classes. The authors concluded that the cost-sensitive Deep Forest reduces the errors margin of error and increases the accuracy.

The investigation of [84] aimed to use various ML algorithms to predict home houses' prices in Virginia. This study compared the results of executing four algorithms: C4.5, RIPPER, Naïve Bayes and AdaBoost. The investigation carried out by the authors resulted in a minimum error in all the algorithms they used. RIPPER had better results, i.e., it managed to be closer to the actual value of each prediction actual value of reality or of prediction.

One of the articles presented in the literature proposed a regression model that uses inference. This study has two datasets: 61.823 elements present from advertisements for home houses' sales in Colombia and 58.888 from the 2011 Metropolitan American Housing Survey data base. The authors developed a new method called Incremental Sample with

Resampling (MINREM) that has the selection of variables as an objective. This study aimed to compare a traditional system with the implementation of MINREM. They were able to get r^2 better value with the new approach [85].

In contrary to many of the studies' the investigation of [50] proposed a recommendations system. This system recommends houses. The authors built the recommendation system with ML and DL techniques. These implemented a RNN and have the lowest recall with the result of 3.60.

Another study based on ML and DL to predict the price of the houses houses' price has a particular aspect: it was supported by economics variables and index. The authors implemented the model, including unsupervised Deep Boltzmann Machine (DBM) Learning , SVM and Back-propagation Neural Network (BPNM). The authors concluded that they had obtained better results than if they had used the models separately [86].

The investigation of [87] compare house price houses' price prediction between ML and Herodic Regression, more specifically the following algorithms: DT, RF, KNN, CatBoost , XGBRegressor, MLP, Linear Regression, Linear Ridge and Linear Lasso. For the model, evaluation parameters used were MAE and RMSE. The authors obtained better results with XGBRegressor.

In Polish Real Estate, the authors of the study [88] chose Search Volume Index (SVI) as a tool for the prediction of houses' price. The investigation had as its purpose to predict if the houses' price would go up, down or stay the same. The authors had as their inspiration a Google tool that predicts the spread of the flu virus through SVI - Google Trends. The investigation combined data from Google Trends with Multiclass HPI Values Classifier. They concluded that the tool they created could be used to predict changes in the Real Estate Market.

Marianthi Stamou, Angelos Mimis and Antonis Rovolis proposed presented a study on predicting houses price prices in Greece, specifically in Athens. This study suggested a theoretical model with the particularity of spatial econometric analysis. The authors aimed to identify the factors that influence the price of houses houses' price[89].

The authors of the study [90] proposed a model for forecasting house prices houses' prices in France. The data contained government information on real estate transactions for a period of five years. This study used ML and DL techniques, combined with spatial attributes. The algorithms used were: MLP, RF, KNN, SVM, Adaboost, Gradient Boosting and Linear regression.

The investigation of [91] presents presented a theoretical model for predicting the price of homes houses' price in Norwegian. The authors proposed the idea that the distance from homes houses to services influences their prices. Services are places that might be of interest to a buyer. We have examples such as the existence of schools, supermarkets, hospitals and other possible external effects. The authors concluded that prices increased and reached their maximum value when there was a short distance from houses to services.

The study of [92] data set consists of real estate transaction data from 2017 to 2018 in Taipei and New Taipei. The data are satellite images from Google Maps. The authors used models from ML and DL: Extreme Gradient Boosting and Light Gradient Boosted Machine.

CHAPTER 4

Research Methodology

Before the project development, it was necessary to apply a process model to the data. Therefore, the present Chapter contains an overview of the Cross Industry Standard Process for Data Mining (CRISP-DM).

Section one includes background about the CRISP-DM and the scenario behind the methodology. Section two until Section seven presents the phases of the CRISP-DM and its description: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. These sections describe the methodology phases related to the building of the system to predict the price of houses in Portugal.

4.1. CRISP-DM Overview

CRISP-DM was created in 1996 by William Vorhies. The model began to have acceptance by experts. Moreover, in 1999, the project received financial support for the first time. Through the years, the model was applied in several DM projects. CRISP-DM has been around for approximately 20 years. During this time, it has contributed to scientific projects and organizations' business processes. Currently, CRISP-DM, is a process model besides being a methodology. Therefore, it is a guide to DM projects. CRISP-DM is adaptable to any area of business or technology [93]–[95]:

CRISP-DM was the research methodology chosen for this thesis. This choice was made due to the characteristics of the methodology. First, it is one of the methodologies specific for DM projects, i.e., it is prepared to analyze the house price prediction system in Portugal developed in this investigation. According to the characteristics of the life cycle discussed in the next sections, it has the benefit of following all the phases of the system and being able to help in its interpretation and development [93]–[95]:

Finally, CRISP-DM was an asset for understanding the business behind the project and the data related to it. Due to the complexity of our solution, CRISP-DM helped us understand the data extracted from Mais Consultores and give it meaning. CRISP-DM was an industry-proven way to conduct DM projects [1].

CRISP-DM as a methodology describes the project phases and the correlations between them. The CRISP-DM is a DM life cycle that contains six phases: Business Understanding (1), Data Understanding (2), Data Preparation (3), Modeling (4), Evaluation (5) and Deployment (6). Each phase has its purpose in a DM project [93]–[95]:

There is a sequence for executing the phases. The sequence is from (1) to (6), but we often need to go back to the previous stage: (1) to (2) and (2) to (1), (3) to (4) and (4) to

(3). There may also be a need for the repetition of the process before completing phase (6), going back from (5) to (1), as illustrated in Figure 1 [93]–[95]:

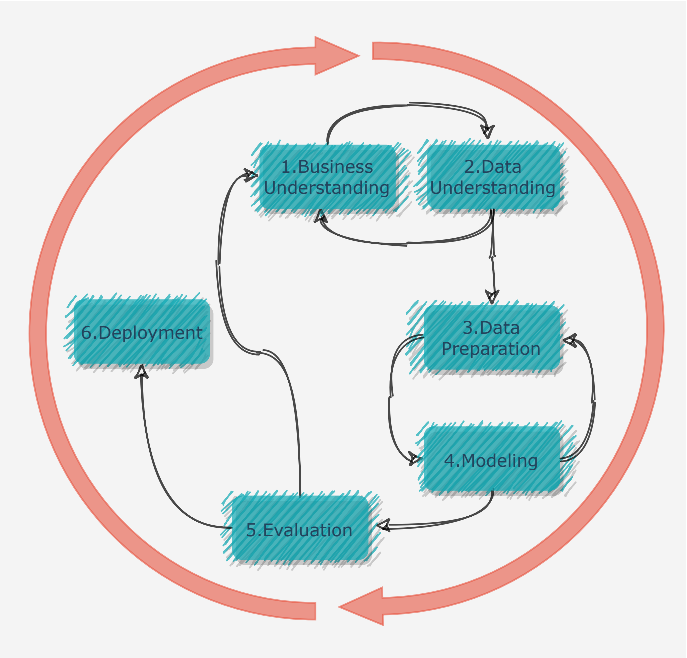


FIGURE 1. Definition of the Six Phases of the Data Mining Cycle

In addition to the advantages mentioned above, CRISP-DM is flexible. It can be applied in traditional projects - Waterfall - as well as in modern projects - Agile. Another advantage is that it can support any type of DM project and be independent of its technology. One disadvantage is the fact that is time-consuming due to the existence of documentation in each phase, which can be a laborious and slow process. Figure 2 presents the six phases of CRISP-DM and their explanation [93].

4.2. CRISP-DM Phases

Business Understanding is the first phase of the CRISP-DM Model Cycle. This phase sets goals and deadlines for the project. Business Understanding is divided into the following tasks: assessing the current situation, defining DM goals and developing a project plan. The objectives of this phase, besides knowing the business and its procedures, consist in realizing how it can be useful for the following phases, i.e., for the building of the system to predict the price of the houses [93]–[95].

In this thesis, the chapters *Background* and *Related Work* were useful for the development of the present phase and the forecast system. The first chapter is composed by an analysis of the technologies that support the system. The second chapter includes an analysis of the studies by several authors concerning the subject of building forecasting systems to predict house prices.

The main objective of this phase is to understand the business to make it easier to process the data in the next stage of CRISP-DM - Data Preparation. To do so, we had to understand the business of the Mais Consultores website and the Imovirtual site [1], [96].

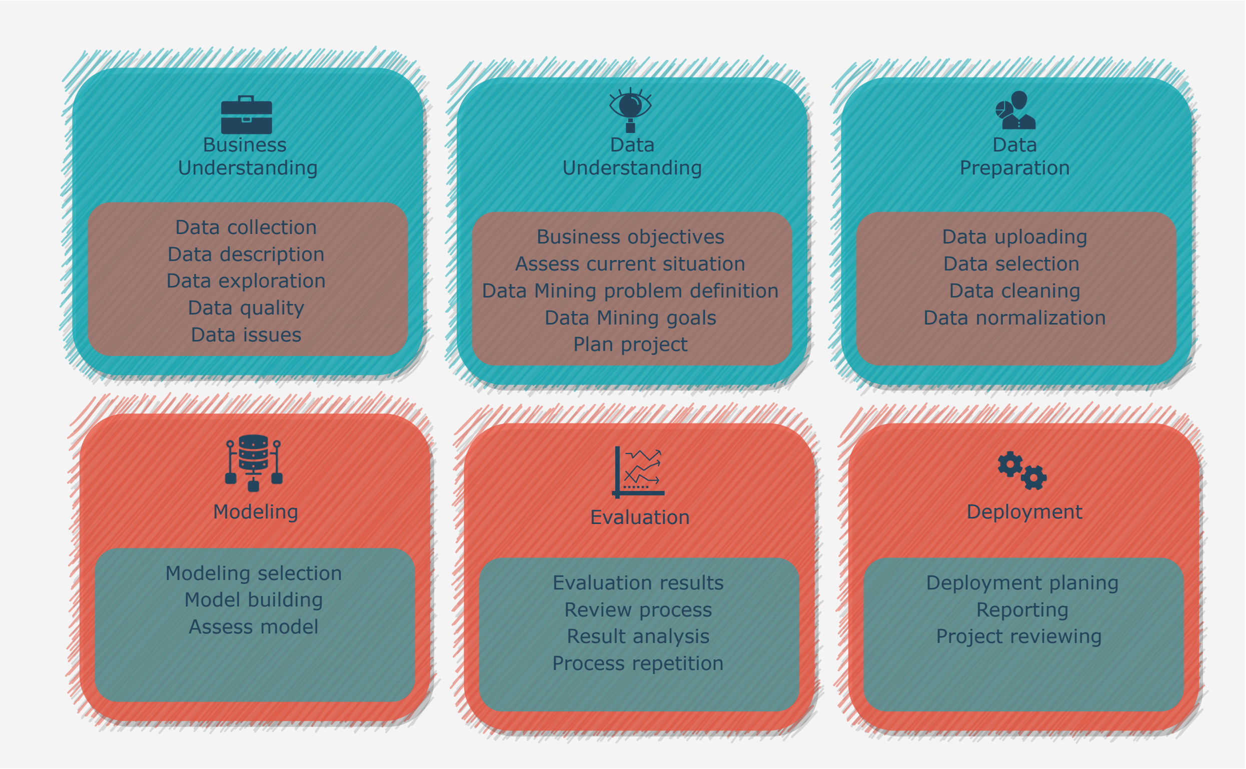


FIGURE 2. Cross Industry Standard Process for Data Mining CRISP-DM Life Cycle Phases

Mais Consultores is a network of real estate consultants. The site can be used by a consultant advertising a house for sale or by a client interested in buying a house.

The site consists of a search section for choosing the house's location. In our project, this field was left blank to search across the country and not in a specific location. Then we can see 18 ads of the house for sale per page. At the time of extraction, there were 122 pages.

Each of the advertisements can contain up to 14 attributes that characterize the house for our study: title, location, number of divisions, number of bedrooms, number of bathrooms, gross area, useful area, energy certificate, sun exposition, year of construction, price, a brief description of the house, the areas of each room, and relevant attributes.

Imovirtual is a Real Estate site in Portugal as Zillow in USA.

Data Preparation is a second phase of the CRISP-DM Model Cycle. This phase consists of understanding the data. It consists of four tasks: Data Collecting, Data Description, Data Exploration and Quality of Data [93]–[95].

Data Collecting consists of two distinct processes: we download the CSV files of Zillow from the Kaggle website regarding the 2016 competition. Then, we extract from the Mais Consultores website 2000 records of houses for sale in Portugal. The subsequent tasks were demonstrated and handled with Python in Google's Colab [1], [97].

The real estate market is highly complex to predict as it is in constant motion. For our system to be always updated date, the data set must be updated, i.e., a new extraction must be performed again. Concerning Zillow, the competition was carried out with data from 2016 and 2017. Unfortunately, Zillow has not made the data available again [7].

We extracted the data from an Excel file with a table composed of 15 columns and 2123 lines. Each line represents one house publicised on Mais Consultores web site. The columns represent features that describe the house and its surroundings. The columns include the title of the publication, the location composed by district, county, parish, and the neighbourhood characterisation; the price in euros, the number of rooms, the number of bedrooms, the number of bathrooms, gross area, useful area, year of construction, solar exposure, energy certificate, areas of the ground floor rooms, areas of the upper floor rooms, description [1].

Data Description consists of the characterization of the data in various aspects, such as the data type, the number of null numbers, the number of elements, etc. First, this subsection presents the Zillow data set and second, the Portuguese data set from Mais Consultores web page.

Zillow separated the data set into two CSV files. One dataset is composed of the features and the other of the target. These features characterize the house or the environment around it. While the target - logerror - is the value we want to predict. The logerror is expressed in the equation X. The target has an average value of 0.006, a minimum value of -4605, and a maximum value of 4737.

The 2016 Zillow dataset is made up of 58 columns or features. In all, the data set consists of 90275 houses. The aim is to predict logerror. So we went to analyze how the features relate to this one.

Figure 2 presents the six phases of CRISP-DM and their explanation. We will present the features with more relevance. The rest of them the images are attached in the annex.

There are columns without null values, these are: parcelid, bathroomcnt, bedroomcnt, fips, latitude, longitude, rawcensustractandblock, regionidcounty, roomcnt, propertylandusetypeid, logerror, transactiondate,assessmentyear.Null values are distributed differently across columns. These are shown in Table 12 and Table 13.

Data Exploration is the task of knowing the data. For example, if the data set has outliers, the correlation value between the variables, and the variance, among others [93]–[95].

The present subsection makes a brief presentation of the data included in the data sets used in this study: Zillow data set and the Portuguese data set. In Section One, Zillow Data is analysed and in Section Two the Portuguese data. The purpose of both sections is to characterize the data because the better we understand them, the better we can prepare them for the following phases: Modeling, Evaluation and Deployment.

Zillow Data Analysis was made from data available in the Zillow Contest site from 2016 and 2017. The competition took place 11 years ago. In this study we chose the data set from 2016. The objective of competition was to predict houses' prices from Zillow actual data. The contest attracted a lot of investigators and had good results concerning the final objective. That was the reason why we chose this example for our investigation.

We wanted to build a data set with Portuguese data with similar characteristics to the Zillow one and be able to have the same performance.

Before starting to create the Portuguese data set, we studied in detail the Zillow data set. First, we accessed Kaggle site for the download. Then, we uploaded it to a Google Colab Notebook. For access to the data download, open: `https://www.kaggle.com/competitions/zillow-prize-1/data`.

The Zillow data are divided in two CSV files: p2016.csv and t2016.csv. The first file contains IDs, otherwise named parcelid, and the remaining features. The second file is composed by parcelid, transactionDate and logerror. The transactionDate is the day of the data extraction. As already mentioned, the logerror is the target, expressed through the Equation 4.1:

$$logerror = \log_{Zestimate} - \log_{SalePrice} \tag{4.1}$$

Logerror is translated by the subtraction between $\log_{Z_{estimate}}$ and $\log_{SalePrice}$. $\log_{Zestimate}$ is the price estimated by Zillow AI sytems. These systems have approximately 7.5 millions of statitical models and ML with the houses' data. $\log_{SalePrice}$ is the actual price of the house. If $\log_{Zestimate}$ equals $\log_{SalePrice}$, then logerror is equal to 0. If $\log_{Zestimate}$ is bigger than $\log_{SalePrice}$, then the logerror value has to be positive. In case $\log_{Zestimate}$ is inferior to $\log_{SalePrice}$, then the logerror value is negative, as shown in Figure 3.

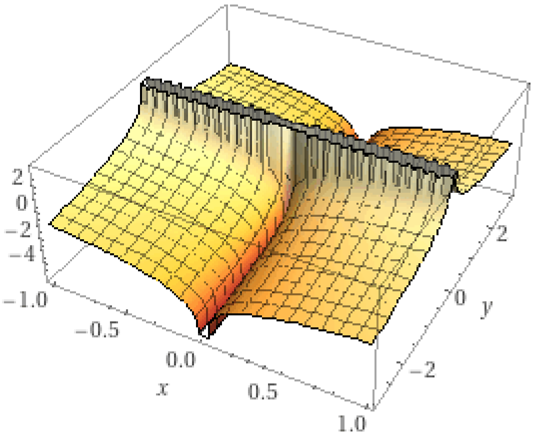


FIGURE 3. Logerror Variation - Analysis between the $\log_{Zestimate}$ (x) and $\log_{SalePrice}$ (y)

Logerror has a maximum value of 4.737, a minimum of -4.605 and an average of 0.006. Afterwards, we analysed the features included in the p2016.csv file. This file is composed by 58 columns. The description of each column is found in Zillow dictionary. See this information in the annexed tables: Table 12, Table 13. Having features characterized by ID, there is a dictionary for each one. These features are: HeatingOrSystemDesc, PropertyLandUseDesc, StoryDesc, AirConditioningDesc, ArchitecturalStyleDesc, TypeConstructionDesc, BuildingClassDesc. They are characterized in the annexed tables: Table 14, Table 15, Table 16, Table 17, Table 18, Table 19 and Table 20.

Data Quality is the task of validating the data and what procedures we have to do to clean it and increase the purity level [93]–[95].

Data Preparation is a third phase of the CRISP-DM Model Cycle. In DM projects, it is like calculating that the time spent on Data Preparation is 80% of the project, the other 20% for the rest of the project [93]–[95].

The importance given to this phase influences the success or failure of the project. How the data is presented to the algorithm. How the algorithm learns and classifies or predicts. It makes all the difference. That's why this was one of Portugal's most laborious phases of the house forecasting system. Regarding Zillow's data, it was not so expensive in terms of time, as the data was cleaner than the data extracted from the More consultants web page.

Data Preparation as a third phase of CRISP-DM is composed of five tasks: Data Selection, Data Cleaning, Data Construction, Data Integration and Data Format. Data Selection consists of separating data. The data will be included and excluded for the new data set.

Data Cleaning is composed of the necessary procedures until the data is as clean as possible to be performed by the algorithm. This cleaning consists of correcting, removing or adding anything that may affect the algorithm's learning and understanding of the data.

Data Construction is manipulating data to create new data from original data. Data Integration consists of creating a new data set with the data already processed. Data Format is a phrase that is always necessary because the data can be formatted. If necessary, data is formatted and normalized.

We used Colab Pro+ to execute the scripts in Python with a view to the data preparation. Some columns were not treated because the extracted data were already clean e ready to be used by the final algorithm: price, room, number of rooms and bathrooms.

From the column title, we extracted two columns: types and status. The column types represents the type of property, since the Mais Consultores web site publicises plots of land, detached houses, buildings, apartments, garages and every type of property that can be sold. For this thesis only the selling of houses if of interest.

The column status describes the type of condition of the house: remodelled, new, needing work, being remodelled, etc.

For each of these columns, we created a validation procedure. This procedure checks if the characterisation was correctly made or not through the comparison between the value of the column and the value of the column title. For this comparison, we used two techniques of TM: Similarity and Rule-Based Entity.

Let us analyse the process of the two columns. To be able to extract knowledge from the column title in order to create the columns types and status, we had to follow several procedures. The process of creating the columns is similar, so we will only explain the procedure concerning the column types.

First, we uploaded the file to a pandas dataframe. Afterwards, we converted the dataframe in a list with the only columns that interested us: parcelid and title. Then, we separated the title in tokens.

Next, we executed a function to estimate the similarity. We created several spacy objects to characterise each word that we were looking for: detached house apartment, restaurant, coffee shop, box room, building, plots of land, garage, home, duplex, farm, estate, office, bar, warehouse, cellar, business, house, property [9]. We tried several values of similarity in order to understand when there was a value inferior to null values. Even with these experiences, we created a validation algorithm for the column types and for the other columns created from the original data set extracted from Mais Consultores.

After that, we converted the result of the similarity function into a new dataframe with two columns: parcelid types and parcelid status. Finally, we created an Excel file with the mentioned information.

The data set with Portuguese properties was created with information from Mais Consultores [1]. This web page contains advertising for properties for sale or rent in Portugal. This investigation focuses on properties for sale in all the cities of Portugal.

Firstly, the process involves extracting features on the advertisement: title with information about the property, such as the topology (T0, T1, T2, etc.) or what kind of property, for example, a land or a house; location with the district, county, and neighborhood from Portugal; price of the property; topology; the number of rooms; the number of bathrooms; brute area; useful area; European Conformation (CE) code; build year; solar explosion; specific reference for the web page; small text with a general description of the property; attribute section with relevant characteristics about the comfort of the property; division section where have the area of each division of the house; and another information not interesting for this study.

The automation procedure for extracting data from the web page maisconsultores.pt has three main functions: connecting to the Chrome Driver, accessing the web page, and searching for all properties in all locations in Portugal (1); extracting individual features (2); extracting lists of features (3); organized the features in a table to an xls file and saved the file in a local directory (4).

Connect driver, allows the local machine to connect Chrome web Driver and access specific HTML elements on `maisconsultores.pt`. The data set uploaded contains Portuguese words. Raw text is required to be normalized and cleaned. Preprocessing work is simple in this thesis since the adoption of the rule-based. The disadvantage of the method implemented is the execution time for some preprocessing tasks.

The fourth phase of CRISP-DM is one of the main phases in a project of DM: Modeling. This phase contains all tasks necessary for executing one or more models. Modeling consists of four sub-tasks [93]–[95]:

- Modeling Selection;
- Design Testing; item Model Building;

Table 1. Definition of the Sets and its size: Training and Test

Set	Quantity	Percentage
Training	0.8	80%
Test	0.2	20%
random_state	42	

- Model Assessing.

The Modeling Selection task consists of selecting the models we want to use. This investigation proposes ML and DL models with the objective of knowing which is the best model for predicting the price of houses. With this in mind and based on articles in the studied literature and a survey *ad hoc*, we chose the models presented in Table 21 for the Zillow data set and in Table 22 for the Portuguese data. Both tables are attached. The tasks of this phase were applied to all data sets. The tasks are similar but not the same. We will present the differences as explainedas we proceed.

The Design Testing task is composed by the tasks before the training of the algorithms ou the algorithms' training: such as the division of data into features and targes plural ou singular, the division of data into training and test sets; any pre-processing that is still needed before running the models. In our case, we developed two dataframes. The first one has two columns: parcelid and price, which will be the target. The second dataframe contains all the features.

We chose not to have a validation set due to the reduced number of data in the Portuguese data set. We divided the data into two sets: training and testing. The training set has 80% of the data and the test set has the remaining 20

The Model Building task can be translated with the fit() function that is present in all ML and DL libraries. The fit() function consists of training the data, i.e., it is when the algorithm is learning. The algorithm gets to know the data, X_train and y_train, which respectively represent the training features and the training target.

The Model Assessing task is performed when we obtain the result of the test set predictions or evaluate the models results with gauges such as MAE. In our study, /dissertation/investigation/thesis we compared the results of several algorithms. We ran each algorithm twice with different parameters to compare results. Each algorithm was evaluated by the following metrics: MAE, MSE, RMSE, Coefficient of Determination and the largest error was calculated. The largest error translates into the greatest difference between the current value and the value predicted by the algorithm. The comparison of results is shown carried out in the next chapter, Chapter 6.

The fifth phase of CRISP-DM is another of the main phases following Modeling in a DM project: Evaluation. This phase contains all tasks necessary for understanding the results. The Evaluation consists of three sub-tasks [93]–[95]:

- Results Evaluation;
- Process Reviewing;

- Future Work.

Results Evaluation is the sub-task where we carry out the results analysis processor for business success. The business was studied in the early stages of CRISP-DM. In the present investigation, the results of the various models do not reflect the truth of the Portuguese Real Estate Market. We will analyse the results in the following chapters: Chapter 5 and Chapter 6.

Process Reviewing consists of reassessing the methods implemented in the models and changing them in case of improved results. In our investigation we implemented three samples for the Zillow data. The choice of three options was due to the desire to increase wish of increasing knowledge about the Zillow characteristics and to implement it in the Portuguese data. The Portuguese data were performe do twice with data from two different sites: Mais Consultores and Imovirtual.

Future Work is the process of reflection and definition of the next steps to take in the investigation. In our study, we discuss the possible parameters for the Portuguese results not being as good as the Zillow results. To improve results, we would choose on creating a data set with the volume of Zillow data with more homogeneity in terms of zones in Portugal. In addition, we would choose a Multithread system so that text analysis, data preparation, pre-processing and processing are less costly and faster tasks.

The last phase of CRISP-DM is Deployment. This phase contains all tasks necessary for the completion of the project. Deployment] consists of four sub-tasks [93]–[95]:

- Deployment Planning;
- Planning Monitoring and Maintenance;
- Production Final Report;
- Conduction Report Review.

The last phase of CRISP-DM is Deployment. This phase contains all tasks for the completion of the project. Deployment] consists of four sub-tasks [93]–[95].

- Deployment Planning;
- Planning Monitoring and Maintenance;
- Production Final Report;
- Driving Reset Review.

Deployment Planning, Planning Monitoring and Maintenance and Production Final Report are sub-tasks concerning the implementation of the projector in the current market. These sub-tasks have the purpose of producing a written document. The content of this document is analysed into Chapter 2 and Chapter 3.

Driving Reset Review is the general project reassessment sub-task. This consists of defining what went well, what can be improved and how can we do it. In order for the results of the Portuguese models to be better in the future, we need to study the Portuguese market in more depth and carry out a more homogeneous data extraction at the local o que - local or national level. If future work manages to cover the whole of Portugal and we have an amount of data identical to that of Zillow, the results will improve and the idea can be implemented for the general public and be a contribution to the scientific community.

CHAPTER 5

Results

The present chapter answers the research questions and presents the results in predicting house prices. In addition, this chapter addresses the three data sets explored throughout the investigation: Zillow, Mais Consultores, and Imovirtual. This chapter is divided into three parts, each with a respective research question.

The research question concerned a house price prediction problem and its scenario. These contained the techniques and algorithms of ML and DL for solving the predicting problem, RQ1. Concerning the predicting house prices application used in Real Estate Sector, RQ2. Lastly, a specific question related to this study: If a predicting model uses a data model and features similar to the Zillow data set, do the results continue to be positive, RQ3.

5.1. Techniques and Algorithms

The literature studied from the SLR of Chapter 3, *Related Work*, we could validate the technologies used by the studies. Most studies that aim to predict house prices use ML and DL as a solution. Research carried out *ad hoc* also demonstrates the same. Increasingly, AI is one of the first tools to be used by scientists to solve home price prediction problems. Remembering the first research question:

RQ1: Which techniques and algorithms are adequate for predicting the price of houses in the Real Estate Sector?

The first research question was answered in Table 10 and Table 11. Additionally, the development of the system for forecasting house prices in Portugal helped to execute . This investigation analyzed which techniques and algorithms were most used in the literature review. These results in the attached tables in Tables 10 and 11.

Undoubtedly, the most used techniques were those of ML. In second place are those of DL and in third place are the merging of the two. Good house price prediction results were presented in the articles presented in the literature, both in ML and in DL. The algorithms used in this investigation were used in the specific context.

For the implementation of the house price prediction system, it is recommended to use either ML or DL. Some algorithms had better results than others. For example, using CNN to predict house prices had worse outcomes than other ML and DL models. Specifically, from neural networks to predict house prices, CNN had worse results. This network is usually implemented for image classification and not for prediction.

Regarding ML algorithms, no study used RF as a solution. In the present investigation, we used RF and this was one of the ML algorithms with the worst prediction results. In addition to a longer training time than the other algorithms, we conclude that this algorithm is unsuitable for predicting house prices. Reinforcing the idea, the RF, both in the execution of the models with the Zillow data set and the Portuguese data sets, is one of the algorithms that have the worst results.

GridSearchCV was one of the algorithms presented in this study, which showed promising results and was not used by researchers in the literature studied. The results, as we will see in the following subsections, was one of the algorithms with the best results in Zillow data set.

5.2. Predicting Houses Price Applications

In order to build a forecasting system for house prices in Portugal, we had to analyze the tools already available. First, we tried to identify them in the literature and then this investigation did an *ad hoc* research to understand the characteristics of each of the solutions.The second research question is as follows:

RQ2: Are there applications/software for predicting the price of houses in the Real Estate Sector? If there are, which are they, where are they from and how do they work?

The articles in the literature contain investigations on the prediction of house prices at an international level. These investigations have not been applied to companies, nor does the solution have a public solution. The literature presents solutions on how to predict prices and not available tools.

There was mention of Zillow in two of the literature articles. These articles addressed the case of Zillow as a case of success and example. In addition to the mention of Zillow, the following price forecasting systems were mentioned in the articles: Redfin, Trulia, Realtor, Redfin and Lianjia.

5.3. Predicting Models

The last research question is practical. This study analyzed the Zillow data set to create a similar data set but with Portuguese data. This data set was constructed using TM Techniques and is explained in Chapter 4. The third research question is as follows:

RQ3: If a predicting model uses a data model and features similar to the Zillow data set, do the results continue to be positive?

After evaluating the results of the Portuguese data sets, the third research question is quickly answered. This investigation will analyze the results first and then answer the research question based on the results. In this investigation, we ran three experiments to test the Zillow data set:

- Removal of columns with less than 60% nulls. Filling null values with the mode of each column. Encoding with labelEncoder();
- Removal of columns with less than 60% nulls. Filling in null values with the mean of each column. Encoding with labelEncoder();
- Removal of all columns not in the Portuguese data sets. Filling in null values with the mean of each column. Encoding with labelEncoder();

The results presented for the first modeling are in Table 1, Table 2, Table 3 and Table 4. In addition to the algorithms presented in the previous chapters, a search method called GridSearchCV was used in the modeling phase. This method adjusts the hyper parameters and aims to find the best parameter combination. In this study, the method is used as a way for adjusting the parameters of the neural network, MPL.

TABLE 1. Results of the evaluation measures: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 1

Zillow Results - Sample 1						
ID	Algorithm and Methods	MAE	MSE	RMSE	r^2	Max error
1	KNN	0.0839	0.0211	0.1452	-0.4929	1.3373
2	KNN	0.0693	0.0693	0.1246	-0.0992	1.3412
3	DT	0.0673	0.0150	0.1226	-0.0636	1.3190
4	DT	0.0582	0.0141	0.1189	-0.0002	1.3190
5	RF	165.0794	31179.9615	176.5784	-2205453.3586	383.1578
6	RF	161.8868	30067.9974	173.4012	-2126800.7229	384.0356
7	SVM	0.0625	0.0157	0.1255	-0.1149	1.3612
8	SVM	169.1312	28623.1242	169.1836	2024600.4057	195.0387
9	CNN	0.0587	0.0141	0.1191	-0.0034	1.3240
10	CNN	0.9930	1.0002	1.0001	-69.7484	2.3240
11	MLP	0.0587	0.0141	0.1191	-0.0034	1.3240
12	MLP	0.0911	0.0229	0.1513	-0.6213	1.3636
13	LSTM	0.9930	1.0002	1.0001	-69.7484	2.3240
14	LSTM	0.0600	0.0242	0.1557	-0.7161	1.3240
15	GridSearchCV	0.0816	0.0192	0.1388	-0.3627	1.4567
16	GridSearchCV	0.0735	0.0170	0.1306	-0.2082	1.4177

As we can see in the tables mentioned, in general, the Zillow model has good results regarding errors and parameters being evaluated. The best results were from the algorithms of ML, more specifically DT, ID 4, and SVM, ID 7. As the worst results, we had the algorithms of RF with the ID 5. The parameters used to run the algorithm are in Table 23 and in Table 24 attached.

The algorithm with the best results, DT, used the following parameters: ccp_alpha=0.1, criterion=absolute_error, max_depth=200, max_features=25, max_leaf_nodes=None, min_samples_leaf=50, min_samples_split=50 and random_state=40. The algorithm DT ID 4 used the following parameters: ccp_alpha = 0.0, criterion='mae', max_depth=200,

Table 2. Results of the evaluation measures with minimum values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 1

ID 4	ID 4	ID 4	ID 5	ID 3 and ID 4
MAE	MSE	RMSE	r^2	Max error
0.0582	0.0141	0.1189	-2205453.359	1.319

max_features=10, max_leaf_nodes=None, min_samples_leaf=12, min_samples_split=2 and random_state=40. ID3 results were worse than ID4, despite having been executed the same number of times. The specification of a larger number of features to be analyzed and an alpha different from zero made the result much better.

Table 3. Results of the evaluation measures with maximum values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 1

ID 7	ID 5	ID 5	ID 8	ID 6
MAE	MSE	RMSE	r^2	Max error
169.1312	31179.9615	176.5784	2024600.406	384.0356

The worst results were presented by the algorithm RF, ID 5 and ID 6. Both have a high value of error and a small value of the coefficient of determination. The parameters used for ID 5 were: n_estimators=200, max_features=auto and random_state=42. For ID 6 the following parameters were used: n_estimators=500, max_features=25 and random_state=42. The best of the RF results was ID 6.

Table 4. Results of the evaluation measures with best values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 1

ID 4	ID 4	ID 4	ID 7	ID 3 and 4
MAE	MSE	RMSE	r^2	Max error
0.0582	0.0141	0.1189	-0.1149	1.319

The second option is represented in Table 5, Table 7,Table 6 and Table 8. The results of sample 1 and sample 2 with Zillow data only differed in filling in the nulls. In sample one, these were filled in with the mode of each column and in sample 2. They were filled in with the average values of each column. Sample 1 values manage to obtain error values lower than sample 1, but the difference is small. This sample continues with RF and SVM with great disappointment compared to the other algorithms.

Regarding the minimum values of the errors and the other measures analyzed, this study presents the results of the algorithms in Zillow Sample 2. The algorithms with the minimum values are DT with ID 4. LSTM with ID 14, the RF with ID 5 and the

Table 5. Results of the evaluation measures: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 2

Zillow Results - Sample 2						
ID	Algoritm and Methods	MAE	MSE	RMSE	r^2	Max error
1	KNN	0.0821	0.0200	0.1414	-0.4140	1.3302
2	KNN	0.0662	0.0154	0.1240	-0.0874	1.3876
3	DT	0.0658	0.0152	0.1232	-0.0729	1.3330
4	DT	0.0583	0.0141	0.1189	-0.0003	1.3290
5	RF	162.0646	30161.0930	173.6695	-2133385.6585	383.1578
6	RF	165.8226	31058.9575	176.2355	-2196894.3725	383.1578
7	SVM	0.0624	0.0156	0.1248	-0.1019	1.3484
8	SVM	168.9066	28551.2362	168.9711	-2019515.5490	195.0387
9	CNN	0.1870	0.1470	0.3834	-9.3979	1.3240
10	CNN	0.9930	1.0002	1.0001	-69.7485	2.3240
11	MLP	0.9930	1.0002	1.0001	-69.7485	2.3240
12	MLP	0.0934	0.0230	0.1517	-0.6271	1.4236
13	LSTM	0.9930	1.0002	1.0001	1.0001	2.3240
14	LSTM	0.0608	0.0162	0.0162	-0.1490	1.3240
15	GridSearchCV	0.0679	0.0151	0.1229	-0.0676	1.3071
16	GridSearchCV	0.0679	0.0151	0.1229	-0.0676	1.3071

GridSearchCV with IDs 15 and 16. The parameters of each algorithm are the same in all Zillow samples and are presented in the attached Table 23 and Table 24.

Table 6. Results of the evaluation measures with Minimum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 2

Minimum Values of Zillow Results - Sample 2				
ID 4	ID 4	ID 14	ID 5	ID 15 e 16
MAE	MSE	RMSE	r^2	Max error
0.05829284	0.014141596	0.016243722	-2196894.372	1.30712467

Regarding the maximum values, i.e., the worst results of the errors and the best result of the coefficient of determination. These values occur in the following algorithms: ID 8, ID 5, ID6 and ID 13. In short, the worst results are those of RF at a very high error level. On the other hand, the best results with the coefficient of determination are those of neural networks, specifically of LSTM.

The minimum values translate into the best values of the errors but the worst value of the coefficient of determination. For example, the algorithms with minor errors are DT and LSTM. Regarding the coefficient of determination, the algorithm with the worst value is RF.

Table 7. Results of the evaluation measures with Maximum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 2

Maximum Values of Zillow Results - Sample 2				
ID 8	ID 6	ID 6	ID 13	ID 5 e 6
MAE	MSE	RMSE	r^2	Max error
168.906596	31058.95755	176.2355173	1.000108957	383.1578

Table 8. Results of the evaluation measures with the best Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 2

Best Zillow Results - Sample 2				
ID 4	ID 4	ID 14	ID 13	ID 15 e 16
MAE	MSE	RMSE	r^2	Max error
0.05829284	0.014141596	0.016243722	1.000108957	1.30712467

The best results are shown in the 8 table. This table shows the best values for the algorithms of DT, LSTM and GridSearchCV. We conclude that in sample 2, where the null values were filled with the average of the columns, the DL algorithms started to generate better results than in sample 1. The only algorithm with good results from ML is the DT.

The third option with Zillow data is in the Table 9, Table 10,Table 11 and Table 12. The values between sample 3 and sample 2 vary, giving good results. It has some similarities in the worst results with RF and with SVM. And both samples show good results with DT. Let's take a closer look at sample 3.

Regarding the maximum error values, the worst results go to RF with ID 6. On the other hand, the best value of the coefficient of determination goes to CNN.

Regarding the maximum error values, the worst results go to RF with ID 6. On the other hand, the best value of the coefficient of determination goes to CNN.

In terms of the best values, as in the previous samples the algorithm with the best results is DT with ID 4.

In this section, therefore, this study concluded that RF was the algorithm with the worst results in all the experiments carried out with the Zillow data set. In terms of better results, we needed more than one algorithm. Several algorithms with the Zillow data set have had good results, both for ML and DL. Subsequently, this investigation will analyze the results of the data set with Portuguese data extracted from the Mais Consultores web page in Table 13 and the Imovirtual web page and in Table 17.

The modeling with the Portuguese data set used seven algorithms instead of eight. The models with the Portuguese data do not use one of the neural networks used in Zillow data processing: LSTM. The non-use of this approach was due to its characteristics. One

Table 9. Results of the evaluation measures: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Mais Consultores Zillow Sample 3

Zillow Results - Sample 3						
ID	Algoritm and Methods	MAE	MSE	RMSE	r^2	Max error
1	KNN	0.0840	0.0235	0.1532	-0.6591	1.2753
2	KNN	0.0705	0.0166	0.1287	-0.1716	1.3136
3	DT	0.0640	0.0149	0.1222	-0.0569	1.3130
4	DT	0.0583	0.0141	0.1189	-0.0003	1.3290
5	RF	170.8029	33389.2581	182.7273	-2361723.6791	383.1578
6	RF	171.4622	33810.9230	183.8775	-2391549.3307	383.1578
7	SVM	0.0611	0.0145	0.1204	-0.0251	1.3441
8	SVM	167.0646	27971.4264	167.2466	-1978503.8176	171.3240
9	CNN	0.0587	0.0142	0.1191	-0.0035	1.3240
10	CNN	0.9930	1.0002	1.0001	-69.7485	2.3240
11	MLP	0.9930	1.0002	1.0001	-69.7485	2.3240
12	MLP	0.0753	0.0177	0.1330	-0.2505	1.3550
13	LSTM	0.9930	1.0002	1.0001	-69.7485	2.3240
14	LSTM	0.0810	0.0217	0.1474	-0.5377	1.3240
15	GridSearchCV	0.0731	0.0162	0.1273	-0.1471	1.3956
16	GridSearchCV	0.0731	0.0162	0.1273	-0.1471	1.3956

Table 10. Results of the evaluation measures with Maximum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 3

Maximmum Values of Zillow Results - Sample 3				
ID 6	ID 6	ID 6	ID 8	ID 6
MAE	MSE	RMSE	r^2	Max error
171.4622	33810.9230	183.8775	-0.0003	383.1578

Table 11. Results of the evaluation measures with Minimum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 3

Minimum Values of Zillow Results - Sample 3				
ID 4	ID 4	ID 4	ID 6	ID 1
MAE	MSE	RMSE	r^2	Max error
0.0583	0.0141	0.1189	-2391549.3307	1.2753

of these features is that LSTM is suitable for time series and in Portuguese data there is no time reference.

Seven ML and DL algorithms were executed with the Portuguese data, both from the Mais Consultores site and the Imovirtual site. For both models, the following ML

Table 12. Results of the evaluation measures with Best Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Zillow Sample 3

Best Zillow Results - Sample 3				
ID 4	ID 4	ID 14	ID 13 and 14	ID 1
MAE	MSE	RMSE	r^2	Max error
0.0583	0.0141	0.1189	-0.1471	1.2753

Table 13. Results of the evaluation measures: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Mais Consultores

Mais Consultores Results						
ID	Algoritm and Methods	MAE	MSE	RMSE	r^2	Max error
1	KNN	315785.6	875240591947.4	935542.9	-1.8	9560000.0
2	KNN	288608.5	424555096797.0	651578.9	-0.4	4810000.0
3	DT	237874.7	431895548006.1	657187.6	-0.4	4750000.0
4	DT	209296.1	273760668258.4	523221.4	0.1	4460000.0
5	RF	394746.5	469604266606.5	685276.8	-0.5	4849712.0
6	RF	394745.3	469605697723.3	685277.8	-0.5	4849712.0
7	SVM	267583.1	328255548763.6	572935.9	0.0	4575000.1
8	SVM	394735.6	469664765807.4	685320.9	-0.5	4849709.0
9	CNN	395026.2	469894719924.4	685488.7	-0.5	4849999.0
10	CNN	395026.2	469894719924.4	685488.7	-0.5	4849999.0
11	MLP	395026.2	469894719924.4	469894719924.4	-0.5	4849999.0
12	MLP	286623.7	555599313322.4	745385.3	-0.8	10232248.2
13	GridSearchCV	370379.8	1234221790266.4	1110955.4	-2.9	20204692.7
14	GridSearchCV	370379.8	1234221790266.4	1110955.4	-2.9	20204692.7

algorithms were used: KNN, DT, RF and SVM. In terms of DL the following algorithms were used: CNN, MLP, LSTM and GridSearchCV. These algorithms can be seen attached in Table 22. In Table 14 is presented the maximum values of the measures parameters:

Table 14. Results of the evaluation measures with Maximum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Mais Consultores

Maximum Values of Mais Consultore Results				
ID 6	ID 6	ID 6	ID 8	ID 6
MAE	MSE	RMSE	r^2	Max error
395026.2406	1234221790266.4000	469894719924.3820	0.1277	20204692.7430

By observing the maximum values we are classifying the worst algorithms. In relation to the other samples from Zillow, the modeling of Mais Consultores has similarities. In both models, the algorithm with the highest values is RF. This one has a high MAE, a

MSE and a RSME compared to the other algorithms. The coefficient of determination despite having a high value for the RF is not significant. This model cannot predict house prices. The minimum values are in the Table 14.

Table 15. Results of the evaluation measures with Minimum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Mais Consultores

Minimum Values of Mais Consultore Results				
ID 4	ID 4	ID 4	ID 6	ID 1
MAE 209296.0672	MSE 273760668258.3790	RMSE 523221.4333	r^2 -2.9325	Max error 4460000.0000

The minimum values are the best error values and the worst values of the coefficient of determination. Again, the algorithm DT with ID 4 has the best values for MAE and for p MSE. The best constraint coefficient value is from the algorithm with ID 13. This algorithm is a neural network named GridSearchCV.

Table 16. Results of the evaluation measures with Best Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Mais Consultores

Best Mais Consultore Results				
ID 4	ID 4	ID 14	ID 13 and 14	ID 1
MAE 209296.0672	MSE 273760668258.3790	RMSE 523221.4333	r^2 -2.9325	Max error 4460000.0000

Subsequently, this investigation will analyze the results of the data set with Portuguese data extracted from the Imovirtual website. The results are presents in Table 17. The last modeling was the one that had the worst results in almost all algorithms. Let's analyze the results of the same.

In Table 18, Table 19 and Table 20 the maximum, minimum and best values found in the modeling of data extracted from the Imovirtual website are presented. These tables present the summary of Table 17. Having intuited a better reading and a quick understanding of which was the best and worst model for predicting house prices in Portugal.

Table 18. Results of the evaluation measures with Maximum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Imovirtual

Maximmum Values of Mais Consultore Results				
ID 6	ID 6	ID 6	ID 8	ID 6
MAE 661183.1033	MSE 9064229609577.6500	RMSE 3010685.9035	r^2 0.0203	Max error 91770454.0000

Table 17. Results of the evaluation measures: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Imovirtual

Imovirtual Results						
ID	Algoritm and Methods	MAE	MSE	RMSE	r^2	Max error
1	KNN	412746.87	9064229609577.65	3010685.90	-0.19	90730637.00
2	KNN	388961.21	7589853510745.15	2754968.88	0.01	90801786.20
3	DT	306879.69	7507060163567.23	2739901.49	0.02	91255455.00
4	DT	302042.06	7477146138670.04	2734437.08	0.02	91190455.00
5	RF	660472.85	8067595929594.00	2840351.37	-0.06	91769538.00
6	RF	302042.06	7477146138670.04	2734437.08	0.02	91190455.00
7	SVM	413479.48	7707607340726.85	2776257.79	-0.01	91384635.00
8	SVM	660670.73	8068037655062.52	2840429.13	-0.06	91769871.00
9	CNN	661183.10	8068890294548.89	2840579.22	-0.06	91770454.00
10	CNN	661183.10	8068890294548.89	2840579.22	-0.06	91770454.00
11	MLP	661183.10	8068890294548.89	2840579.22	-0.06	91770454.00
12	MLP	660670.73	8068037655062.52	2840429.13	-0.06	91769871.00
13	GridSearchCV	434676.31	7500830767507.53	2738764.46	0.02	90643247.21
14	GridSearchCV	405537.90	7452450963790.10	2729917.80	0.02	90551934.80

Let's start by analyzing the maximum values, that is, the results of the algorithms with the worst results. This is the algorithm with ID 6, the RF. Again this has a very high value of MAE, MSE and RMSE. The best value of the coefficient of determination is from the algorithm with ID 8, CNN. This value is also not significant as it exceeds 1 and does not translate into a good price forecast.

Table 19. Results of the evaluation measures with Minimum Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Imovirtual

Minimum Values of Mais Consultore Results				
ID 4	ID 4	ID 4	ID 6	ID 1
MAE	MSE	RMSE	r^2	Max error
302042.0624	7452450963790.1000	2729917.8000	-0.1877	90551934.8000

The minimum values are very high and are not representative of a good classification. The difference between the current price and the predicted price is very high, so high that the model does not predict like the Zillow model, and is more wrong than right. Even with high error values and non-coherent values of the coefficient of determination, the algorithm with the lowest results was DT, with ID 4. the best values are presented in the Table 20:

Answering the research questions was of added value to the success of this study - first, an understanding of the scenario behind the theme of forecasting house prices. This investigation gave theoretical concepts for constructing a forecasting system for house

Table 20. Results of the evaluation measures with Best Values: Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Coefficient of determination and Max Error - Imovirtual

Best Imovirtual Results				
ID 4	ID 4	ID 14	ID 13 and 14	ID 1
MAE	MSE	RMSE	r^2	Max error
0.0000	302042.0624	7452450963790.1000	2738764.4600	-0.1877

prices in Portugal and the best and worst results with different approaches. Finally, through the development of the system for Portugal, we managed to understand that the data sometimes behave in different ways due to the many similarities they have to successful cases. This summary is observed in Table 21.

Table 21. Answer to the Three Research Questions Defined for the Present Study

RQ	Key Findings
1	- Frequent use of ML and DL algorithms - Less frequent use of theoretical studies - Most used algorithms in ML - Algorithms most used in DL were neural networks, more specifically ANN and the combination of neural networks with Geo-spatial systems. - Most of the algorithms implemented their solutions with accurate data from advertisements of houses for sale - text format - Few studies used satellite images or street view as data
2	- All solutions in the studies made a scientific contribution when it comes to predicting the price of homes that are for sale - No study implemented its solution and made it available to the general public or any company - The tools mentioned and analyzed in the studies were Zillow, Redfin, Trulia, Realtor, Redfin and Lianjia
3	- This study proves that the use of similar features and a structure identical to the Zillow model is not guaranteed - For many similarities that the data set has with the Zillow data set, this one can present results as good as this one or not - Independence of features and data composition for successful forecasting

CHAPTER 6

Conclusion

The present chapter concerns the conclusions of this investigation, the limitations and the reflections on future work. This research aimed to estimate the price of houses in Portugal with Machine Learning and Deep Learning techniques. The purpose of this investigation was divided into data extraction, preparation, and comparison.

Data extraction consisted in the extraction of data from the Mais Consultants and Imovirtual web pages. Mais Consultores extraction was developed in Python with a Selenium library. Although the extraction was successfully done, we consider that Selenium is not the best choice for data extraction on sites that advertise houses for sale. Selenium made the extraction time-consuming. The extraction made by Selenium extremely time-consuming. In addition, the page updates were challenging to synchronize the extractions and doubled the execution time.

Imovirtual extraction was developed in Python with the BeautifulSoup library. BeautifulSoup, unlike Selenium, has no timing-consuming problems and no page updates. Instead, Selenium tries to simulate a user using the page and BeautifulSoup does the actions. The program that used the BeautifulSoup was developed with a combination of BeautifulSoup and Multithreading. The extraction with BeautifulSoup was successful and took approximately less than half the time than Selenium and with twice more data.

Zillow results were as expected and the Portuguese data set had a lower error and a higher accuracy. However, the results obtained with the Portuguese data were not expected nor similar to Zillow. Moreover, even with similar features and identical correlation weights, the algorithms performed worse results.

The three data sets with Portuguese data had results. However, the difference between actual and predicted prices is significant. Thus, the techniques used for data set construction with Portuguese data identical to Zillow cannot get results as good as this one. Zillow results are similar to those obtained by some of the researchers who participated in its competition. The results of the Portuguese set date/data set cannot be compared to Zillow's.

As a contribution to the scientific community, Text Mining techniques and Text Analysis techniques are not deterministic factors for predicting the price of houses with Portuguese data. Adding that to the use of language with similar data alters the results and should be analysed as distinct cases. These techniques no effective results with the application of Machine Learning and Deep Learning algorithms. On the other hand, the extraction of data from a browser with advertising homes for sale is carried out faster and with better results with the Beautiful Soup library and not with the Selenium library. Since the combination with Multithread the BeautifulSoup is a better solution than Selenium.

Although the results of the models with Portuguese data do not match those of the Zillow data, both had similar results. We thus state that to predict the price of houses, both in Portugal and in the USA, the best technology is ML and the best algorithm is DT. These results may not hold true for other data types. The data have an influence on the presentation of good results as well as the language in which they are written. Each language will require the use of different pre-processing techniques.

This investigation had, as a limitation, insufficient local resources. The machine used for the processing had limited resources, as do Google Colab Pro+ subscriptions. This investigation used text analysis and similarity and rule-based matching functions. These functions making various iterations and comparisons, the processing being time-consuming. The more data a data set contains, the more processing time and resources it will take. In this study, thesis we were able to obtain more data than those presented, but we did not have the processing resources to achieve pre-processing and data processing of the whole data.

This study conclude that, although the quantitative results presented do not allow us to infer any relationship between improving the price of houses forecast in Portugal and using Text Mining to prepare the data, this study proposes one of the techniques. For future work, there must be a back end mechanism for processing the data. This should be previously studied, so that in the modeling phase it will be possible to carry out the experiments more quickly and without requiring as many resources as in the current study. One of the options considered is the use of an external cloud that can run a notebook with more processing and memory capacity. After discussing the results, we present the exploration of two future. The first will be data extraction from multiple sites and, if possible, creating some partnerships to obtain data. The second will be using Text Mining techniques with the help of Multithread. These two proposed paths have a better aim at the quantity and quality of the data and improving the time of implementation of the data preparation.

Appendix

Table 1. Definition of the Keywords Present in the Systematic Literature Review

Keyword	Definition
KW1	House Value Prediction
KW2	Real Estate Price Prediction
KW3	Predicting House Prices and Machine Learning or Predicting Real Estate House Prices and Machine Learning and Predicting House Prices and Deep Learning or Real Estate House Price and Deep Learning

Table 2. Definition of the Filters Present in the Systematic Literature Review

Filter	Definition
F1	Search in all metadata
F2	Choose Q1 Racking
F3	Search between 2013 and 2022
F4	Choose the articles with manual reading

Table 3. Articles Present in the Systematic Literature Review - from ID 1 to ID 10

ID	Article
1	Crosby, H., Damoulas, T., Caton, A., Davis, P., Porto de Albuquerque, J., & Jarvis, S. A. (2018). Road distance and travel time for an improved house price Kriging predictor. Geo-Spatial Information Science, 21(3), 185–194. https://doi.org/10.1080/10095020.2018.1503775
2	Das, S. S. S., Ali, M. E., Li, Y. F., Kang, Y. Bin, & Sellis, T. (2021). Boosting house price predictions using geo-spatial network embedding. Data Mining and Knowledge Discovery, 35(6), 2221–2250. https://doi.org/10.1007/s10618-021-00789-x
3	Embaye, W. T., Zereyesus, Y. A., & Chen, B. (2021). Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. PLoS ONE, 16(2 February), 1–20. https://doi.org/10.1371/journal.pone.0244953
4	Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. Automation in Construction, 41, 33–39. https://doi.org/10.1016/j.autcon.2014.02.002
5	Hei-Ling Lam, C., & Chi-Man Hui, E. (2018). How does investor sentiment predict the future real estate returns of residential property in Hong Kong? Habitat International, 75(July 2017), 1–11. https://doi.org/10.1016/j.habitatint.2018.02.009
6	Helbich, M., & Griffith, D. A. (2016). Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches. Computers, Environment and Urban Systems, 57, 1–11. https://doi.org/10.1016/j.compenvurbsys.2015.12.002
7	Helbich, M., Jochem, A., Mücke, W., & Höfle, B. (2013). Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning. Computers, Environment and Urban Systems, 39, 81–92. https://doi.org/10.1016/j.compenvurbsys.2013.01.001
8	Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558
9	Iwata, S., Sumita, K., & Fujisawa, M. (2019). Price competition in the spatial real estate market: allies or rivals? Spatial Economic Analysis, 14(2), 174–195. https://doi.org/10.1080/17421772.2019.1532596
10	Jafari, A., & Akhavian, R. (2019). Driving forces for the US residential housing price: a predictive analysis. Built Environment Project and Asset Management, 9(4), 515–529. https://doi.org/10.1108/BEPAM-07-2018-0100

Table 4. Articles Present in the Systematic Literature Review - from ID 11 to ID 20

ID	Article
11	Jiang, L., Phillips, P. C. B., & Yu, J. (2015). New methodology for constructing real estate price indices applied to the Singapore residential market. Journal of Banking and Finance, 61, S121–S131. https://doi.org/10.1016/j.jbankfin.2015.08.026
12	Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. Land Use Policy, 111(September 2019), 104919. https://doi.org/10.1016/j.landusepol.2020.104919
13	Kuntz, M., & Helbich, M. (2014). Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. International Journal of Geographical Information Science, 28(9), 1904–1921. https://doi.org/10.1080/13658816.2014.906041
14	Lin, R. F. Y., Ou, C., Tseng, K. K., Bowen, D., Yung, K. L., & Ip, W. H. (2021). The Spatial neural network model with disruptive technology for property appraisal in real estate industry. Technological Forecasting and Social Change, 173(August). https://doi.org/10.1016/j.techfore.2021.121067
15	Liu, L., & Wu, L. (2020). Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm. Socio-Economic Planning Sciences, 72(June), 100916. https://doi.org/10.1016/j.seps.2020.100916
16	Ma, C., Liu, Z., Cao, Z., Song, W., Zhang, J., & Zeng, W. (2020). Cost-sensitive deep forest for price prediction. Pattern Recognition, 107, 107499. https://doi.org/10.1016/j.patcog.2020.107499
17	Park, B., & Kwon Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
18	Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. Journal of Property Research, 36(1), 59–96. https://doi.org/10.1080/09599916.2019.1587489
19	Polohakul, J., Chuangsuwanich, E., Suchato, A., & Punyabukkana, P. (2021). Real Estate Recommendation Approach for Solving the Item Cold-Start Problem. IEEE Access, 9, 68139–68150. https://doi.org/10.1109/ACCESS.2021.3077564
20	Posa, D., Perna, S., Resch, Y., Lupinek, C., Panetta, V., Hofmaier, S., Rohrbach, A., Hatzler, L., Grabenhenrich, L., Tsilochristou, O., Chen, K. W., Bauer, C. P., Hoffman, U., Forster, J., Zepp, F., Schuster, A., Wahn, U., Keil, T., Lau, S., … Matricardi, P. M. (2017). Evolution and predictive value of IgE responses toward a comprehensive panel of house dust mite allergens during the first 2 decades of life. Journal of Allergy and Clinical Immunology, 139(2), 541-549.e8. https://doi.org/10.1016/j.jaci.2016.08.014

Table 5. Literature Description: Articles' Year of Publication and Journal

ID	Year	Journal
1	2018	Geo-spatial Information Science
2	2021	Data Mining and Knowledge Discovery
3	2021	PLoS One
4	2014	Automation in Construction
5	2018	Habitat International
6	2016	Computers, Environment and Urban System
7	2013	Computers, Environment and Urban System
8	2021	Journal of Property Analysis
9	2019	Spatial Economic Analysis
10	2019	Built Environment Project and Asset Management
11	2015	Journal of Banking and Asset Management
12	2021	Land Use Policy
13	2014	International Journal of Geographical Information Science
14	2021	Technological Forecasting and Social Change
15	2020	Socio-Economic Planning Sciences
16	2020	Pattern Recognition
17	2015	Expert System with Application
18	2019	Journal of Property Research
19	2021	IEEE Access
20	2017	Journal of Allergy and Clinical Immunology
21	2018	Journal of Construction Engineering and Management
22	2021	Expert System with Application
23	2021	IEEE Access
24	2017	Journal of Property Research
25	2019	Data Mining and Knowledge Discovery
26	2022	Annals of Operations Research
27	2018	Journal of Property Research
28	2021	IEEE Access

Table 6. Literature Description: Articles' Countries of Origin

ID	Country
1	United State Of America
2	United State Of America, Bangladesh and Australia
3	United State Of America
4	Turkey
5	China
6	United State Of America and Netherlands
7	Germany and Austria
8	United State Of America
9	United State Of America
10	United State Of America
11	United State Of America, Signapure, New Zeland and United Kigdom
12	United State Of America and Brazil
13	Germany and Netherlands
14	China and United State Of America
15	China
16	China and Singapore
17	Korea
18	Colombia
19	Thailand
20	Germany, Vienna, Austria, Rome and Italy
21	United State Of America
22	United State Of America and Spain
23	Poland
24	United State Of America and Spain
25	United State Of America
26	United State Of America
27	United State Of America
28	Taiwan

Table 7. Literature Description: Articles' Title - from ID 1 to ID 28

ID	Title
1	Road distance and travel time for an improved house price Kriging predictor
2	Geo-spatial network embedding,Graph neural networks,House-price predictions,Real estate queries
3	Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches
4	House selling price assessment using two different adaptive neuro-fuzzy techniques
5	How does investor sentiment predict the future real estate returns of residential property in Hong Kong?
6	Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches
7	Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning
8	Predicting property prices with machine learning algorithms
9	Price competition in the spatial real estate market: allies or rivals?
10	Driving forces for the US residential housing price: a predictive analysis
11	New methodology for constructing real estate price indices applied to the Singapore residential market
12	Understanding house price appreciation using multi-source big geo-data and machine learning
13	Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging
14	The Spatial neural network model with disruptive technology for property appraisal in real estate industry
15	Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm
16	Cost-sensitive deep forest for price prediction
17	Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data
18	A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes
19	Real Estate Recommendation Approach for Solving the Item Cold-Start Problem
20	Evolution and predictive value of IgE responses toward a comprehensive panel of house dust mite allergens during the first 2 decades of life
21	Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes
22	Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain
23	Big data utilization,Google trends (GT),cross-correlation analysis,house price index (HPI),machine learning classification,prediction,real estate market,search volume index (SVI),time-lag
24	Can Web Search Queries Predict Prices Change on the Real Estate Market?
25	House price determinants in Athens: a spatial econometric approach
26	Real estate price estimation in French cities using geocoding and machine learning
27	House prices and proximity to kindergarten–costs of distance and external effects?
28	Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism

Table 8. Literature Description: Articles' Keywords - from ID 1 to ID 14

ID	Keywords
1	Kriging, Minkowski, real-estate valuation, road distance, travel time
2	Geo-spatial network embedding, Graph neural networks, House-price predictions, Real estate queries
3	NA
4	Adaptive neuro-fuzzy, House price, Prediction model, Real estate market
5	Behavioral finance, Boom and bust, Investor sentiment, Property bubbles, Real estate, Residential property, Sentiment index
6	Eigenvector spatial filtering, Geographically weighted regression, Hedonic models, Housing, Moving window regression, Prediction accuracy, Spatial expansion method
7	Airborne laser scanning, GIS, Generalized additive model,Hedonic regression, LiDAR, Real estate, Solar radiation, Vienna (Austria)
8	GBM, Machine Learning algorithms, RF, SVM, property valuation
9	Tokyo, real estate competition, spatial econometrics, strategic pricing
10	Data analytics, Hedonic pricing method, Housing prices, Predictive model, Residential property, Stepwise regression
11	Cooling measures, Explosive behavior, Hedonic models, Prediction, Real estate price index, Repeat sales
12	Geographically weighted regression, House photos, House price appreciation rate, Human mobility patterns, Street view images
1	Kriging, Minkowski, real-estate valuation, road distance, travel time
2	Geo-spatial network embedding, Graph neural networks, House-price predictions, Real estate queries
3	NA
4	Adaptive neuro-fuzzy, House price, Prediction model, Real estate market
5	Behavioral finance, Boom and bust, Investor sentiment, Property bubbles, Real estate, Residential property, Sentiment index
6	Eigenvector spatial filtering, Geographically weighted regression, Hedonic models, Housing, Moving window regression, Prediction accuracy, Spatial expansion method
7	Airborne laser scanning, GIS, Generalized additive model,Hedonic regression, LiDAR, Real estate, Solar radiation, Vienna (Austria)
8	GBM, Machine Learning algorithms, RF, SVM, property valuation
9	Tokyo, real estate competition, spatial econometrics, strategic pricing
10	Data analytics, Hedonic pricing method, Housing prices, Predictive model, Residential property, Stepwise regression
11	Cooling measures, Explosive behavior, Hedonic models, Prediction, Real estate price index, Repeat sales
12	Geographically weighted regression, House photos, House price appreciation rate, Human mobility patterns, Street view images
13	Vienna (Austria), geostatistics, housing, kriging, price prediction, real estate
14	Class activation mapping, Deep-Automated Optical Inspection (AOI), Disruptive technology, Real estate valuation, Spatial information, Spatial neural network

Table 9. Literature Description: Articles' Keywords - from ID 15 to ID 28

ID	Keywords
15	Housing prices, MHES, Predict, Time series, WOA
16	Cost-sensitive Deep Forest, Ensemble Deep Learning, Modified K-means, Price Prediction
17	AdaBoost, C4.5, Housing price index, Housing price prediction model, Machine learning algorithms, Naïve Bayes, RIPPER
18	Regression analysis, big data, machine learning, real estate, variable selection
19	Context awareness, machine learning, recommender systems, recurrent neural networks
20	Dermatophagoides pteronyssinus, House dust mite allergy, IgE, allergic rhinitis, asthma, birth cohort, children, component-resolved diagnostics, microarray, prediction, recombinant allergens
21	NA
22	Forecasting housing prices, Hedonic tools, Machine Learning, Models' explainability
23	Big data utilization, Google trends (GT), cross-correlation analysis, house price index (HPI), machine learning classification, prediction,real estate market, search volume index (SVI), time-lag
24	Spatial econometric, hedonic method, housing prices, price determination, weight matrices
25	External component, House prices, Internal component, Neighborhood value, Spatiotemporal effects
26	Artificial intelligence, Automated valuation models, French cities, Geocoding, Investment, Machine learning, Real estate market
27	House-prices, external effects, kindergarten, pre-school, transportation costs
28	Google satellite map, Heterogeneous data, House price prediction, Joint self-attention mechanism, Spatial transformer network

Table 10. Literature Description: Articles' Technologies

ID	Technology
1	Machine Learning
2	Deep Learning
3	Deep Learning
4	Neuro-fuzzy Techniques
5	Machine Learning
6	Machine Learning
7	Machine Learning
8	Machine Learning
9	Theoretical Study
10	Machine Learning
11	Theoretical Study
12	Machine Learning
13	Empirical Comparison
14	Machine Learning and Deep Learning
15	Machine Learning
16	Machine Learning
17	Machine Learning
18	Machine Learning
19	Machine Learning and Deep Learning
20	Theoretical Study
21	Machine Learning and Deep Learning
22	Machine Learning and Deep Learning
23	Machine Learning and Deep Learning
24	Machine Learning and Deep Learning
25	Machine Learning and Deep Learning
26	Machine Learning and Deep Learning
27	Theoretical Study
28	Machine Learning and Deep Learning

TABLE 11. Literature Description: Articles' Algorithms

ID	Algorithm
1	Geographically Eeighted Regression (GWR)
2	Geo-spatial Network Embedding (GSNE) and Graph Neural Network (GNN)
3	Geo-spatial Network Embedding (GSNE) and Graph Neural Network (GNN)
4	ANFIS-SC and ANFIS-GP
5	Sentiment Analysis
6	Spatial Expansion Method (SEM), Moving Window Regression (MWR), Geographically Weighted Regression (GWR), and Genetic Algorithm-based Eigenvector Spatial Filtering (ESF)
7	Hedonic Regression
8	Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM)
9	NA
10	American Housing Survey (AHS) and Hedonic Pricing Method (HPM)
11	NA
12	Geographically Weighted Regression (GWR)
13	Geo-statistical mapping
14	Support vector machine (SVM) and a Spatial Neural Network (SNN)
15	Holt's Exponential Smoothing (MHES) Method
16	Cost-sensitive Deep Forest; Ensemble Deep Learning
17	C4.5, RIPPER, Naïve Bayesian, and AdaBoost
18	Regression
19	Recommendation Systems, Recurrent Neural Networks (RNN) and Short-term Memory (LSTM)
20	NA
21	Back-propagation Neural Network (BPNN) and Support Vector Machine (SVM)
22	Nearest Neighbours (KNN), Decision Tree, Random Forest, AdaBoost
23	Web Search Queries, Geographically Weighted Regression (GWR), and Artificial Neural Networks (ANN)
24	Artificial Neural Network (ANN), Nearest Neighborhood (KNN), Spatial-temporal Lag (STLAG)
25	Artificial Neural Network (ANN)
26	Geographically Weighted Regressions (GWR) and Artificial Neural Network (ANN)
27	NA
28	Support vector machine (SVM), Extreme Gradient Boosting (XGBoost), Long Sort-term Memory (LSTM) and Convolutional Neural Networks (CNNs)

Table 12. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary I

Feature	Feature
airconditioningtypeid	Type of cooling system present in the home (if any)
architecturalstyletypeid	Architectural style of the home (i.e. ranch, colonial, split-level, etc...)
basementsqft	Finished living area below or partially below ground level
bathroomcnt	Number of bathrooms in home including fractional bathrooms
bedroomcnt	Number of bedrooms in home
buildingqualitytypeid	Overall assessment of condition of the building from best (lowest) to worst (highest)
buildingclasstypeid	The building framing type (steel frame, wood frame, concrete/brick)
calculatedbathnbr	Number of bathrooms in home including fractional bathroom
decktypeid	Type of deck (if any) present on parcel
threequarterbathnbr	Number of 3/4 bathrooms in house (shower + sink + toilet)
finishedfloor1squarefeet	Size of the finished living area on the first (entry) floor of the home
calculatedfinishedsquarefeet	Calculated total finished living area of the home
finishedsquarefeet6	Base unfinished and finished area
finishedsquarefeet12	Finished living area
finishedsquarefeet13	Perimeter living area
finishedsquarefeet15	Total area
finishedsquarefeet50	Size of the finished living area on the first (entry) floor of the home
fips	Federal Information Processing Standard code - see https://en.wikipedia.org/wiki/FIPS_county_code for more details
fireplacecnt	Number of fireplaces in a home (if any)
fireplaceflag	Is a fireplace present in this home
fullbathcnt	Number of full bathrooms (sink, shower + bathtub, and toilet) present in home
garagecarcnt	Total number of garages on the lot including an attached garage
garagetotalsqft	Total number of square feet of all garages on lot including an attached garage
hashottuborspa	Does the home have a hot tub or spa
heatingorsystemtypeid	Type of home heating system
latitude	Latitude of the middle of the parcel multiplied by 10e6
longitude	Longitude of the middle of the parcel multiplied by 10e6
lotsizesquarefeet	Area of the lot in square feet
numberofstories	Number of stories or levels the home has
parcelid	Unique identifier for parcels (lots)

Table 13. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary II

Feature	Feature
poolcnt	Number of pools on the lot (if any)
poolsizesum	Total square footage of all pools on property
pooltypeid10	Spa or Hot Tub
pooltypeid2	Pool with Spa/Hot Tub
pooltypeid7	Pool without hot tub
propertycountylandusecode	County land use code i.e. its zoning at the county level
propertylandusetypeid	Type of land use the property is zoned for
propertyzoningdesc	Description of the allowed land uses (zoning) for that property
rawcensustractandblock	Census tract and block ID combined - also contains blockgroup assignment by extension
censustractandblock	Census tract and block ID combined - also contains blockgroup assignment by extension
regionidcounty	County in which the property is located
regionidcity	City in which the property is located (if any)
regionidzip	Zip code in which the property is located
regionidneighborhood	Neighborhood in which the property is located
roomcnt	Total number of rooms in the principal residence
storytypeid	Type of floors in a multi-story house (i.e. basement and main level, split-level, attic, etc.). See tab for details.
typeconstructiontypeid	What type of construction material was used to construct the home
unitcnt	Number of units the structure is built into (i.e. 2 = duplex, 3 = triplex, etc...)
yardbuildingsqft17	Patio in yard
yardbuildingsqft26	Storage shed/building in yard
yearbuilt	The Year the principal residence was built
taxvaluedollarcnt	The total tax assessed value of the parcel
structuretaxvaluedollarcnt	The assessed value of the built structure on the parcel
landtaxvaluedollarcnt	The assessed value of the land area of the parcel
taxamount	The total property tax assessed for that assessment year
assessmentyear	The year of the property tax assessment
taxdelinquencyflag	Property taxes for this parcel are past due as of 2015
taxdelinquencyyear	Year for which the unpaid propert taxes were due

Table 14. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Heating System IDs and their Description

HeatingOrSystemTypeID	HeatingOrSystemDesc
1	Baseboard
2	Central
3	Coal
4	Convection
5	Electric
6	Forced air
7	Floor/Wall
8	Gas
9	Geo Thermal
10	Gravity
11	Heat Pump
12	Hot Water
13	None
14	Other
15	Oil
16	Partial
17	Propane
18	Radiant
19	Steam
20	Solar
21	Space/Suspended
22	Vent
23	Wood Burning
24	Yes
25	Zone

Table 15. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Property Land System IDs and their Description

PropertyLandUseTypeID	PropertyLandUseDesc
31	Commercial/Office/Residential Mixed Used
46	Multi-Story Store
47	Store/Office (Mixed Use)
246	Duplex (2 Units, Any Combination)
247	Triplex (3 Units, Any Combination)
248	Quadruplex (4 Units, Any Combination)
260	Residential General
261	Single Family Residential
262	Rural Residence
263	Mobile Home
264	Townhouse
265	Cluster Home
266	Condominium
267	Cooperative
268	Row House
269	Planned Unit Development
270	Residential Common Area
271	Timeshare
273	Bungalow
274	Zero Lot Line
275	Manufactured, Modular, Prefabricated Homes
276	Patio Home
279	Inferred Single Family Residential
290	Vacant Land - General
291	Residential Vacant Land

Table 16. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Story Type IDs and their Description

StoryTypeID	StoryDesc
1	Attic & Basement
2	Attic
3	Bi-Level with Attic & Basement
4	Bi-Level
5	Bi-Level with Attic
6	Bi-Level with Basement
7	Basement
8	Split Entry with Attic & Basement
9	Split Foyer with Attic & Basement
10	Level with Attic & Basement
11	Level with Attic
12	Level with Basement
13	Level
14	Multi-Level with Attic & Basement
15	Multi-Level
16	Multi-Level with Attic
17	Multi-Level with Basement
18	Split Level with Attic & Basement
19	Single Level with Attic & Basement
20	Split Entry with Attic
21	Split Entry with Basement
22	Split Foyer with Attic
23	Split Foyer with Basement
24	Single Level with Attic
25	Single Level with Basement
26	Single Level
27	Split Level with Attic
28	Split Level with Basement
29	Split Entry
30	Split Foyer
31	Split Level
32	Tri-level with Attic & Basement
33	Tri-level with Attic
34	Tri-level with Basement
35	Tri-level

TABLE 17. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Air Conditioning Types IDs and their Description

AirConditioningTypeID	AirConditioningDesc
1	Central
2	Chilled Water
3	Evaporative Cooler
4	Geo Thermal
5	None
6	Other
7	Packaged AC Unit
8	Partial
9	Refrigeration
10	Ventilation
11	Wall Unit
12	Window Unit
13	Yes

TABLE 18. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Architectural Style Type IDs and their Description

ArchitecturalStyleTypeID	ArchitecturalStyleDesc
1	A-Frame
2	Bungalow
3	Cape Cod
4	Cottage
5	Colonial
6	Custom
7	Contemporary
8	Conventional
9	Dome
10	French Provincial
11	Georgian
12	High Rise
13	Historical
14	Log Cabin/Rustic
15	Mediterranean
16	Modern
17	Mansion
18	English
19	Other
20	Prefab
21	Ranch/Rambler
22	Raised Ranch
23	Spanish
24	Traditional
25	Tudor
26	Unfinished/Under Construction
27	Victorian

Table 19. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Type Construction Type IDs and their Description

TypeConstructionTypeID	TypeConstructionDesc
1	Adobe
2	Brick
3	Concrete Block
4	Concrete
5	Dome
6	Frame
7	Heavy
8	Log
9	Light
10	Metal
11	Manufactured
12	Mixed
13	Masonry
14	Other
15	Steel
16	Stone
17	Tilt-Up
18	Wood

Table 20. Description of the Features Present in the Zillow Data Set - Zillow Data Dictionary Concerning the Type Building Class Type IDs and their Description

BuildingClassTypeID	BuildingClassDesc
1	Buildings having fireproofed structural steel frames carrying all wall, floor and roof loads. Wall, floor and roof structures are built of non-combustible materials.
2	Buildings having fireproofed reinforced concrete frames carrying all wall floor and roof loads which are all non-combustible.
3	Buildings having exterior walls built of a non-combustible material such as brick, concrete, block or poured concrete. Interior partitions and roof structures are built of combustible materials. Floor may be concrete or wood frame.
4	Buildings having wood or wood and steel frames
5	Specialized buildings that do not fit in any of the above categories

Table 21. Add caption

ID	Algoritm
1	KNN
2	KNN
3	DT
4	DT
5	RF
6	RF
7	SVM
8	SVM
9	CNN
10	CNN
11	MLP
12	MLP
13	LSTM
14	LSTM
15	GridSearchCV
16	GridSearchCV

Table 22. Add caption

ID	Algoritm
1	KNN
2	KNN
3	DT
4	DT
5	RF
6	RF
7	SVM
8	SVM
9	CNN
10	CNN
11	MLP
12	MLP
13	GridSearchCV
14	GridSearchCV

Table 23. Definition of Algorithm Parameters - from ID 1 to ID 10 - Zillow Data Set

ID	Algoritm	Algoritms Parameters
1	KNN	k = 5
2	KNN	k = 25

3	DT	ccp_alpha = 0.0,
		criterion='mae'
		max_depth=200
		max_features=10
		max_leaf_nodes=None
		min_samples_leaf=12
		min_samples_split=2
		random_state=40

4	DT	ccp_alpha=0.1
		criterion='absolute_error'
		max_depth=200
		max_features=25
		max_leaf_nodes=None
		min_samples_leaf=50
		min_samples_split=50
		random_state=40

5	RF	n_estimators=200
		max_features=auto
		random_state=42

6	RF	n_estimators=500
		max_features=25
		random_state=42

7	SVM	SVR: Without Parameters
8	SVM	SVC: Without Parameters

9	CNN	input=100 and activation='relu'
		input=60 and activation='relu'
		input=45 and activation='relu'
		input=1 and activation='sigmoid'

10	CNN	input=100 and activation='relu'
		input=60 and activation='relu'
		input=45 and activation='relu'
		input=1 and activation='softmax'

TABLE 24. Definition of Algorithm Parameters - from ID 11 to ID 16 - Zillow Data Set

ID	Algoritm	Algoritms Parameters

11	MLP	hidden_layer_sizes=(100,50)
		max_iter = 200
		activation = 'relu'
		solver = 'adam'

12	MLP	hidden_layer_sizes=(150,100,50)
		max_iter = 1000,
		activation = 'relu'
		solver = 'adam'

13	LSTM	input=100 and activation='relu'
		input=60 and activation='relu'
		input=45 and activation='relu'
		input=1 and activation='softmax'

14	LSTM	input=100 and activation='relu'
		input=60 and activation='relu'
		input=45 and activation='relu'
		input=1 and activation='sigmoid'

15	GridSearchCV	hidden_layer_sizes: [(100,50), (80,40), (50,30)] max_iter: [50, 100] activation: ['tanh', 'relu'] solver: ['sgd', 'adam'] alpha: [0.1, 0.05] learning_rate: ['constant','adaptive']
16	GridSearchCV	hidden_layer_sizes: [(150,100,50), (120,80,40), (100,50,30)] max_iter: [50, 100] activation: ['tanh', 'relu'] solver: ['sgd', 'adam'] alpha: [0.1, 0.05] learning_rate: ['constant','adaptive']

References

[1] M. Consultores. “Mais consultores.” (2022), [Online]. Available: https://www.maisconsultores.pt/.

[2] I. C. Education, “Artificial intelligence (ai),” 2020. [Online]. Available: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

[3] Oracle, “Artificial intelligence (ai),” 2022. [Online]. Available: https://www.oracle.com/artificial-intelligence/what-is-ai/.

[4] Microsoft, “Machine intelligence,” 2022. [Online]. Available: https://www.microsoft.com/en-us/research/theme/machine-intelligence/.

[5] E. Cohe and A. Jaffri, “Quick answer: What is the true return on ai investment?,” 2022. [Online]. Available: https://www.oracle.com/explore/oci-itdm/gartner-quick-answer?topic=%5C%20Artificial%5C%20Intelligence%5C&lb-mode=overlay%5C&source=:ow:o:s:mt:::RC_WWMK220613P00075:ArtIntel_pt%5C&intcmp=:ow:o:s:mt:::RC_WWMK220613P00075:ArtIntel_pt.

[6] Zillow, “Zillow,” 2022. [Online]. Available: https://www.zillow.com/.

[7] Zillow, “Zillow prize: Zillow’s home value prediction (zestimate),” 2017. [Online]. Available: https://www.kaggle.com/c/zillow-prize-1.

[8] Alfredo, “Alfredo,” 2022. [Online]. Available: https://alfredo.pt/s.

[9] Spacy. “Spacy.” (2022), [Online]. Available: https://spacy.io/.

[10] R. Coase, *If you Torture the Data Long Enough It Will Confess*. 2020.

[11] R. P. João Antão and R. Ribeiro, “Mobile crm development for real estate agents,” *Property Management*, pp. 938–957, 2022. DOI: 10.1108/PM-05-2021-0029. [Online]. Available: http://hdl.handle.net/10071/24578.

[12] B. Muthukadan. “Selenium-python.” (2022), [Online]. Available: https://selenium-python.readthedocs.io/.

[13] B. Muthukadan. “Selenium documentation.” (2022), [Online]. Available: https://www.selenium.dev/selenium/docs/api/py/api.html.

[14] Spacy. “Rule-based matching.” (2022), [Online]. Available: https://spacy.io/usage/rule-based-matching#entityruler.

[15] Spacy. “Vectors-similarity.” (2022), [Online]. Available: https://spacy.io/usage/linguistic-features/#vectors-similarity.

[16] Cambridge, *Cambridge Dictionary*. 2022. [Online]. Available: https://dictionary.cambridge.org/pt/dicionario/.

[17] N. Otsuka and M. Matsushita, "Constructing knowledge using exploratory text mining," pp. 1392–1397, 2014. DOI: 10.1109/SCIS-ISIS.2014.7044806. [Online]. Available: https://doi.org/10.1109/SCIS-ISIS.2014.7044806.

[18] K. Balog. "The standford natural language processing group." (2018), [Online]. Available: https://nlp.stanford.edu/projects/kbp/.

[19] A. Y. A. A. Tamanna Siddiqui and N. A. Khan, "Entity-oriented search.the information retrieval series," 2019. DOI: https://doi.org/10.1007/978-3-319-93935-3_6.

[20] X. Lin, H. Li, H. Xin, Z. Li, and L. Chen, "Kbpearl: A knowledge base population system supported by joint entity and relation linking," *Proc. VLDB Endow.*, vol. 13, no. 7, pp. 1035–1049, 2020. DOI: 10.14778/3384345.3384352. [Online]. Available: http://www.vldb.org/pvldb/vol13/p1035-lin.pdf.

[21] J. Ellis, J. Getman, J. Mott, *et al.*, "Linguistic resources for 2013 knowledge base population evaluations," in *Proceedings of the Sixth Text Analysis Conference, TAC 2013, Gaithersburg, Maryland, USA, November 18-19, 2013*, NIST, 2013. [Online]. Available: https://tac.nist.gov/publications/2013/additional.papers/KBP2013%5C_annotation%5C_overview.TAC2013.proceedings.pdf.

[22] P. P.Shinde and S. Shah, "A review of machine learning and deep learning applications," *IEEE*, 2018.

[23] I. B. M. C. (C. Education, *Machine Learning*, https://www.ibm.com/cloud/learn/machine-learning, Accessed: 2022-01-20, 2020.

[24] W. Blake, *The true method of knowledge is experiment.* 2021. [Online]. Available: https://nlp.stanford.edu/projects/kbp/.

[25] I. B. M. C. (C. Education, *Artificial Intelligence*, https://www.ibm.com/cloud/learn/what-is-artificial-intelligence, Accessed: 2022-01-20, 2020.

[26] L. Shiloh-Perl and R. Giryes, "Introduction to deep learning," *CoRR*, vol. abs/2003.03253, 2020. arXiv: 2003.03253. [Online]. Available: https://arxiv.org/abs/2003.03253.

[27] C. C. Aggarwal, "Neural networks and deep learning," *Springer*, 2018. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-94463-0.

[28] P. Tan, M. S. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining (Second Edition).* Pearson, 2019. [Online]. Available: https://www-users.cse.umn.edu/%5C%7Ekumar001/dmbook/index.php.

[29] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques, 3rd Edition.* Morgan Kaufmann, Elsevier, 2011, ISBN: 9780123748560. [Online]. Available: https://www.worldcat.org/oclc/262433473.

[30] I. H. Witten, E. Frank, and M. A. Hall, *Encyclopedia of Machine Learning.* Springer, 2011, ISBN: 978-0-387-30768-8. [Online]. Available: https://www.worldcat.org/oclc/262433473.

[31] "Coefficient of determination," in *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, pp. 88–91, ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_62. [Online]. Available: https://doi.org/10.1007/978-0-387-32833-1_62.

[32] G. Guo, H. Wang, D. A. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," R. Meersman, Z. Tari, and D. C. Schmidt, Eds., ser. Lecture Notes in Computer Science, vol. 2888, Springer, 2003, pp. 986–996. DOI: 10.1007/978-3-540-39964-3_62. [Online]. Available: https://doi.org/10.1007/978-3-540-39964-3%5C_62.

[33] O. Kramer, *Dimensionality Reduction with Unsupervised Nearest Neighbors* (Intelligent Systems Reference Library). Springer, 2013, vol. 51, ISBN: 978-3-642-38651-0. DOI: 10.1007/978-3-642-38652-7. [Online]. Available: https://doi.org/10.1007/978-3-642-38652-7.

[34] G. Santos, "Knn regression model in python," 2022. [Online]. Available: https://towardsdatascience.com/knn-regression-model-in-python-9868f21c9fa2.

[35] L. Liberti and C. Lavor, *Euclidean Distance Geometry - an Introduction* (Springer undergraduate texts in mathematics and technology). Springer, 2017, ISBN: 978-3-319-60791-7. DOI: 10.1007/978-3-319-60792-4. [Online]. Available: https://doi.org/10.1007/978-3-319-60792-4.

[36] D. Xu, W. Gui, and J. He, Eds., *A novel Minkowski-distance-based consensus clustering algorithm*. Springer US, 2017. DOI: 10.1007/s11633-016-1033-z. [Online]. Available: https://doi.org/10.1007/s11633-016-1033-z.

[37] S. Z. Li and A. K. Jain, Eds., *Encyclopedia of Biometrics, Second Edition*. Springer US, 2015, ISBN: 978-1-4899-7487-7. DOI: 10.1007/978-1-4899-7488-4. [Online]. Available: https://doi.org/10.1007/978-1-4899-7488-4.

[38] S. Wang and W. Shi, "Data mining and knowledge discovery," in *Die Dynamik sozialer und sprachlicher Netzwerke, Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*, B. Frank-Job, A. Mehler, and T. Sutter, Eds., Springer, 2012, pp. 49–58. DOI: 10.1007/978-3-540-72680-7_5. [Online]. Available: https://doi.org/10.1007/978-3-540-72680-7%5C_5.

[39] L. Liu and M. T. Özsu, Eds., *Encyclopedia of Database Systems, Second Edition*. Springer, 2018, ISBN: 978-1-4614-8266-6. DOI: 10.1007/978-1-4614-8265-9. [Online]. Available: https://doi.org/10.1007/978-1-4614-8265-9.

[40] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *Information Computing and Applications - Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings*, B. Liu, M. Ma, and J. Chang, Eds., ser. Lecture Notes in Computer Science, vol. 7473, Springer, 2012, pp. 246–252. DOI: 10.1007/978-3-642-34062-8_32. [Online]. Available: https://doi.org/10.1007/978-3-642-34062-8%5C_32.

[41] L. Breiman, "Randon florest," 2001.

[42] E. Bayro-Corrochano and N. Arana-Daniel, "Theory and applications of clifford support vector machines," *J. Math. Imaging Vis.*, vol. 28, no. 1, pp. 29–46, 2007. DOI: 10.1007/s10851-007-0008-7. [Online]. Available: https://doi.org/10.1007/s10851-007-0008-7.

[43] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in *Machine Learning and Its Applications, Advanced Lectures*, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos, Eds., ser. Lecture Notes in Computer Science, vol. 2049, Springer, 2001, pp. 249–257. DOI: 10.1007/3-540-44673-7_12. [Online]. Available: https://doi.org/10.1007/3-540-44673-7%5C_12.

[44] I. B. M. C. (C. Education, *Deep Learning*, https://www.ibm.com/cloud/learn/deep-learning, Accessed: 2022-01-20, 2020.

[45] S. Skansi, *Introduction to Deep Learning.* 2018. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-73004-2.

[46] F. Tan, C. Cheng, and Z. Wei, "Modeling and elucidation of housing price," *Data Min. Knowl. Discov.*, vol. 33, no. 3, pp. 636–662, 2019. DOI: 10.1007/s10618-018-00612-0. [Online]. Available: https://doi.org/10.1007/s10618-018-00612-0.

[47] S. S. S. Das, M. E. Ali, Y. Li, Y. Kang, and T. Sellis, "Boosting house price predictions using geo-spatial network embedding," *CoRR*, vol. abs/2009.00254, 2020. arXiv: 2009.00254. [Online]. Available: https://arxiv.org/abs/2009.00254.

[48] S. S. S. Das, M. E. Ali, Y. Li, Y. Kang, and T. Sellis, "Boosting house price predictions using geo-spatial network embedding," *CoRR*, vol. abs/2009.00254, 2020. arXiv: 2009.00254. [Online]. Available: https://arxiv.org/abs/2009.00254.

[49] S. Law, B. Paige, and C. Russell, "Take a look around: Using street view and satellite images to estimate house prices," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, 54:1–54:19, 2019. DOI: 10.1145/3342240. [Online]. Available: https://doi.org/10.1145/3342240.

[50] J. Polohakul, E. Chuangsuwanich, A. Suchato, and P. Punyabukkana, "Real Estate Recommendation Approach for Solving the Item Cold-Start Problem," *IEEE Access*, vol. 9, pp. 68 139–68 150, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3077564.

[51] P. Wang, C. Chen, J. Su, T. Wang, and S. Huang, "Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism," *IEEE Access*, vol. 9, pp. 55 244–55 259, 2021. DOI: 10.1109/ACCESS.2021.3071306. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3071306.

[52] F. Baeta, J. Correia, T. Martins, and P. Machado, "Exploring genetic programming in tensorflow with tensorgp," *SN Comput. Sci.*, vol. 3, no. 2, p. 154, 2022. DOI: 10.1007/s42979-021-01006-8. [Online]. Available: https://doi.org/10.1007/s42979-021-01006-8.

[53] J. C. Kunz, I. F. C. Smith, and T. Tomiyama, "Advanced engineering informatics editorial," *Adv. Eng. Informatics*, vol. 23, no. 2, p. 139, 2009. DOI: 10.1016/j.aei.2008.12.002. [Online]. Available: https://doi.org/10.1016/j.aei.2008.12.002.

[54] B. Kristjansson and D. Lee, "The mpl modeling system," in *Modeling Languages in Mathematical Optimization*, J. Kallrath, Ed. Boston, MA: Springer US, 2004, pp. 239–266, ISBN: 978-1-4613-0215-5. DOI: 10.1007/978-1-4613-0215-5_13. [Online]. Available: https://doi.org/10.1007/978-1-4613-0215-5_13.

[55] C. Ma, Z. Liu, Z. Cao, W. Song, J. Zhang, and W. Zeng, "Cost-sensitive deep forest for price prediction," *Pattern Recognit.*, vol. 107, p. 107 499, 2020. DOI: 10.1016/j.patcog.2020.107499. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107499.

[56] Q. Ren, M. Li, H. Li, and Y. Shen, "A novel deep learning prediction model for concrete dam displacements using interpretable mixed attention mechanism," *Adv. Eng. Informatics*, vol. 50, p. 101 407, 2021. DOI: 10.1016/j.aei.2021.101407. [Online]. Available: https://doi.org/10.1016/j.aei.2021.101407.

[57] R. M. Razeira and I. A. Rodello, "A LSTM recurrent neural network implementation for classifying entities on brazilian legal documents," in *Computational Science and Its Applications - ICCSA 2021 - 21st International Conference, Cagliari, Italy, September 13-16, 2021, Proceedings, Part II*, O. Gervasi, B. Murgante, S. Misra, *et al.*, Eds., ser. Lecture Notes in Computer Science, vol. 12950, Springer, 2021, pp. 648–656. DOI: 10.1007/978-3-030-86960-1_48. [Online]. Available: https://doi.org/10.1007/978-3-030-86960-1%5C_48.

[58] L. B. Moreira and A. A. Namen, "A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia," *Computer Methods and Programs in Biomedicine*, vol. 165, pp. 139–149, 2018, ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2018.08.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718307569.

[59] M. Kim, Y. Xu, O. R. Zaïane, and R. Goebel, "Recognition of patient-related named entities in noisy tele-health texts," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 4, 59:1–59:23, 2015. DOI: 10.1145/2651444. [Online]. Available: https://doi.org/10.1145/2651444.

[60] X. Xu, X. Yin, and X. Chen, "A large-group emergency risk decision method based on data mining of public attribute preferences," *Knowledge-Based Systems*, vol. 163, pp. 495–509, 2019, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2018.09.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705118304611.

[61] N. Otsuka and M. Matsushita, "Constructing knowledge using exploratory text mining," in *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems*

(ISIS), Kita-Kyushu, Japan, December 3-6, 2014, IEEE, 2014, pp. 1392–1397. DOI: 10.1109/SCIS-ISIS.2014.7044806. [Online]. Available: https://doi.org/10.1109/SCIS-ISIS.2014.7044806.

[62] M. Lamba and M. Madhusudhan, *Text Mining for Information Professionals - An Uncharted Territory*. Springer, 2022, ISBN: 978-3-030-85084-5. DOI: 10.1007/978-3-030-85085-2. [Online]. Available: https://doi.org/10.1007/978-3-030-85085-2.

[63] Scimago. "Scimago journal & country rank xplore." (2022), [Online]. Available: https://www.scimagojr.com/.

[64] G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*, H. van Halteren, Ed. Dordrecht: Springer Netherlands, 1999, pp. 117–133, ISBN: 978-94-015-9273-4. DOI: 10.1007/978-94-015-9273-4_9. [Online]. Available: https://doi.org/10.1007/978-94-015-9273-4_9.

[65] H. R. Kouchaksaraei and H. Karl, "Service function chaining across openstack and kubernetes domains," *DEBS 2019 - Proceedings of the 13th ACM International Conference on Distributed and Event-Based Systems*, pp. 240–243, 2019. DOI: 10.1145/3328905.3332505.

[66] IEEE. "Ieee xplore." (2022), [Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp.

[67] W. of Science. "Clarivate." (2022), [Online]. Available: https://www.webofscience.com/wos/woscc/basic-search.

[68] H. Crosby, T. Damoulas, A. Caton, P. Davis, J. Porto de Albuquerque, and S. A. Jarvis, "Road distance and travel time for an improved house price Kriging predictor," *Geo-Spatial Information Science*, vol. 21, no. 3, pp. 185–194, 2018, ISSN: 10095020. DOI: 10.1080/10095020.2018.1503775. [Online]. Available: https://doi.org/10.1080/10095020.2018.1503775.

[69] S. S. S. Das, M. E. Ali, Y. F. Li, Y. B. Kang, and T. Sellis, "Boosting house price predictions using geo-spatial network embedding," *Data Mining and Knowledge Discovery*, vol. 35, no. 6, pp. 2221–2250, 2021, ISSN: 1573756X. DOI: 10.1007/s10618-021-00789-x. arXiv: 2009.00254. [Online]. Available: https://doi.org/10.1007/s10618-021-00789-x.

[70] W. T. Embaye, Y. A. Zereyesus, and B. Chen, "Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches," *PLoS ONE*, vol. 16, no. 2 February, pp. 1–20, 2021, ISSN: 19326203. DOI: 10.1371/journal.pone.0244953. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0244953.

[71] I. H. Gerek, "House selling price assessment using two different adaptive neuro-fuzzy techniques," *Automation in Construction*, vol. 41, pp. 33–39, 2014, ISSN: 09265805. DOI: 10.1016/j.autcon.2014.02.002. [Online]. Available: http://dx.doi.org/10.1016/j.autcon.2014.02.002.

[72] C. Hei-Ling Lam and E. Chi-Man Hui, "How does investor sentiment predict the future real estate returns of residential property in hong kong?" *Habitat International*, vol. 75, no. July 2017, pp. 1–11, 2018, ISSN: 01973975. DOI: 10.1016/j.habitatint.2018.02.009. [Online]. Available: https://doi.org/10.1016/j.habitatint.2018.02.009.

[73] M. Helbich and D. A. Griffith, "Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches," *Computers, Environment and Urban Systems*, vol. 57, pp. 1–11, 2016, ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2015.12.002.

[74] M. Helbich, A. Jochem, W. Mücke, and B. Höfle, "Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning," *Computers, Environment and Urban Systems*, vol. 39, pp. 81–92, 2013, ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2013.01.001. [Online]. Available: http://dx.doi.org/10.1016/j.compenvurbsys.2013.01.001.

[75] W. K. Ho, B. S. Tang, and S. W. Wong, "Predicting property prices with machine learning algorithms," *Journal of Property Research*, vol. 38, no. 1, pp. 48–70, 2021, ISSN: 14664453. DOI: 10.1080/09599916.2020.1832558. [Online]. Available: https://doi.org/10.1080/09599916.2020.1832558.

[76] S. Iwata, K. Sumita, and M. Fujisawa, "Price competition in the spatial real estate market: allies or rivals?" *Spatial Economic Analysis*, vol. 14, no. 2, pp. 174–195, 2019, ISSN: 17421780. DOI: 10.1080/17421772.2019.1532596. [Online]. Available: https://doi.org/10.1080/17421772.2019.1532596.

[77] A. Jafari and R. Akhavian, "Driving forces for the US residential housing price: a predictive analysis," *Built Environment Project and Asset Management*, vol. 9, no. 4, pp. 515–529, 2019, ISSN: 20441258. DOI: 10.1108/BEPAM-07-2018-0100.

[78] L. Jiang, P. C. Phillips, and J. Yu, "New methodology for constructing real estate price indices applied to the Singapore residential market," *Journal of Banking and Finance*, vol. 61, S121–S131, 2015, ISSN: 03784266. DOI: 10.1016/j.jbankfin.2015.08.026. [Online]. Available: http://dx.doi.org/10.1016/j.jbankfin.2015.08.026.

[79] Y. Kang, F. Zhang, W. Peng, *et al.*, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy*, vol. 111, no. September 2019, p. 104 919, 2021, ISSN: 02648377. DOI: 10.1016/j.landusepol.2020.104919. [Online]. Available: https://doi.org/10.1016/j.landusepol.2020.104919.

[80] M. Kuntz and M. Helbich, "Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1904–1921, 2014, ISSN: 13623087. DOI: 10.1080/13658816.2014.906041.

[81] R. F. Y. Lin, C. Ou, K. K. Tseng, D. Bowen, K. L. Yung, and W. H. Ip, "The Spatial neural network model with disruptive technology for property appraisal in real estate industry," *Technological Forecasting and Social Change*, vol. 173, no. August, 2021, ISSN: 00401625. DOI: 10.1016/j.techfore.2021.121067.

[82] L. Liu and L. Wu, "Predicting housing prices in China based on modified Holt's exponential smoothing incorporating whale optimization algorithm," *Socio-Economic Planning Sciences*, vol. 72, no. June, p. 100 916, 2020, ISSN: 00380121. DOI: 10.1016/j.seps.2020.100916. [Online]. Available: https://doi.org/10.1016/j.seps.2020.100916.

[83] C. Ma, Z. Liu, Z. Cao, W. Song, J. Zhang, and W. Zeng, "Cost-sensitive deep forest for price prediction," *Pattern Recognition*, vol. 107, p. 107 499, 2020, ISSN: 00313203. DOI: 10.1016/j.patcog.2020.107499. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107499.

[84] B. Park and J. Kwon Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications*, vol. 42, no. 6, pp. 2928–2934, 2015, ISSN: 09574174. DOI: 10.1016/j.eswa.2014.11.040. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2014.11.040.

[85] J. I. Pérez-Rave, J. C. Correa-Morales, and F. González-Echavarría, "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes," *Journal of Property Research*, vol. 36, no. 1, pp. 59–96, 2019, ISSN: 14664453. DOI: 10.1080/09599916.2019.1587489. [Online]. Available: https://doi.org/10.1080/09599916.2019.1587489.

[86] M. H. Rafiei and H. Adeli, "Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes," *Journal of Construction Engineering and Management*, vol. 144, no. 12, pp. 1–9, 2018, ISSN: 0733-9364. DOI: 10.1061/(asce)co.1943-7862.0001570.

[87] J. R. Rico-Juan and P. Taltavull de La Paz, "Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain," *Expert Systems with Applications*, vol. 171, no. October 2020, 2021, ISSN: 09574174. DOI: 10.1016/j.eswa.2021.114590.

[88] N. Rizun and A. Baj-Rogowska, "Can Web Search Queries Predict Prices Change on the Real Estate Market?" *IEEE Access*, vol. 9, pp. 70 095–70 117, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3077860.

[89] M. Stamou, A. Mimis, and A. Rovolis, "House price determinants in Athens: a spatial econometric approach," *Journal of Property Research*, vol. 34, no. 4, pp. 269–284, 2017, ISSN: 14664453. DOI: 10.1080/09599916.2017.1400575. [Online]. Available: https://doi.org/10.1080/09599916.2017.1400575.

[90] D. Tchuente and S. Nyawa, *Real estate price estimation in French cities using geocoding and machine learning*. Springer US, 2022, vol. 308, pp. 571–608, ISBN:

0123456789. DOI: 10.1007/s10479-021-03932-5. [Online]. Available: https://doi.org/10.1007/s10479-021-03932-5.

[91] T. Theisen and A. W. Emblem, “House prices and proximity to kindergarten–costs of distance and external effects?” *Journal of Property Research*, vol. 35, no. 4, pp. 321–343, 2018, ISSN: 14664453. DOI: 10.1080/09599916.2018.1513057. [Online]. Available: https://doi.org/10.1080/09599916.2018.1513057.

[92] P. Y. Wang, C. T. Chen, J. W. Su, T. Y. Wang, and S. H. Huang, “Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism,” *IEEE Access*, vol. 9, pp. 55 244–55 259, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2021.3071306.

[93] I. C. Education, “Crisp-dm help overview,” 2021. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview.

[94] S. Vision, “What is the crisp-dm methodology?,” 2022. [Online]. Available: https://www.sv-europe.com/crisp-dm-methodology/.

[95] N. HOTZ, “What is crisp dm?,” 2022. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/.

[96] Imovirtual. “Imovirtual.” (2022), [Online]. Available: https://www.imovirtual.com/.

[97] Google. “Colab.” (2022), [Online]. Available: https://colab.research.google.com/.