

iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Modelos Automatizados de Avaliação de Imóveis: Aplicação à Área Metropolitana de Lisboa

João Miguel Pequito Zacarias

Mestrado em Métodos Analíticos para Gestão

Orientador:

Professor Doutor Raul Manuel da Silva Laureano, Prof. Associado, Iscte Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Outubro, 2022

iscte

**BUSINESS
SCHOOL**

Departamento de Métodos Quantitativos para Gestão e Economia

Modelos Automatizados de Avaliação de Imóveis: Aplicação à Área Metropolitana de Lisboa

João Miguel Pequito Zacarias

Mestrado em Métodos Analíticos para Gestão

Orientador:

Professor Doutor Raul Manuel da Silva Laureano, Prof. Associado, Iscte Business School, Departamento de Métodos Quantitativos para Gestão e Economia

Outubro, 2022

Aos meus pais, à minha irmã e aos meus avós.

Agradecimentos

A elaboração desta dissertação não seria possível sem o contributo de alguns intervenientes, a quem importa agradecer.

Em primeiro lugar, agradecer, particularmente, a disponibilidade, o apoio e o conhecimento demonstrados desde o primeiro até ao último momento por parte do meu orientador e, também, diretor de mestrado, Professor Doutor Raul Manuel da Silva Laureano.

Agradecer, também, a todos os professores do Iscte que permitiram aumentar o meu conhecimento nesta e noutras áreas, fazendo de mim um aluno preparado para concluir esta investigação.

Deixo o meu agradecimento especial, também, aos meus pais, à minha irmã e aos restantes familiares que sempre me apoiaram ao longo da vida e me incentivaram sempre a concluir todas as etapas a que me propunha mesmo com os desafios exigentes adjacentes às mesmas.

Por fim, mas igualmente importante, deixar o meu agradecimento à minha namorada e aos meus amigos que sempre me apoiaram na realização desta dissertação.

Resumo

A avaliação fornece uma medida dos benefícios e responsabilidades de um imóvel e, desde sempre, depende muito dos métodos tradicionais, aplicados por peritos avaliadores. Apesar destes resistirem à evolução dos métodos de previsão do valor, têm surgido sistemas especializados na determinação do mesmo, os modelos automatizados de avaliação (AVM), estimando-se que já serviram mais de 100 milhões de entidades. Os AVM são programas informáticos que estimam o valor dos imóveis através da análise de atributos relacionados com os imóveis.

Além disso, os peritos são acusados de subjetividade nas avaliações, surgindo a necessidade de melhorar a qualidade da previsão através de sistemas independentes.

Assim, esta dissertação foca-se na construção de modelos preditivos com dados de uma plataforma imobiliária, da Área Metropolitana de Lisboa, recorrendo a técnicas de *data mining*. Para se construir modelo, procede-se a uma análise estatística, como forma de ajustamento da base de dados, e aplicam-se técnicas preditivas. A utilização de diversos algoritmos visa a obtenção do melhor modelo possível, com a melhor capacidade preditiva do valor do imóvel, e tendo como variáveis predictoras as características do imóvel e da sua localização. O melhor modelo obtido, é uma rede neuronal e apresenta um erro médio de previsão de 0,06 e um coeficiente de determinação de 0,90.

Um modelo eficaz é uma ferramenta muito útil de apoio aos *stakeholders*, sendo eles peritos, compradores, vendedores, bancos, Estado, entre outros. Em suma, esta investigação é muito importante para o setor imobiliário, que tem muita importância em qualquer economia.

Palavras-Chave: data mining; modelos preditivos; imobiliário; AVM; avaliação; Área Metropolitana de Lisboa.

(JEL Classification System: C1, R31)

Abstract

Valuation provides a measure of the benefits and responsibilities of a property and has always depended heavily on traditional methods applied by expert appraisers. Despite they resist the evolution of value forecasting methods, specialized systems have emerged to determine the value, the automated valuation models (AVM), estimated to have served more than 100 million entities. AVMs are computer programs that estimate the value of properties through the analysis of attributes related to the properties.

In addition, experts are accused of subjectivity in valuations, which raises the need to improve the quality of forecasting through independent systems.

Thus, this dissertation focuses on the construction of predictive models with data from a real estate platform, in the Lisbon Metropolitan Area, using data mining techniques. In order to build a model, statistical analyzes is carried out, as a way of adjusting the database, and predictive techniques are be applied. The use of these techniques aims to obtain the best possible model, with the best predictive capacity of the value of the property and having as predictors the characteristics of the property and its location. The best model obtained is a neural network and presents an average prediction error of 0.06 and a coefficient of determination of 0.90.

An effective model is a very useful tool to support stakeholders, including experts, buyers, sellers, banks, the State, among others. In short, this investigation is very important for the real estate sector, which is very important in any economy.

Keywords: data mining; predictive models; real estate; AVM; valuation; Lisbon Metropolitan Area.

(JEL Classification System: C1, R31)

Índice

Agradecimentos	i
Resumo	iii
Abstract	v
Índice de Figuras	ix
Índice de Tabelas	xi
Lista de Acrónimos e Siglas	xiii
1. Introdução	1
1.1. Enquadramento	1
1.2. Problema, questão de investigação e objetivos.....	1
1.3. Abordagem Metodológica	2
1.4. Organização	3
2. Revisão da Literatura	5
2.1. Revisão Sistemática da Literatura sobre AVM	5
2.1.1. Protocolo da Revisão Sistemática da Literatura	5
2.1.2. Enquadramento dos AVM.....	9
2.1.3. Contextos de utilização dos AVM	12
2.1.4. Construção de AVM: Metodologia	13
2.1.4.1. Compreensão e preparação dos dados	13
2.1.4.2. Algoritmos.....	16
2.1.4.3. Avaliação da qualidade dos AVM.....	18
2.1.5. Avaliação da qualidade dos artigos da RSL	19
2.2. Contexto nacional e questões legais associadas	20
3. Metodologia	23
3.1. Compreensão do Negócio	23
3.2. Compreensão dos Dados	24
3.2.1. Análise Exploratória	25
3.3. Preparação dos Dados	30
3.4. Modelação	33
3.5. Avaliação.....	34
3.6. Implementação.....	35
4. Resultados e Discussão	37
4.1. Caracterização dos valores dos imóveis	37
4.2. Relação entre os valores dos imóveis e as suas características	37
4.3. Modelos preditivos dos valores dos imóveis.....	40
4.3.1. Qualidade dos modelos preditivos	40
4.3.2. Perfis de imóveis com avaliações certas e erradas.....	49
5. Conclusões e recomendações	51
5.1. Síntese.....	51

5.2.	Contributos	52
5.3.	Limitações	52
5.4.	Trabalho Futuro	53
6.	<i>Bibliografia</i>	55
7.	<i>Anexos</i>	61
7.1.	Anexo A: Dicionário de dados	61
7.2.	Anexo B: Correlação de variáveis	63
7.3.	Anexo C: Equação da Regressão	69
7.4.	Anexo D: Regras das árvores	69
7.5.	Anexo E: <i>Rule set</i> do melhor modelo	70
7.6.	Anexo F: <i>Rule set</i> do modelo para perfis ajustado, subestimado e sobrestimado	73
7.7.	Anexo G: Regras de perfis de imóveis ajustados, subestimados e sobrestimados	76

Índice de Figuras

Figura 2.1 - Processo de seleção de artigos alvo da RSL	6
Figura 2.2 - Nuvem de palavras das palavras-chave dos artigos	9
Figura 3.1 - Valor (média) dos imóveis por concelho	27
Figura 4.1 - Relação entre a área do imóvel e o logaritmo do valor previsto	44
Figura 4.2 - Relação entre a área do imóvel e o logaritmo do valor observado.....	44
Figura 4.3 - Logaritmo médio do valor por lugares de estacionamento	44
Figura 4.4 - Relação entre o poder de compra per capita e o logaritmo do valor previsto	45
Figura 4.5 - Relação entre o poder de compra per capita e o logaritmo do valor real	45
Figura 4.6 - Média do logaritmo do valor previsto por concelho	46
Figura 4.7 - Média do logaritmo do valor real por concelho	46
Figura 4.8 - Média do logaritmo do valor previsto por freguesia	47
Figura 4.9 - Média do logaritmo do valor real por freguesia	47
Figura 4.10 - Frequência do tipo de estimativa do modelo	50

Índice de Tabelas

Tabela 2.1 - Critérios de inclusão e exclusão	6
Tabela 2.2 - Artigos recolhidos.....	7
Tabela 2.3 - Critérios de qualidade	9
Tabela 2.4 - Vantagens e desvantagens dos AVM	12
Tabela 2.5 - Aplicações dos AVM	12
Tabela 2.6 - Contexto atual dos AVM em alguns países	13
Tabela 2.7 - Variáveis relacionadas com venda dos imóveis	14
Tabela 2.8 - Variáveis relacionadas com as características dos imóveis	14
Tabela 2.9 - Variáveis relacionadas com a localização dos imóveis	15
Tabela 2.10 - Caracterização da componente empírica dos estudos	16
Tabela 2.11 - Avaliação dos artigos.....	20
Tabela 3.1 - Fontes de dados	25
Tabela 3.2 - Estatísticas descritivas de variáveis quantitativas relativas aos imóveis	26
Tabela 3.3 - Estatísticas descritivas das variáveis qualitativas.....	26
Tabela 3.4 - Estatísticas descritivas de variáveis relativas a características da freguesia do imóvel	28
Tabela 3.5 - Estatísticas descritivas de variáveis relativas a características do concelho do imóvel	28
Tabela 3.6 - Associação entre variáveis predictoras e o valor do imóvel.....	29
Tabela 3.7 - Dicionário de dados e procedimento das novas variáveis relativas ao imóvel	31
Tabela 3.8 - Estatísticas descritivas das novas variáveis qualitativas	32
Tabela 3.9 - Estatísticas descritivas das novas variáveis quantitativas.....	32
Tabela 3.10 - Parametrizações do algoritmo CART.....	34
Tabela 3.11 - Parametrizações do algoritmo CHAID	34
Tabela 3.12 - Descrição das métricas de avaliação dos modelos	35
Tabela 4.1 - Valor do imóvel pelas suas características.....	38
Tabela 4.2 - Valor do imóvel por distrito, concelho e freguesia (top 10)	39
Tabela 4.3 - Métricas de avaliação do modelo de regressão linear múltipla	41
Tabela 4.4 - Métricas de avaliação dos modelos CART.....	41
Tabela 4.5 - Métricas de avaliação de treino dos modelos CHAID	42
Tabela 4.6 - Métricas de avaliação dos modelos das redes neuronais.....	42
Tabela 4.7 - Variáveis predictoras mais importantes	43
Tabela 4.8 - Estatísticas descritivas das variáveis do processo de análise de resíduos	49
Tabela 7.1 - Dicionário de dados de variáveis relativas ao imóvel	61
Tabela 7.2 - Dicionário de dados de variáveis relativas à população da freguesia do imóvel	61
Tabela 7.3 - Dicionário de dados de variáveis relativas a características da freguesia do imóvel	62
Tabela 7.4 - Dicionário de dados de variáveis de indicadores económicos do concelho do imóvel	62
Tabela 7.5 - Correlação com variáveis do imóvel e da população.....	63
Tabela 7.6 - Correlação com variáveis da população.....	64
Tabela 7.7 - Correlação com variáveis da população e características da freguesia.....	65
Tabela 7.8 - Correlação com variáveis de escolas, hospitais, farmácias e transportes	66
Tabela 7.9 - Correlação com variáveis de transportes, estabelecimentos policiais e praias	67
Tabela 7.10 - Correlação com variáveis económicas	68

Lista de Acrónimos e Siglas

Sigla/Acrónimo	Designação
AML	Área Metropolitana de Lisboa
AVM	Modelos Automatizados de Avaliação
COD	Coeficiente de Dispersão
CART	<i>Classification And Regression Trees</i>
CHAID	<i>Chi-squared Automatic Interaction Detection</i>
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
DM	<i>Data Mining</i>
EVIP	<i>European Valuation Information Paper</i>
EVS	<i>European Valuation Standards</i>
GNR	Guarda Nacional Republicana
IAAO	<i>International Association of Assessing Officers</i>
INE	Instituto Nacional de Estatística
IFRS	<i>International Financial Reporting Standards</i>
IVS	<i>International Valuation Standards</i>
IVSC	<i>International Valuation Standards Committee</i>
KDD	<i>Knowledge Discovery in Databases</i>
MAE	Erro médio absoluto
MSE	Erro quadrático médio
PSP	Polícia de Segurança Pública
R	Coeficiente de correlação
R ²	Coeficiente de determinação
RICS	<i>Royal Institution of Chartered Surveyors</i>
RL	Revisão da Literatura
RSL	Revisão Sistemática da Literatura
SEMMA	<i>Sample, Explore, Modify, Model and Assess</i>
TEGoVA	<i>The European Group of Valuers' Associations</i>
UE	União Europeia

1. Introdução

1.1. Enquadramento

A investigação foca os Modelos Automatizados de Avaliação (AVM), que são programas informáticos que estimam o valor de um imóvel através da análise de diversos fatores relacionados com os imóveis e sua localização a partir de dados históricos do comportamento do mercado imobiliário (IAAO, 2018).

Os AVM, aplicados ao setor imobiliário, têm progredido nos últimos anos, estimando-se que já serviram mais de 100 milhões de entidades (Steurer *et al.*, 2021).

Atualmente, as avaliações são efetuadas por peritos avaliadores, porém, podem existir fatores que os façam avaliar um imóvel num valor superior ou inferior ao normal se a avaliação for feita de forma justa (Bogin & Shui, 2020). A avaliação gera uma medida dos benefícios e responsabilidades de um imóvel (Fu *et al.*, 2013) e é importante para os compradores, vendedores, investidores ou para os credores (Bellotti, 2017). Além disso, é importante em vários aspetos no setor como transações de imóveis, transferências de propriedade, questões fiscais ou heranças (Poursaeed *et al.*, 2018).

Esta investigação é um contributo importante para o mercado imobiliário, que é um dos setores com maior importância em qualquer economia nacional, pois a previsão dos valores dos imóveis é uma tarefa muito complexa devido a diversos fatores que influenciam a sua qualidade (Pai & Wang, 2020), com a inteligência artificial, que nas últimas duas décadas tem visto o seu “caminho” cruzar-se com o setor imobiliário, provando ser uma ferramenta importante para o mesmo (Birkeland *et al.*, 2021).

Além disso, no prisma dos peritos avaliadores, pode ser útil, permitindo estabelecer uma comparação entre valor do imóvel atribuído pelo modelo preditivo e o valor formado por eles, confirmando a validade da avaliação e, assim, ser um complemento na tarefa dos peritos (Diem, 2019).

Em relação à comunidade científica, este estudo contribui para a confirmação da análise preditiva como uma ferramenta importante para qualquer área da sociedade, incluindo a área imobiliária, dando a conhecer mais uma aplicação no mercado português dos AVM, isto é, uma ferramenta de apoio à decisão do perito imobiliário (Su *et al.*, 2021) e de outros *stakeholders*.

1.2. Problema, questão de investigação e objetivos

Os AVM têm sido, por um lado, bastante utilizados para avaliações de grandes carteiras de imóveis, por exemplo, na banca (Steurer *et al.*, 2021), mas, por outro, têm sido alvo de críticas por parte dos peritos avaliadores e das associações que os representam e supervisionam (Su *et al.*, 2021). De facto, existe por parte dos peritos uma resistência à mudança e ao reconhecimento de que estes modelos podem ser bastante úteis nos processos de avaliação. Perante este problema, surge a questão de investigação: Como prever os valores de venda de imóveis com modelos preditivos?

Para responder à questão, e tendo o contexto do mercado imobiliário da Área Metropolitana de Lisboa (AML), pretendem-se alcançar vários objetivos, nomeadamente, caracterizar os valores de venda de imóveis, caracterizar a relação entre os valores de venda dos imóveis e as suas características, criar modelo preditivos do valor de venda em função das características do imóvel e do mercado onde se insere e identificar perfis de imóveis em que o modelo acerta e, também, em que erra na previsão.

Para tal, recolhem-se e analisam-se os dados relacionados com imóveis que não sejam confidenciais, identificam-se os melhores algoritmos de *data mining* (DM) e as melhores métricas para os avaliar para alcançar o melhor modelo possível e determinam-se quais são os atributos com maior impacto na variável alvo, o valor do imóvel (Renigier-Bižozor *et al.*, 2019).

Este estudo pretende evidenciar que os AVM são uma ferramenta muito útil de apoio aos *stakeholders*, sendo eles peritos avaliadores, credores, compradores, vendedores, mediadores, investidores, bancos, Estado, ou outras entidades. Além disso, este modelo, em termos científicos, é inovador em Portugal, não havendo, ainda, muitos estudos que comprovem a sua utilidade, servindo esta investigação como uma das primeiras nesse aspeto a nível nacional.

1.3. Abordagem Metodológica

Esta investigação, para além da análise da literatura existente relacionada com o tema em investigação, AVM no setor imobiliário, recorre à metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

Esta metodologia consiste na conversão de dados em informação, através de DM (Chapman *et al.*, 1999), sendo a mais utilizada nestas questões, revelando-se a mais completa quando comparada com as restantes na análise de dados e na construção de modelos preditivos (Caetano, 2013; Laureano *et al.*, 2014). Consiste em seis etapas: Compreensão do Negócio (onde se definem os objetivos e se enquadra o problema a resolver), Compreensão dos Dados (recolha e análise de dados), Preparação dos Dados (procedimentos de preparação dos dados), Modelação (construção de modelos), Avaliação (análise da *performance* dos modelos) e Implementação (implementação do projeto).

Assim, para a análise empírica, é essencial recolher dados relacionados com imóveis. Neste estudo determina-se o valor dos imóveis, através da extração da informação presente na página de uma agência imobiliária, ao invés da utilização de uma base de dados com informação de avaliações imobiliárias realizadas por peritos, inacessíveis devido a questões legais de confidencialidade dos dados.

O objetivo de negócio diz respeito ao melhoramento da qualidade da estimativa dos valores dos imóveis, de modo a fornecer estimativas mais fiáveis do valor do imóvel aos clientes, peritos e outros *stakeholders*.

Para este trabalho foram selecionados os dados de apartamentos, destinados a venda, resultando em 4013 registos. Foram incluídos atributos relacionados com os imóveis, como, por exemplo, o número de quartos, a área, a existência de elevador. Também foram recolhidos dados sobre a população da freguesia e outras características da mesma como o número de praias e de estações de transporte. Além disso, recolheu-se, também, informação sobre o poder de compra dos habitantes do concelho.

Neste estudo, foram testadas várias técnicas preditivas: regressão linear, as árvores de decisão e redes neurais, sendo os resultados dos modelos interpretados e analisados de acordo com os objetivos.

1.4. Organização

A dissertação segue a estrutura tradicional encontrando-se organizada em vários capítulos, para além desta introdução.

No Capítulo 2 é apresentada a revisão da literatura (RL) relacionada com o tema da investigação. Através da revisão sistemática da literatura (RSL), são abordados conceitos de AVM, como são elaborados os modelos, as questões legais que impactam a sua utilização e, ainda, a sua aplicação a nível nacional e internacional. Além disso, abordam-se conceitos de avaliação, de métodos usados para determinar o valor de imóveis e de algoritmos de inteligência artificial.

O Capítulo 3 representa o processo experimental da investigação, contendo a metodologia científica utilizada, o CRISP-DM.

O Capítulo 4 é dedicado à análise e discussão dos resultados, sendo as conclusões, os contributos, as limitações e as pistas para investigações futuras apresentados no Capítulo 5.

2. Revisão da Literatura

Neste capítulo é possível encontrar conceitos de avaliação de imóveis e de AVM e como os vários estudos sobre o tema abordam a construção dos mesmos. Em primeiro lugar, procede-se à realização da RSL, que consiste na identificação, na recolha e na interpretação de todos os estudos relevantes que consigam responder à(s) questão(ões) de investigação (Kitchenham, 2004). Para tal, estabelece-se o Protocolo da RSL, que consiste no planeamento da metodologia a utilizar na mesma.

Após estes procedimentos, avaliam-se as abordagens teóricas e práticas da literatura recolhida, sendo que ainda se averigua o estado de utilização dos AVM no panorama internacional. Por fim, avalia-se a qualidade dos artigos recolhidos na RSL, mediante os critérios definidos. Além da bibliografia recolhida para a RSL, é considerada outra literatura relevante, indicada por especialistas no setor, que ajudam a responder a questões legais da utilização dos AVM.

2.1. Revisão Sistemática da Literatura sobre AVM

Para a elaboração da RSL, é necessário elaborar-se o protocolo da mesma, que descreve o processo de obtenção da literatura relevante para a investigação. São formuladas questões às quais a RSL terá de dar resposta. Depois, a estrutura segue a ordem das questões e, assim, aborda os seus conceitos.

2.1.1. Protocolo da Revisão Sistemática da Literatura

O protocolo tem como intuito a planificação de métodos a utilizar na RSL, reduzindo o possível enviesamento da pesquisa (Kitchenham, 2004). Este descreve os critérios que são usados para selecionar e extrair os estudos, como é que a sua qualidade é avaliada e que comparações devem ser feitas (Gates, 2002). Além disso, serve para que outros investigadores possam replicar e seguir o mesmo estudo.

A RSL desta investigação tem como objetivo contribuir para o conhecimento acerca do tema dos AVM, assim como da avaliação de imóveis, através do recurso a literatura sobre componentes teóricas e práticas do tema. Assim, a RSL pretende responder à seguinte questão: Como construir AVM, ou seja, modelos preditivos para o valor de avaliação de imóveis? Existem, ainda assim, três questões mais específicas, entre as quais: Como se caracterizam os AVM; Em que contexto se podem utilizar AVM?; e Como selecionar a melhor metodologia de DM.

A literatura é selecionada através da extração de conteúdo presente em bases de dados científicas. De uma forma geral, a *web of science* (<https://www.webofknowledge.com>) e a *scopus* (<https://www.scopus.com>) oferecem a melhor cobertura ao nível dos *journals*, artigos e citações, sendo estas duas as escolhidas (Norris & Oppenheim, 2007).

Para extrair os artigos das bases de dados, aplica-se uma *query*, validada por especialistas, académicos e do setor imobiliário, que incide na pesquisa de conceitos presentes em campos como o título, o *abstract* e as palavras-chave. Estes conceitos são sinónimos de AVM, de técnicas preditivas ou *machine learning* e de imóveis.

A *query* construída é: (“Automated valu*” or “AVM”) AND (“Machine Learning” or “Algorithm*” or “Artificial Intelligence” or “Techniques” or “AI” or “Predict*” or “Forecast*” or “Estimat*” or “Analytics” or “Data Science” or “Data Mining” or “Big Data”) AND (“Home” or “Hous*” or “Building*” or “Property” or “Apartment*” or “Residence*” or “Habitation” or “Domicile” or “Real Estate”).

Da introdução da *query* nas duas bases de dados consultadas obtiveram-se 245 artigos, sendo que para a extração e sumarização de resultados foram considerados dados como o título, o(s) autor(es), o *abstract*, o tipo de documento, a área de pesquisa, as palavras-chave, o idioma, o *journal* e o ano.

Na sequência deste processo, definiram-se os critérios de inclusão e exclusão na Tabela 2.1.

Tabela 2.1 - Critérios de inclusão e exclusão

Critérios de Inclusão	Critérios de Exclusão
Artigos científicos com <i>peer review</i> (teóricos ou práticos) sobre AVM ou avaliação de imóveis	<i>Papers</i> , livros, tutoriais, editoriais, entre outros tipos de documentos
Artigos dos últimos 10 anos	Artigos duplicados
Artigos com <i>download</i> disponível	Artigos que não estão em língua inglesa

Através da aplicação dos critérios, excluíram-se 211 dos 245 artigos extraídos, como consta na Figura 2.1 que sintetiza o processo de seleção dos artigos alvo da RSL. Desta forma, são considerados como elegíveis para a revisão 34 artigos, apresentados na Tabela 2.2.

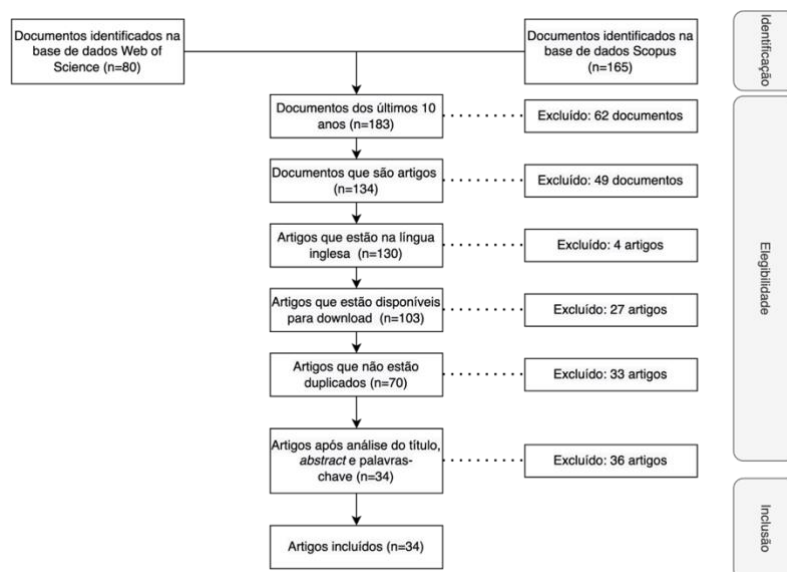


Figura 2.1 - Processo de seleção de artigos alvo da RSL

Tabela 2.2 - Artigos recolhidos

ID	Ano	Título	Autores	Journal	Quartil
1	2013	A reexamination of current hotel valuation techniques – which approach is more realistic?	Fu, J.; Sheel, A.; Lang, J.	Journal of Hospitality Financial Management	Q4
2	2014	Automated valuation modelling: a specification exercise	Schulz, R.; Wersing, M.; Werwatz, A.	Journal of Property Research	Q2
3	2017	The need to reference automatic valuation models to the valuation process	Epley, D.	Journal of Real Estate Literature	Q3
4	2017	An analysis of the use of mass appraisal methods for agricultural properties	Walt, K.; Boshoff, D.	Acta Structilia	Q4
5	2017	Reliable region predictions for automated valuation models	Bellotti, A.	Annals of Mathematics and Artificial Intelligence	Q3
6	2018	Road distance and travel time for an improved house price Kriging predictor	Crosby, H.; Damoulas, T.; Caton, A.; Davis, P.; Albuquerque, J.; Jarvis, S.	Geo-Spatial Information Science	Q1
7	2018	Vision-based real estate price estimation	Poursaeed, O.; Matera, T.; Belongie, S.	Machine Vision and Applications	Q3
8	2019	Automated Valuation Model based on fuzzy and rough set theory for real estate market with insufficient source data	Renigier-Biłozor, M.; Janowski, A.; d'Amato, M.	Land Use Policy	Q1
9	2019	Contextualized property market models vs. Generalized mass appraisals: An innovative approach	Morano, P.; Rosato, P.; Tajani, F.; Manganelli, B.; Liddo, F.	Sustainability	Q2
10	2019	Sensitivity analysis of machine learning models for the mass appraisal of real estate. Case study of residential units in Nicosia, Cyprus	Dimopoulos, T.; Bakas, N.	Remote Sensing	Q1
11	2020	Appraisal Accuracy and Automated Valuation Models in Rural Areas	Bogin, A.; Shui, J.	Journal of Real Estate Finance and Economics	Q1
12	2020	Combining Property Price Predictions from Repeat Sales and Spatially Enhanced Hedonic Regressions	Oust, A.; Hansen, S.; Pettrem, T.	Journal of Real Estate Finance and Economics	Q1
13	2020	Genetic Algorithm as Automated Valuation Model Component in Real Estate Investment Decisions System	Chmielewska, A.; Adamiczka, J.; Romanowski, M.	Real Estate Management and Valuation	Q3
14	2020	Who performs better? AVMs vs hedonic models	Valier, A.	Journal of Property Investment and Finance	Q2
15	2020	An innovative GIS-based territorial information tool for the evaluation of corporate properties: An application to the Italian context	Locurcio, M.; Morano, P.; Tajani, F.; Liddo, F.	Sustainability	Q2
16	2020	Automated valuation methods through the cost approach in a BIM and GIS integration framework for smart city appraisals	Arcuri, N.; De Ruggiero, M.; Salvo, F.; Zinno, R.	Sustainability	Q2
17	2020	Uncertainty in automated valuation models: Error-based versus model-based approaches	Krause, A.; Martin, A.; Fix, M.	Journal of Property Research	Q2
18	2021	A computer-assisted expert algorithm for real estate valuation in Spanish cities	Larraz, B.; Alfaro-Navarro, J.; Cano, E.; Alfaro-Cortes, E.; Garcia, N.; Gámez, M.	Environment and Planning B: Urban Analytics and City Science	Q1
19	2021	Automated Valuation Services: A case study for Aberdeen in Scotland	Schulz, R.; Wersing, M.	Journal of Property Research	Q2
20	2021	Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches	Doumpos, M.; Papastamos, D.; Andritsos, D.; Zopounidis, C.	Annals of Operations Research	Q1

Tabela 2.2 – Artigos recolhidos (continuação)

ID	Ano	Título	Autores	Journal	Quartil
21	2021	Hybrid predictions of the homogenous properties' market value with the use of ann	Anysz, H.; Podwórna, M.; Ibadov, N.; Lennerts, K.; Dikarev, K.	Archives of Civil Engineering	Q4
22	2021	Ignoring Spatial and Spatiotemporal Dependence in the Disturbances Can Make Black Swans Appear Grey	Pace, R.; Calabrese, R.	Journal of Real Estate Finance and Economics	Q1
23	2021	Normalized nonconformity measures for automated valuation models	Lim, Z.; Bellotti, A.	Expert Systems with Applications	Q1
24	2021	Genetic algorithm application for real estate market analysis in the uncertainty conditions	Renigier-Biłozor, M.; Chmielewska, A.; Walacik, M.; Janowski, A.; Lepkova, N.	Journal of Housing and the Built Environment	Q1
25	2021	Boosted Tree Ensembles for Artificial Intelligence Based Automated Valuation Models (AI-AVM)	Sing, T.; Yang, J.; Yu, S.	Journal of Real Estate Finance and Economics	Q1
26	2021	The Predictability of House Prices: "Human Against Machine"	Füss, R.; D'Silva, A.; Oust, A.; Birkeland, K.	International Real Estate Review	Q4
27	2021	Collateral misreporting in the residential mortgage-backed security market	Kruger, S.; Maturana, G.	Management Science	Q1
28	2021	Metrics for evaluating the performance of machine learning based automated valuation models	Steurer, M.; Hill, R.; Pfeifer, N.	Journal of Property Research	Q2
29	2021	A BIM and machine learning integration framework for automated property valuation	Su, T.; Li, H.; An, Y.	Journal of Building Engineering	Q1
30	2021	Property mass valuation on small markets	Gnat, S.	Land	Q2
31	2021	Real estate listings and their usefulness for hedonic regressions	Kolbe, J.; Schulz, R.; Wersing, M.; Werwatz, A.	Empirical Economics	Q2
32	2021	Using shrinkage for data-driven automated valuation model specification—a case study from Berlin	Hinrichs, N.; Kolbe, J.; Werwatz, A.	Journal of Property Research	Q2
33	2022	Real estate price estimation in French cities using geocoding and machine learning	Tchuente, D.; Nyawa, S.	Annals of Operations Research	Q1
34	2022	International acceptance of automated modern tools use must-have for sustainable real estate market development	Renigier-Biłozor, M.; Żróbek, S.; Walacik, M.; Borst, R.; Grover, R.; d'Amato, M.	Land Use Policy	Q1

Constata-se que existem 24 *journals*, sendo que o *Journal of Property Research* é o que tem mais artigos. Dos 34 artigos, 15 pertencem ao quartil 1 e 11 pertencem ao quartil 2, ou seja, são artigos provenientes de *journals* conceituados. O *Journal of Real Estate Finance and Economics* é o mais representado (quatro artigos) do primeiro quartil, sendo o escolhido no caso da publicação da investigação. Há 86 autores, sendo que os que mais se repetem são Rainer Schulz, Martin Wersing, Axel Wewatz e Małgorzata Renigier-Biłozor (três artigos). É de destacar, também, a evolução do estudo deste tema ao longo dos anos, onde metade dos artigos foram elaborados desde 2021.

De modo a perceber o conteúdo dos artigos e avaliar a sua pertinência para o tema, fez-se uma nuvem de palavras das palavras-chave dos artigos, como consta na Figura 2.2. Destacam-se as palavras *valuation*, *automated*, *model(s)*, *real*, *estate*, *market*, *machine* e *learning*, pelo que se conclui que os artigos estão alinhados com o tema da investigação e, assim, são úteis para elaborar o estado da arte.

Atualmente, as avaliações são realizadas por peritos avaliadores, que podem enfrentar pressões para atingir um valor maior do que o que seria estipulado de forma imparcial para garantir que a venda seja realizada (Bogin & Shui, 2020). Além disso, o processo tradicional tem sido questionado pela falta de transparência, imprecisão e ineficiência (Su *et al.*, 2021), assim como pelo custo elevado e pela demora no processo de obtenção do valor de um imóvel. As avaliações de imóveis são, na maioria, estimadas com base nos valores de imóveis semelhantes, levando à subjetividade (Bogin & Shui, 2020).

O setor foi sempre conservador, pelo que, novos métodos, por norma, tiveram muitas dificuldades na implementação (Arcuri *et al.*, 2020). Contudo, está a sofrer alterações devido à inovação tecnológica. Este processo foi acelerado pela pandemia COVID-19 e pelas consequências das restrições impostas como, por exemplo, a proibição de visita dos peritos ao imóvel (Renigier-Bifozor *et al.*, 2022).

Muitos peritos aprovam a existência de ferramentas para apoiar as suas funções, mas também têm medo da redução/perda de trabalho, por poderem ser substituídos. Além disso, temem não conseguir aprender a utilizar os métodos.

Tradicionalmente, existem várias formas de avaliação de imóveis, sendo as principais: o Método de Rendimento, o Método do Custo e o Método de Comparação de Mercado (Fu *et al.*, 2013). Estes métodos são definidos pelas *International Financial Reporting Standards* (IFRS) 13, normas reconhecidas a nível internacional, adotadas pela União Europeia (UE) em 2013.

O Método de Rendimento indica que o valor de um imóvel diz respeito ao seu retorno líquido, ou seja, ao valor atual dos benefícios futuros. Esses benefícios correspondem às receitas líquidas estimadas através de uma previsão de receitas e despesas, assim como de receitas antecipadas de uma venda no futuro. A técnica mais usada neste método é o *Cash Flow* Descontado porque permite ao comprador/investidor analisar a variação do fluxo financeiro ao longo de um determinado período (Fu *et al.*, 2013). Em termos algébricos, este método apresenta-se da seguinte forma:

$$V = r \times \frac{12}{t} \quad (1)$$

Em que V representa o valor do imóvel; r expressa o rendimento mensal bruto do imóvel; e t reflete a taxa de capitalização. Esta taxa é calculada pela relação entre os rendimentos anuais do imóvel e o valor do mesmo ou de imóveis idênticos.

O Método do Custo é calculado através da diferença entre os custos e as depreciações. Este método é mais indicado quando aplicado a imóveis recém-construídos (Fu *et al.*, 2013) e assenta no princípio da substituição: o comprador não está disposto a investir mais num imóvel do que o valor do custo de reconstrução, tendo em conta o estado de depreciação (Arcuri *et al.*, 2020).

A parte mais complexa deste método é a estimativa da depreciação acumulada, que pode ser feita pelo perito ou através de um manual de custos (Epley, 2017). Em termos algébricos, representa-se da seguinte forma:

$$V = v + Cc - D \quad (2)$$

Onde V corresponde ao valor do imóvel; v representa o valor do terreno; Cc expressa o custo de construção; e D reflete a depreciação (apenas no caso de os imóveis serem usados).

Por sua vez, o Método de Comparação de Mercado baseia-se na ideia de que os ativos com características semelhantes devem ter valores semelhantes. Este método requer conhecimento do mercado, ao nível dos valores das transações de imóveis. Contudo, se houver pouca informação disponível, torna-se difícil estimar um valor fiável, sendo que deve haver um foco especializado na recolha de informação de imóveis à venda ou já vendidos (Fu *et al.*, 2013).

Esta abordagem é recomendada pela *International Valuation Standards* (IVS) e pela *European Valuation Standards* (EVS) como o método principal para estimar o valor de um imóvel (Anysz *et al.*, 2021). A fórmula utilizada neste método é a seguinte:

$$V = \frac{v}{m^2} \times A \quad (3)$$

Onde V representa o valor do imóvel; v representa o valor de outros imóveis semelhantes; m^2 expressa os metros quadrados da área bruta; e A reflete a área do imóvel.

Apesar da avaliação de imóveis se basear na comparação entre imóveis com características semelhantes, ao longo da última década desenvolveram-se sistemas alternativos de avaliação capazes de aplicar estes métodos (Larraz *et al.*, 2021). Um exemplo destes sistemas de avaliação são os AVM.

Os AVM são programas informáticos, baseados em métodos matemáticos, que estimam o valor de um imóvel através da análise de fatores como a localização, as condições de mercado ou as características do imóvel, a partir de informações recolhidas. Este tipo de *software* tem sido desenvolvido, principalmente, pelos bancos e por empresas de *analytics* (Bellotti, 2017).

Uma das principais diferenças entre as avaliações tradicionais e as avaliações com AVM é que os peritos realizam inspeções da própria propriedade e os AVM dependem dos dados que são recolhidos (Renigier-Biłozor *et al.*, 2022).

O computador imita a perceção humana, através de modelos matemáticos que competem com humanos para cumprir tarefas como, por exemplo, a análise de um sistema e previsões de observações fora da amostra (Dimopoulos & Bakas, 2019).

Através da análise à literatura recolhida na RSL identificam-se algumas das principais vantagens e desvantagens dos AVM, como consta na Tabela 2.4.

Tabela 2.4 - Vantagens e desvantagens dos AVM

Vantagens	Desvantagens
<ul style="list-style-type: none"> • Aumento da transparência do setor; • Baixo custo; • Consistência; • Velocidade; • Objetividade; • Capacidade de abrangência; • Ferramenta útil para peritos avaliadores através da comparação de valores. 	<ul style="list-style-type: none"> • Forte dependência dos dados (quantidade e qualidade) para precisão dos modelos; • Contradição em relação aos métodos tradicionais; • A legalização da utilização ainda está em estágios iniciais em muitos países; • Dificil conhecimento da estrutura e funcionamento (considerado <i>black box</i>); • É difícil introduzir característica intangíveis como, por exemplo, a vista.

Fonte: Adaptado de Belotti (2017), Renigier-Biřozor *et al.* (2022), Schulz & Wersing (2021), Steurer *et al.* (2021) e Walt & Boshoff (2017).

2.1.3. Contextos de utilização dos AVM

Apesar dos AVM possuírem algumas limitações, é claro que estes são muito úteis para o setor imobiliário e podem significar um grande avanço na determinação do valor dos imóveis. Estes programas podem ser aplicados em diversos cenários e podem ser utilizados por diversos *stakeholders* para várias finalidades, como se observa na Tabela 2.5.

Tabela 2.5 - Aplicações dos AVM

Stakeholder	Objetivos
Perito avaliador	Complemento do método de avaliação tradicional (permite a comparação dos valores obtidos nos dois métodos); Revisão de avaliações
Comprador	Percepção do valor a pagar
Vendedor	Percepção do valor a receber
Credor	Definir o valor do crédito à habitação ou da hipoteca
Investidor	Percepção do valor a investir (é possível elaborar AVM para um imóvel individual ou para uma carteira de investimentos)
Seguradora	Avaliação de risco e assim definir o valor do seguro do imóvel
Estado	Determinar o valor do imposto que tem a receber pelo imóvel; Análise Financeira de despesa pública

Fonte: Adaptado de Belotti (2017), Schulz *et al.* (2014) e Steurer *et al.* (2021).

Da análise da tabela, conclui-se que estes programas se alargam a muitas áreas de interesse e podem aplicar-se em diversos contextos. Porém, o seu panorama atual não é igual a nível internacional.

Atualmente, os AVM já são utilizados em quase todo o mundo. Alguns países encontram-se, ainda, numa fase inicial da utilização deste *software*, enquanto, por exemplo, os Estados Unidos da América (EUA) já se encontram numa fase avançada de uso de AVM (Walt & Boshoff, 2017). Renigier-Biřozor *et al.* (2022) realizaram uma análise comparativa entre alguns países sobre a aplicação e a regulamentação dos AVM (Tabela 2.6).

Tabela 2.6 - Contexto atual dos AVM em alguns países

País	Utilização/Regulamentação
Itália	Aplicados segundo os regulamentos do Banco da Itália; Usados, principalmente, na gestão de garantias hipotecárias;
Estados Unidos da América	Usados segundo os padrões da <i>International Association of Assessing Officers</i> , da Prática de Avaliação Profissional e os regulamentos de cada Estado; Geralmente, são utilizados para a tributação da propriedade, segurança hipotecária e seguros;
Reino Unido	São utilizados, maioritariamente, por bancos; Alguns serviços construíram estes sistemas para que os avaliadores possam verificar as suas avaliações;
Polónia	Não há legislação aplicada, não podendo ser usados para relatórios de avaliação; Usados na área da consultoria de investimentos e análise de mercado;
Bulgária	Apenas uma minoria de peritos os utiliza; Não há, atualmente, regulamentos, sendo mais utilizados em bancos;
Bielorrússia	Não existe legislação; São usados por empresas imobiliárias para avaliação de habitações;
Eslováquia	Não há regulamentos nem aplicações;
Noruega	Existe legislação, onde os imóveis são registados e avaliados; No setor público são apenas utilizados por autoridades fiscais; Existem empresas privadas que construíram para análise e previsão de preços;
Geórgia	Não existem regulamentos; São utilizados para análise de mercado principalmente em moradias e apartamentos.

Fonte: Adaptado de Renigier-Biřozor *et al.* (2022: p. 8-9)

É possível comprovar que já existem alguns países que recorrem aos AVM para diversas finalidades, sendo que existe ainda muito espaço para esta ferramenta avançar no setor imobiliário. Ainda assim, um modelo eficaz num país não é necessariamente bem-sucedido noutro, exigindo-se adaptação às condições do mercado da região (Walt & Boshoff, 2017).

2.1.4. Construção de AVM: Metodologia

Os AVM são treinados com recurso a bases de dados de grandes dimensões para prever os valores de imóveis. Belotti (2017) refere que os AVM são desenvolvidos com variáveis relacionadas com os imóveis (por exemplo, a área, o número de quartos, o número de pisos ou se tem elevador), relacionadas com o local (por exemplo, proximidade a escolas ou aos transportes), relacionadas com os preços e com o contexto económico (por exemplo, o Produto Interno Bruto ou o poder de compra). O principal desafio na elaboração de um AVM é definir uma abordagem analítica viável que, através dos dados que usa, forneça avaliações confiáveis (Schulz *et al.*, 2014). Contudo, se o conjunto de dados disponíveis for pequeno, a eficácia dos métodos analíticos é reduzida. (Renigier-Biřozor *et al.*, 2019).

2.1.4.1. Compreensão e preparação dos dados

Uma das principais questões levantadas acerca da elaboração de AVM diz respeito à avaliação de qualidade, quer seja do local, quer seja do imóvel. Para lidar com isto, alguns estudos realizam vários procedimentos. Foi feito um levantamento das variáveis mais utilizadas, isto é, que se repetiram em, pelo menos, três estudos alvo da RSL, apresentado entre a Tabela 2.7 e a Tabela 2.9.

Tabela 2.7 - Variáveis relacionadas com venda dos imóveis

Variável	ID																																
	2	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19	20	21	22	23	24	25	26	28	29	30	31	32	33				
Valor	X	X		X	X		X	X	X		X	X			X	X		X		X	X		X	X	X		X	X	X				
Logaritmo do valor			X											X			X		X														
Preço/m ²						X						X					X					X				X							
Data		X		X										X								X	X	X					X				

Fonte: Elaboração Própria

Tabela 2.8 - Variáveis relacionadas com as características dos imóveis

Variável	ID																																
	2	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19	20	21	22	23	24	25	26	28	29	30	31	32	33				
Idade	X						X	X	X					X			X		X	X		X	X	X	X		X	X					
Ano de construção										X																							
Área útil	X					X	X	X	X	X		X		X	X	X	X	X		X		X	X	X	X	X	X	X	X				
Área do lote	X	X												X						X					X				X				
Tipo de imóvel	X		X	X							X					X							X				X		X				
Piso						X	X					X					X	X				X	X			X			X				
Estado de conservação	X		X	X			X	X	X	X		X		X			X										X	X					
Vista	X						X							X	X																		
Número de quartos					X				X	X	X			X	X	X		X		X	X		X			X			X				
Número de casas de banho					X		X		X		X			X	X	X				X	X					X							
Número de cozinhas																				X	X												
Número de despensas																						X											
Estacionamento																X	X						X										
Jardim											X					X	X																
Tipo de propriedade (proprietário ou arrendamento)												X			X					X		X		X									
Aquecimento																X	X	X									X						
Piscina																X										X		X					

Fonte: Elaboração Própria

Tabela 2.9 - Variáveis relacionadas com a localização dos imóveis

Variável	ID																													
	2	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19	20	21	22	23	24	25	26	28	29	30	31	32	33	
Acessos	X	X																										X		
Qualidade da localização	X		X	X	X		X		X		X			X	X		X				X		X	X	X	X	X	X		X
Distância a pontos de interesse			X				X															X	X							
Estações de transporte			X									X															X			

Fonte: Elaboração Própria

Não se consideraram os estudos 1, pois aborda a avaliação de hotéis através das receitas por quarto; e 27, que aborda o impacto da inflação em AVM, já que não é possível retirar conclusões nos atributos selecionados. Os restantes estudos, que não se encontram nas tabelas, não têm componente empírica.

A análise das tabelas permite concluir que existe muita diversidade de atributos no processo da obtenção do valor dos imóveis. No entanto, a maioria dos estudos tem em comum, além da variável dependente que é o valor do imóvel, ou uma sua transformação, as variáveis preditoras: área, estado, número de quartos, número de casas de banho, localização e tipo de imóvel. Alguns artigos recorrem a várias abordagens de análise exploratória como, por exemplo, a análise univariada de variáveis, aferindo a qualidade dos dados ao nível de estatísticas com recurso à média, ao desvio padrão, à mediana, entre outros. São também analisados fatores que causem enviesamento nos modelos como sejam os *outliers* e os *missings values*, e é elaborado um dicionário de dados (Morano *et al.*, 2019).

Após esta análise, procede-se ao tratamento da base de dados. Esta etapa é muito importante na realização de AVM. Por exemplo, os *outliers* podem causar grandes erros de avaliação. Alguns artigos removem os *outliers* e os *missing values*, pois sem esta eliminação a qualidade das avaliações pioram drasticamente (Schulz *et al.*, 2014).

Além disso, a maioria dos estudos consultados faz análises de correlação das variáveis de modo a analisar o comportamento das mesmas com as restantes e eliminar aquelas cujos valores de correlação são irrelevantes para o modelo. Por exemplo, concluiu-se que um modelo de regressão melhora através da eliminação de variáveis independentes que não têm influência significativa na variável dependente (Walt & Boshoff, 2017).

2.1.4.2. Algoritmos

Após o tratamento de dados, a maioria dos artigos selecionados aplicou modelos de regressão aos seus dados, pelo que será utilizado nesta investigação. Além disso, os algoritmos mais utilizados foram as árvores de decisão e as redes neuronais, como consta na Tabela 2.10.

Tabela 2.10 - Caracterização da componente empírica dos estudos

Componente Empírica						
ID	Âmbito avaliação	Nº de registos	Países da amostra	Período	Técnica(s) Preditiva(s)	Métricas usadas
1	Hotéis	59	EUA	1999-2009	N.E. ⁽¹⁾	Medidas descritivas
2	Habitação	19 553	Alemanha	2000-2011	Regressão Linear	MPE, MdPE, MAPE e MSE
4	Propriedades agrícolas	60	África do Sul	N.E.	Regressão Linear	R, R ² , R ² ajustado e medidas descritivas
5	Habitação	22 145	Inglaterra	2009-2014	Regressão Linear	Medidas descritivas
6	Habitação	3 669	Inglaterra	2016	Regressão Linear	R ² , MSE e MAPE
7	Habitação	149 000	EUA	N.R.	Regressão Linear	Medidas descritivas

Tabela 2.10 - Caracterização da componente empírica dos estudos (continuação)

Componente Empírica						
ID	Âmbito	Nº de registos	Países da amostra	Período	Técnica(s) Preditiva(s)	Métricas usadas
8	Habituação	42	Itália e Polónia	2016	Regressão Linear	R ² , MPE e medidas descritivas
9	Habituação	600	Itália	2016-2017	Regressão Linear	MSE, MAPE, MaxAPE, R ² , R ² ajustado e medidas descritivas
10	Habituação	4 261	Chipre	2008-2014	Regressão Linear e Árvores de Decisão	MAE, MSE, MAPE, MaxAPE e medidas descritivas
11	Habituação	20,1 milhões	EUA	2012-2016	Regressão Linear e Árvores de Decisão	R ² e MSE
12	Habituação	16 417	Noruega	2016-2017	Regressão Linear	MAPE e medidas descritivas
13	Habituação	2 578	Polónia	2018	Algoritmo Genético e Método de Hellwig	Medidas descritivas
15	Habituação	358	Itália	2004-2016	Regressão Linear	MaxAPE, MAPE e MSE
16	Habituação	N.E.	Itália	N.E.	N.E.	N.E.
17	Habituação	485 044	EUA	1999-2019	Modelo Linear e Não Linear	MdAPE, MdPE e medidas descritivas
18	Habituação	1,78 milhões	Espanha	2018	Inverso da Potência das Distâncias	MAPE, MSE e medidas descritivas
19	Habituação	12 032	Escócia	2017	Regressão Linear e Árvores de Decisão	MRE, MdRE, REV, MSE, MARE e RER
20	Habituação	36 529	Grécia	2012-2016	Regressão Linear	R ² , MAPE e MdAPE
21	Habituação	222	Polónia	2013-2014	Redes Neurais	MAPE
22	Habituação	62 266	EUA	2000	Regressão Linear	MAE
23	Habituação	2 930	EUA	2006-2010	Regressão Linear	Medidas descritivas
24	Habituação	3 247	Polónia	2017	Algoritmo Genético e Método de Hellwig	Medidas descritivas
25	Habituação	378 032	Singapura	1995-2019	Regressão e Árvores de Decisão	R ² , MAPE e MSE
26	Habituação	220 898	Noruega	1993-2018	Regressão Linear, <i>XGBoost</i> e Árvores de Decisão	MdAPE e MAPE
27	Empréstimos	11,86 milhões	EUA	2001-2007	Regressão Linear	R ² e Medidas descritivas
28	Habituação	18 957	Áustria	2015-2020	Regressão Linear, Árvores de Decisão e Redes Neurais	MPE, MAE, MAPE, COD, MSE, R ² , LMDPE e LRMSE
29	Habituação	18 597	EUA	2000-2017	Regressão Linear	MAPE, MAE e MSPE
30	Habituação	318	Polónia	2018	Regressão Linear e <i>XGBoost</i>	MSE e medidas descritivas
31	Habituação	161 924	Alemanha	2007-2015	Regressão Linear	MSE, MAE e Medidas descritivas
32	Habituação	190 000	Alemanha	1996-2013	Regressão Linear	MSE, MAPE e MdAPE
33	Habituação	N.E.	França	2015-2019	Rede Neurais, Árvores de Decisão, Regressão Linear, <i>Support Vector Machine</i>	MAE, MdAE, MSE, LMRSE e R ²

Notas: N.E.: Não existe/Não referido; MPE: Erro médio percentual; MdPE: Erro mediano percentual; MAPE: Erro médio absoluto percentual; MSE: Erro quadrático médio; R: Coeficiente de correlação; R²: Coeficiente de determinação; MaxAPE: Erro máximo absoluto percentual; MAE: Erro médio absoluto; MdAPE: Erro percentual absoluto mediano; MRE: Erro médio relativo; MdRE: Erro mediano relativo; VER: Variância relativa do erro; MARE: Erro médio absoluto relativo; RER: Taxa de erro relativa; COD: Coeficiente de Dispersão; LMDPE: Logaritmo do erro quadrático médio da previsão; LRMSE: Logaritmo do erro quadrático médio; MdAE: Erro mediano absoluto.

Fonte: Elaboração Própria

Não foram considerados os artigos que não tinham componente empírica e não há referência a metodologias de DM como, por exemplo, CRISP-DM, o *Sample, Explore, Modify, Model and Assess* (SEMMA) ou o *Knowledge Discovery in Databases* (KDD).

O algoritmo da regressão linear é uma ferramenta estatística que identifica relações entre a variável dependente e as independentes (Sing *et al.*, 2021), assumindo que o valor de mercado é ponderado sobre características (Poursaeed *et al.*, 2018), por exemplo, a existência de garagem, de ar condicionado, o número de metros quadrados de área útil, e a localização (Epley, 2017). Este algoritmo não lida bem com *outliers*, podendo enviesar o modelo (Poursaeed *et al.*, 2018).

As redes neuronais são algoritmos que identificam padrões na avaliação (Epley, 2017). Estes modelos são construídos para imitar o cérebro humano, através da repetição de estímulos, emparelhando dados de *input* e *output*. Têm como principal vantagem a *performance* ser superior a outros algoritmos, na maioria dos casos, no entanto existem limitações nesta técnica. A principal crítica é o facto de ser uma técnica *black box*, devido à sua estrutura interna que conecta os nós nos modelos não ser conhecida. Além disso, ao contrário da regressão, que identifica relações entre as variáveis predictoras e o valor do imóvel, os modelos deste algoritmo são opacos e é muito difícil interpretar os resultados (Sing *et al.*, 2021).

Por sua vez, as árvores de decisão são compostas por várias regras de divisão, começando no topo da árvore e indo até ao fundo. Esta técnica tenta replicar a forma como o ser humano toma decisões. Este algoritmo utiliza uma representação gráfica de fácil interpretação. No entanto uma pequena mudança nos dados pode ter muito impacto nas previsões (Sing *et al.*, 2021).

Deste modo, quanto maior for a capacidade que os dados têm em identificar as variáveis mais importantes para estimar os principais atributos do valor de um imóvel, maior é a qualidade das previsões dos modelos obtidos (Oust *et al.*, 2020).

2.1.4.3. Avaliação da qualidade dos AVM

Segundo Steurer *et al.* (2021), para avaliar a qualidade dos modelos, recorre-se, por exemplo, ao *cross validation*, nomeadamente o *k-fold*, que, para lidar com o sobreajustamento, dividem o conjunto de dados para ajustar o modelo e avaliar o desempenho.

Assim que se escolher o algoritmo a utilizar, são necessárias, também, métricas para avaliar a qualidade dos modelos (Steurer *et al.*, 2021).

Através da análise da Tabela 2.10, é possível perceber que os estudos apresentam várias métricas, como forma de avaliação aos modelos. Existem, então, muitas métricas para medir a qualidade dos modelos em problemas de regressão (onde a variável dependente é quantitativa).

As principais métricas utilizadas são o erro médio absoluto (MAE), o erro quadrático médio (MSE), o coeficiente de determinação (R^2) e o coeficiente de dispersão (COD), cujas fórmulas são, respetivamente:

$$MAE = \frac{1}{N} \sum_{n=1}^N |v_n - \hat{v}_n| \quad (4)$$

$$MSE = \frac{1}{N} \sum_{n=1}^N (v_n - \hat{v}_n)^2 \quad (5)$$

$$R^2 = 1 - \frac{\sum_{n=1}^N (v_n - \hat{v}_n)^2}{\sum_{n=1}^N (v_n - \bar{v})^2} \quad (6)$$

$$COD = \frac{1}{N} \sum_{n=1}^N \left| \left[\left(\frac{v_n}{\hat{v}_n} \right) / med \left(\frac{v}{\hat{v}} \right) \right] - 1 \right| \quad (7)$$

Onde os valores observados dos imóveis correspondem ao v_n , os valores previstos correspondem ao \hat{v}_n e a média aritmética dos valores observados é refletida pelo \bar{v} , onde o n varia entre 1 e N , que é o número total de observações (Steurer *et al.*, 2021). Em relação ao R^2 , o valor, nos estudos, varia entre 0,75 e 0,90 na maioria dos casos.

Além disso, na avaliação dos modelos, recorre-se à análise de sensibilidade para avaliar o impacto de cada variável independente na variável dependente, o preço do imóvel (Dimopoulos & Bakas, 2019). É através desta análise que, por exemplo, a interpretabilidade das redes neuronais aumenta, comparando o efeito de um preditor com a variável dependente. São obtidas curvas de sensibilidade para cada modelo, o que ilustra a complexidade dos mesmos (Dimopoulos & Bakas, 2019).

2.1.5. Avaliação da qualidade dos artigos da RSL

O processo de avaliação dos artigos, tal como definido no protocolo da RSL, tem por base pontuações atribuídas a cada critério. Assim, a avaliação dos artigos, nos diferentes critérios e respetivas dimensões de análise (questões de investigação), é apresentada na Tabela 2.11.

A avaliação resultou na atribuição máxima da pontuação a três artigos, o “An analysis of the use of mass appraisal methods for agricultural properties” (Walt & Boshoff, 2017), “Genetic Algorithm as Automated Valuation Model Component in Real Estate Investment Decisions System” (Renigier-Bižozor *et al.*, 2021) e “Uncertainty in automated valuation models: Error-based versus model-based approaches” (Krause *et al.*, 2020).

Além disso, o artigo que obteve menor pontuação foi o “Who performs better? AVMs vs hedonic models” (Valier, 2020), muito devido a não ter componente empírica no seu estudo. O critério com maior pontuação foi o Q4 (“Descreve minuciosamente cenários de utilização dos AVM?”), em todos os artigos foram descritos cenários de uso de AVM. O critério que obteve menor pontuação foi o Q7 (“A metodologia está bem definida e descrita?”), onde além de nenhum artigo ter adotado nenhuma metodologia de DM, foram poucos os que explicaram de forma detalhada a metodologia do estudo.

Por outro lado, a questão de investigação mais bem respondida pelos artigos é “Em que contexto se podem utilizar AVM?”, cuja média é 34 pontos por critério e a pior é “Como se caracterizam os AVM?” cuja média é 25,5 pontos por critério.

Tabela 2.11 - Avaliação dos artigos

ID	Caracterizar AVM			Aplicação de AVM	Selecionar a melhor metodologia de DM						Total
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
1	1	0,5	1	1	1	1	0,5	0	1	1	8
2	0,5	1	0	1	1	1	1	1	1	1	8,5
3	0,5	1	0	1	0	0	0	0	0	0	2,5
4	1	1	1	1	1	1	1	1	1	1	10
5	0	1	0	1	1	1	1	1	1	1	8
6	0,5	1	0	1	1	1	1	1	1	1	8,5
7	1	1	0	1	1	1	0	0,5	0	0	5,5
8	0	1	1	1	1	0,5	0,5	0,5	0,5	1	7
9	1	1	0	1	1	0,5	0	1	1	1	7,5
10	0	1	1	1	1	1	1	1	1	1	9
11	1	0,5	1	1	1	1	0,5	1	1	1	9
12	1	1	1	1	1	1	1	0	1	1	9
13	1	1	1	1	1	1	1	1	1	1	10
14	1	1	0	1	0	0	0	0	0	0	3
15	1	1	1	1	1	0,5	0	1	1	1	8,5
16	1	1	1	1	1	1	0	0	1	0	7
17	1	1	1	1	1	1	1	1	1	1	10
18	1	1	0,5	1	1	1	1	1	1	1	9,5
19	1	1	0	1	1	1	0	1	1	1	8
20	1	1	0	1	1	1	0	1	1	1	8
21	1	0,5	0	1	1	0,5	0	1	1	1	7
22	1	1	1	1	1	1	0	0	1	1	8
23	1	1	0	1	1	1	0,5	1	1	1	8,5
24	1	1	0	1	1	1	0,5	1	1	0,5	8
25	1	1	1	1	1	1	0	1	1	1	10
26	1	1	1	1	1	1	0,5	1	1	1	9,5
27	1	1	1	1	1	1	0	0	0	1	7
28	0,5	1	0	1	1	1	1	1	1	1	8,5
29	1	0,5	1	1	1	1	1	1	1	1	9,5
30	1	1	0	1	1	1	0	1	1	1	8
31	1	0,5	0	1	1	1	0,5	1	1	1	8
32	0,5	1	0	1	1	1	0,5	1	1	1	8
33	1	1	0	1	1	1	0	1	1	1	8
34	1	1	1	1	0	0	0	0	0	0	4

2.2. Contexto nacional e questões legais associadas

No panorama internacional, há três organizações que representam os profissionais do setor imobiliário de maior relevância quanto à avaliação de imóveis, sendo elas a *The European Group of Valuers' Associations* (TEGoVA), a *Royal Institution of Chartered Surveyors* (RICS) e a *International Valuation Standards Committee* (IVSC). De acordo com o Banco de Portugal *et al.* (2013), estes organismos têm como função a emissão de várias normas internacionais de avaliação, como a EVS, o *RedBook* e a IVS.

Contudo, nenhuma norma prevalece sobre as restantes, em termos internacionais, como referência para os peritos em avaliação de imóveis (Banco de Portugal *et al.*, 2013).

Adicionalmente, na *European Valuation Information Paper 7* (EVIP 7), da EVS 2020, são estabelecidas as condições de utilização dos AVM, onde pode ser usado por avaliadores qualificados para análise do mercado (TEGoVA, 2020). Porém, o valor obtido não pode ser considerado um valor de avaliação nem um valor de mercado. Deste modo, os AVM constituem uma ferramenta de apoio à avaliação, não substituindo o perito (Vaz & Anjos, 2022).

Por sua vez, o IVSC defende que os utilizadores dos AVM ainda não compreendem a finalidade, o funcionamento e as limitações dos mesmos e que não estão alinhados com as exigências das IVS, tendo o perito que se responsabilizar pelo *output* do AVM (Vaz & Anjos, 2022).

A legislação de Portugal e da UE, permite que as instituições de crédito utilizem métodos estatísticos considerados adequados. Todavia, em Portugal, esta é uma matéria que está numa fase inicial, onde alguns projetos se encontram em estudo para colmatar lacunas existentes na análise da evolução dos mercados imobiliários (Banco de Portugal *et al.*, 2013).

De acordo com o Banco de Portugal (2006), as instituições de crédito podem recorrer a métodos estatísticos para a verificação de hipoteca ou revisão da avaliação, sendo que deve ser elaborado um relatório que descreve os métodos estatísticos e a utilização. Além disso, carece de um parecer, elaborado por um perito com independência, acerca da adequação dos métodos.

Assim, em Portugal, os AVM ainda não podem ser utilizados de forma isolada para a obtenção de um valor de avaliação. Podem, no entanto, ser utilizados como uma ferramenta de apoio ao avaliador, por exemplo, para análise de mercado ou para análise comparativa com os valores determinados por ele.

3. Metodologia

O trabalho prático nesta dissertação corresponde à construção de um AVM aplicado a imóveis da AML. Devido às elevadas restrições na privacidade de dados, quer por parte dos bancos, quer por parte de empresas peritas em avaliação de imóveis, não é possível utilizar um *dataset* cujo valor do imóvel diz respeito ao valor de avaliação do imóvel. Como tal, a solução encontrada que mais se adequa é a utilização do valor que o vendedor pede pelo imóvel, o *asking price*. A previsão deste valor é muito importante para *stakeholders*, como vendedor e comprador.

Para realizar este processo, utiliza-se a metodologia CRISP-DM que consiste, essencialmente, na transformação de dados em conhecimento/informação útil para as partes interessadas, através do DM (Chapman *et al.*, 1999). Esta metodologia é a mais utilizada em problemas de analítica, revelando-se a mais completa, focada no negócio, quando comparada com as restantes metodologias a aplicar a problemas de análise de dados e na construção de modelos preditivos (Caetano, 2013; Laureano *et al.*, 2014). É constituída por seis etapas, Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelação, Avaliação e Implementação, e são iterativas, ou seja, em qualquer situação é possível recuar para as etapas anteriores. A utilização do CRISP-DM é um dos fatores inovadores desta investigação, pois na literatura recolhida nenhum estudo recorreu ao CRISP-DM, podendo, assim, ser um avanço nesta temática.

3.1. Compreensão do Negócio

A primeira etapa diz respeito à compreensão do negócio, ao problema que tem de ser resolvido e aos objetivos estipulados no projeto, sendo este um dos aspetos mais importantes nesta fase (Schröer *et al.*, 2021). É pela definição de objetivos e métricas nesta fase que, depois, se avalia o cumprimento dos mesmos e, assim, perceber se este processo corresponde às expectativas delineadas previamente.

Esta fase consiste no enquadramento do mercado imobiliário, realizado nas secções anteriores, assim como na definição de objetivos do projeto, sendo o principal a previsão, bem-sucedida, do valor dos imóveis, comprovando-se a eficácia e a capacidade preditiva e, assim, uma previsão o mais correta possível para todos os *stakeholders*.

Assim, de modo a resolver o problema desta investigação, estipulam-se objetivos, de negócio e analítico. O objetivo de negócio é o aumento da qualidade da estimativa dos valores dos imóveis e, assim, melhorar as estimativas do valor do imóvel fornecidas aos clientes e o objetivo analítico é a criação de um modelo preditivo com boa qualidade na previsão do valor dos imóveis.

Ainda nesta etapa, foram consultados vários especialistas no setor imobiliário. Tendo em conta o objetivo de negócio definido, os especialistas aconselharam a não se estudar simultaneamente o segmento de imóveis de luxo com os imóveis comuns, nem a estudar em simultâneo apartamentos e

moradias, pelo que se definiu que o estudo se foca em apartamentos em que o valor (*asking price*) não excede um milhão de euros.

Os principais riscos deste projeto são a extração dos dados, pois podem ocorrer erros que afetem a qualidade dos dados e a privacidade dos dados, que pode resultar na indisponibilização dos mesmos, ou de alguns atributos, e, assim, prejudicar a análise e eficácia das previsões.

No que diz respeito às metas do projeto, define-se um valor mínimo para o R^2 de 0,8 e deve escolher-se o melhor modelo que satisfaz este critério e, ao mesmo tempo, possa ajudar à explicação da formação do valor estimado, atendendo ao prazo do projeto, que não excede um ano.

3.2. Compreensão dos Dados

Esta etapa consiste na recolha, identificação, exploração e aferição da qualidade dos dados necessários para a realização do modelo (Schröer *et al.*, 2021), o que corresponde à recolha dos dados da avaliação de imóveis, organizando-os de forma estruturada. Após a estruturação dos dados, cria-se um dicionário de dados, onde consta o nome da variável, o tipo e a sua descrição, para melhor entendimento.

Assim, procedeu-se à extração dos dados relativos aos imóveis, extraídos da *webpage*, com recurso à plataforma *ParseHub*, da agência imobiliária Remax, sobre apartamentos destinados a venda na AML, em maio de 2022.

Após esta extração, apesar de não existirem atributos relativos ao contexto da população e dos alojamentos nos estudos consultados, vários especialistas no setor aconselharam a extração de informação relativa aos mesmos da zona do imóvel, sendo que os dados mais adequados dizem respeito aos resultados provisórios dos Censos de 2021, relativos à população e a alojamentos das freguesias, extraídos do *website* do Instituto Nacional de Estatística (INE).

Adicionalmente, na literatura existem vários estudos que consideraram relevantes as distâncias a pontos de interesse como escolas, serviços de saúde, serviços de segurança, entre outros como, por exemplo, Bellotti (2017) ou Morano *et al* (2019). Portanto, recolheram-se dados sobre as escolas existentes, à data, na freguesia de cada imóvel através de informação presente na Rede Escolar da GesEdu, assim como dos hospitais e farmácias através de dados provenientes do Infarmed. Foi, ainda, recolhida a informação acerca da presença de estações de metropolitano e de comboio em cada freguesia, através das plataformas dos Comboios de Portugal, do Metropolitano de Lisboa e do Metro Transportes do Sul, à data. Ainda foi extraída informação quanto ao número de estabelecimentos da Polícia de Segurança Pública (PSP) e da Guarda Nacional Republicana (GNR), através das suas plataformas, assim como das praias de cada freguesia com recurso aos *websites* das autarquias e da *PlayOcean*. Por fim, os especialistas aconselharam a recolher atributos que medissem o contexto

económico à data, recolhendo, através da Pordata, os dados mais recentes, 2019, sobre o poder de compra e, através da Confidencial Imobiliário, recolheram-se dados sobre valores de referência para venda e arrendamento, em 2021. Na Tabela 3.1, encontram-se todas as fontes de dados e o seu ano de referência.

Tabela 3.1 - Fontes de dados

Tipo de dados	Fonte de dados	Ano de referência
Imóvel	Remax	2022
População e alojamentos	INE	2021
Escolas	GesEdu	2022
Hospitais e farmácias	Infarmed	2022
Estações de transportes públicos	CP, Metropolitano de Lisboa e Metro Transportes do Sul	2022
Estabelecimentos policiais	PSP e GNR	2022
Praias	Autarquias e <i>PlayOcean</i>	2022
Poder de compra	Pordata	2019
Valores de venda e arrendamento	Confidencial Imobiliário	2021

Notas: INE: Instituto Nacional de Estatística; CP: Comboios de Portugal; PSP: Polícia de Segurança Pública; e GNR: Guarda Nacional Republicana.

Deste modo, o *dataset* contém 43 variáveis com 5268 registos, em formato csv¹. No entanto, foram excluídos imóveis com erros, como sejam os imóveis com vários quartos e casas de banho com áreas reduzidas (menos que 50 m²), imóveis sem casas de banho, imóveis com várias casas de banho e nenhum quarto, imóveis duplicados ou imóveis cujo valor é muito reduzido para as características da casa. Além disso, foram removidos registos cujo número de casas de banho era negativo, cujo ano de construção é muito antigo (significando um erro), por exemplo, ano 1000, e superior a 2024 (onde os registos eram 4179 e 20202) e cujo número de lugares de estacionamento apresentava o registo “sim”. Para lidar com *missing values*, nas variáveis quantitativas, como estes representam uma percentagem muito reduzida em qualquer variável, foram eliminados. Destas alterações resultou um *dataset* com 4103 registos (imóveis).

3.2.1. Análise Exploratória

Nesta fase, realiza-se uma análise descritiva de todas as variáveis, quantitativas e qualitativas, no SPSS *Statistics* (v. 28), para aferir a qualidade dos dados, que consta na Tabela 3.2 a Tabela 3.5 e na **Error! Reference source not found.**

Como se observa na Tabela 3.2, os imóveis têm entre 25 m² e 427 m² de área útil. Todos os imóveis têm, pelo menos, uma casa de banho, sendo que metade dos imóveis têm um valor pedido, no máximo, de 269 mil euros. Um imóvel custa, em média, cerca de 335 mil euros e tem cerca de 102 m² de área útil. Existe muita variedade de apartamentos, havendo imóveis que têm entre 0 e 6 quartos.

¹ O dicionário de dados está presente no Anexo A: Dicionário de dados.

Tabela 3.2 - Estatísticas descritivas de variáveis quantitativas relativas aos imóveis

Variável (unidade)	Média	Desvio Padrão	Mínimo	Máximo	Percentis		
					25	50	75
Valor (x1000€) (€)	335,82	212,10	49,00	1000,00	170,00	269,00	445,00
Área (útil) (m ²)	101,72	43,79	25,00	427,00	72,00	92,00	121,00
Casas de Banho	1,71	0,78	1	6	1	2	2
Quartos	2,35	0,93	0	6	2	2	3

Nota: n=4013 imóveis.

Em relação às variáveis qualitativas, na Tabela 3.3, verifica-se que a maior parte dos apartamentos foram construídos no presente ano de 2022 (7,8%), havendo 75 anos de construção diferentes, o que revela que muitos imóveis no mercado já foram reabilitados (que se utiliza apenas na etapa seguinte e não entrando nos modelos). Também se constata que a maior parte dos apartamentos se situam no 1º andar (19,9%) e no concelho de Lisboa (80,3%). No entanto, a freguesia mais comum é Algés, Linda-a-Velha e Cruz Quebrada-Dafundo, concelho de Oeiras, com apenas 3,3% dos imóveis da amostra. Além disso, a maioria dos apartamentos não tem lugar de estacionamento associado (59,9%) e os seus edifícios têm elevador (60,3%).

Tabela 3.3 - Estatísticas descritivas das variáveis qualitativas

Variável	Categoria mais frequente		Categoria menos frequente		Número de categorias
	Categoria	%	Categoria	%	
Ano de Construção	2022	7,80	1952	0,05	75
Piso	1	19,90	9	1,50	12
Distrito	Lisboa	79,60	Setúbal	20,40	2
Concelho	Lisboa	30,50	Alcochete	0,90	18
Freguesia	Algés, Linda-a-Velha e Cruz Quebrada-Dafundo	3,50	Atalaia e Alto Estanqueiro-Jardia; Igreja Nova e Cheleiros; Sado; e São João das Lampas e Terrugem	0,02	109
Eficiência Energética	D	40,10	G	0,20	9
Elevador	Sim	60,30	Não	39,70	2
Número de Lugares de Estacionamento	Não	59,90	3 ou mais	2,80	4

Nota: n=4103 imóveis.

No que diz respeito à análise do valor dos imóveis por município (Figura 3.1) comprova-se que os concelhos a norte do rio Tejo (Amadora, Cascais, Lisboa, Loures, Mafra, Odivelas, Oeiras, Sintra e Vila Franca de Xira) têm, em média, valor mais elevado do que os restantes. Lisboa é o concelho que tem, em média, imóveis mais valorizados (477316 euros) e Moita o concelho que tem menor valor (115945 euros). Pelo contrário, na região norte, Sintra é o concelho menos valioso (192779 euros) e Alcochete é o concelho mais valioso na região sul (404076 euros).

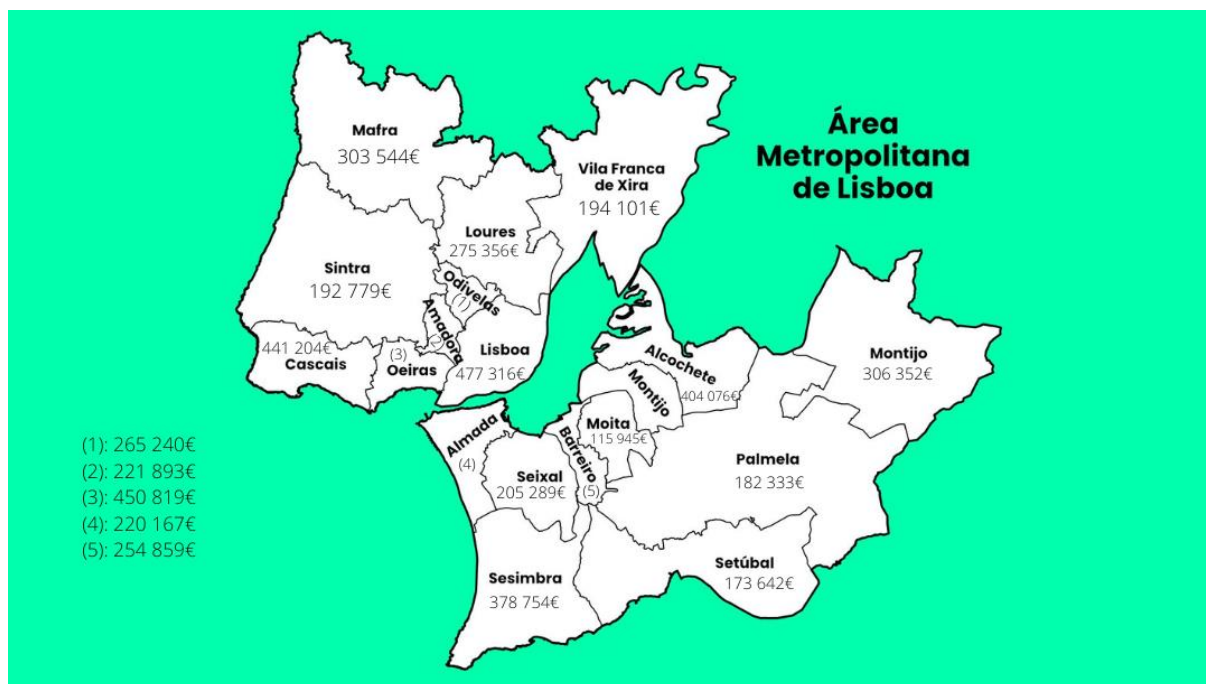


Figura 3.1 - Valor (média) dos imóveis por concelho

Continuando a análise da macro localização (Tabela 3.4) destaca-se a diferença entre freguesias, onde existem umas com cerca de 2500 habitantes e outras com cerca de 68600, umas com cerca de 25% de decréscimo da população e outras com acréscimo de cerca de 32%, sendo a média igual a 0,37%, entre 2011 e 2021. É de notar a falta de jovens nas freguesias, pois os jovens entre os 15 e os 24 anos representam a menor fatia das freguesias, em média. Além disso, a maior parte da população nas freguesias, em média, é aquela que não detém qualquer nível de escolaridade.

Verifica-se que existem, em média, cerca de 18190 alojamentos por freguesia sendo que o crescimento se encontra estagnado (média de -0,18%). Além disso, o maior decréscimo (-28,4 %) é, praticamente, o dobro, em valor absoluto, do maior crescimento (14,5%). Um cenário possível de análise a estes números seria que uma zona que apresente crescimento de alojamentos poderá significar investimento na região, o que a valoriza e, consequentemente, valoriza o imóvel.

Observa-se, também, que, em média, as freguesias têm uma área de cerca de 14 km², apesar de metade dos imóveis pertencerem a freguesias que não excedem os 7,18 km². Todas as freguesias têm, pelo menos, uma escola pública, componente que pode ponderar na decisão de compra de um imóvel. No entanto, existem freguesias que não têm hospitais nem farmácias. Existem, em média, cerca de duas estações de transporte e uma praia por freguesia. Contudo, destaca-se o facto de haver freguesias sem nenhum posto policial. Os fatores da segurança e saúde também podem pesar no processo de aquisição de um imóvel e não apenas as características do mesmo.

Tabela 3.4 - Estatísticas descritivas de variáveis relativas a características da freguesia do imóvel

Variável	Média	Desvio Padrão	Mínimo	Máximo	Percentis		
					25	50	75
População da Freguesia	34716,48	15713,97	2571	68654	21262	35482	46532
Varição da População da Freguesia (%)	0,37	6,70	-25,90	32,20	-1,30	0,50	3,60
População até aos 14 anos	4873,32	2398,90	344	10505	3047	4699	6746
População entre os 15 e os 24 anos	3732,11	1817,50	283	8376	2355	3721	5039
População entre os 25 e os 64 anos	18561,50	8386,30	1377	38483	11283	18793	24490
População a partir dos 65 anos	7552,44	3598,78	372	16627	4735	7567	10463
População Sem Escolaridade	13142,85	6653,01	914	28686	7894	13002	18646
População com o 1º Ciclo Básico	5093,68	2502,15	355	9629	2877	5190	7056
População com o 2º Ciclo Básico	2853,65	1487,19	232	6635	1586	2806	3958
População com o 3º Ciclo Básico	5220,05	2744,30	327	12539	3052	4968	7333
População com o Secundário e o Pós Secundário	8229,87	4328,75	576	19782	4484	7262	11452
População com o Ensino Superior	8751,99	5295,83	369	22723	5059	7306	11941
Alojamentos da Freguesia	18190,99	7563,05	1082	38739	12063	18133	24100
Varição dos Alojamentos da Freguesia (%)	-0,18	5,02	-28,40	14,50	-0,80	0,20	1,90
Área da Freguesia (km ²)	13,99	22,85	1,49	212,86	3,09	7,18	16,74
Escolas da Freguesia	23,60	12,23	1	68	15	24	30
Escolas da Freguesia	23,60	12,23	1	68	15	24	30
Escolas Privadas da Freguesia	12,99	8,20	0	46	7	13	17
Escolas Públicas da Freguesia	10,59	5,15	1	22	7	11	14
Hospitais da Freguesia	0,54	0,82	0	4	0	0	1
Farmácias da Freguesia	9,97	4,99	0	21	7	10	14
Estações de Transportes da Freguesia	1,91	1,86	0	9	0	2	3
Estações de Metro da Freguesia	0,95	1,74	0	9	0	0	1
Estações de Comboio da Freguesia	0,97	1,15	0	5	0	1	1
Estabelecimentos Policiais da Freguesia	1,80	1,55	0	7	1	1	2
Estabelecimentos da PSP da Freguesia	1,56	1,54	0	7	0	1	2
Estabelecimentos da GNR da Freguesia	0,23	0,43	0	2	0	0	0
Praias da Freguesia	0,60	2,20	0	20	0	0	0

Nota: n=4103 imóveis.

Analisando a Tabela 3.5, verifica-se que a média do peso do poder de compra é cerca de 4,5%, em que o máximo é 10,8%, ou seja, este município representa 10,8% do poder de compra nacional e a média do poder de compra *per capita* é 136,16%, o que significa que nestes municípios existe maior riqueza do que a média nacional. Possivelmente, concelhos com mais riqueza, terão indicadores de emprego mais elevados, com maior poder de compra, possibilitando a aquisição de um imóvel.

Tabela 3.5 - Estatísticas descritivas de variáveis relativas a características do concelho do imóvel

Variável	Média	Desvio Padrão	Mínimo	Máximo	Percentis		
					25	50	75
Peso do Poder de Compra (%)	4,48	3,86	0,23	10,18	1,49	2,64	10,18
Poder de Compra per capita (%)	136,16	48,65	83,40	205,60	97,40	108,00	205,60
Valor de Oferta de Venda por m ² (€)	3586,88	1724,97	1275	5898	2178,00	2506	5898
Valor de Renda Pedida por m ² (€)	11,24	2,42	7,30	14,20	9,20	10,50	14,20

Nota: n=4103 imóveis.

Para verificar a existência de *outliers* recorre-se ao critério mais utilizado na literatura para a sua análise. Consideram-se *outliers* os valores que se afastam, para cima ou para baixo, mais do que três desvios padrão (DP) da média (Jacinto, 2019). Para complementar a análise também se recorre ao

boxplot, (identifica os *outliers* pelo afastamento à mediana). Porém, o primeiro critério foi prioritário na análise. Além disso, avalia-se a relação entre as variáveis preditoras e o valor do imóvel, através da correlação de Pearson, quando a preditora é quantitativa, ou de Spearman, quando é qualitativa ordinal. Os valores da correlação variam entre 1 e -1 e, segundo Cohen (1992), a intensidade da relação é fraca entre $\pm 0,10$ e $\pm 0,29$, moderada entre $\pm 0,30$ e $\pm 0,49$ e forte a partir de $\pm 0,5$. No caso das preditoras qualitativas nominais, a associação é avaliada com o Eta, que mede a intensidade na mesma escala, de zero (ausência de relação) a um (relação perfeita), tal como se apresenta na Tabela 3.6.

Tabela 3.6 - Associação entre variáveis preditoras e o valor do imóvel

Variável preditora	Medida de associação	Coefficiente de associação
Área do Imóvel	Pearson	0,607*
Número de Casas de Banho	Pearson	0,595*
Número de Quartos	Pearson	0,324*
Ano de Construção	Spearman	0,469*
Piso	Spearman	0,256*
Distrito	Eta	0,242
Concelho	Eta	0,588
Freguesia	Eta	0,685
Eficiência Energética	Spearman	0,363*
Elevador	Eta	0,423
Lugares de Estacionamento	Eta	0,541
População da Freguesia	Pearson	-0,164*
Crescimento da População da Freguesia	Pearson	-0,079*
População até aos 14 anos	Pearson	-0,200*
População entre os 15 e os 24 anos	Pearson	-0,208*
População entre os 25 e os 64 anos	Pearson	-0,185*
População a partir dos 65 anos	Pearson	-0,047*
População Sem Escolaridade	Pearson	-0,361*
População com o 1º Ciclo	Pearson	-0,365*
População com o 2º Ciclo	Pearson	-0,365*
População com o 3º Ciclo	Pearson	-0,341*
População com o Secundário e o Pós Secundário	Pearson	-0,291*
População com o Ensino Superior	Pearson	0,294*
Alojamentos da Freguesia	Pearson	-0,067*
Crescimento dos Alojamentos da Freguesia	Pearson	-0,126*
Área da Freguesia	Pearson	-0,150*
Escolas da Freguesia	Pearson	0,025
Escolas Privadas da Freguesia	Pearson	0,159*
Escolas Públicas da Freguesia	Pearson	-0,195*
Hospitais da Freguesia	Pearson	0,304*
Farmácias da Freguesia	Pearson	0,304*
Estações de Transportes da Freguesia	Pearson	0,293*
Estações de Metro da Freguesia	Pearson	0,264*
Estações de Comboio da Freguesia	Pearson	0,076*
Estabelecimentos Policiais da Freguesia	Pearson	0,181*
Estabelecimentos da PSP da Freguesia	Pearson	0,213*
Estabelecimentos da GNR da Freguesia	Pearson	-0,105*
Número de Praias da Freguesia	Pearson	0,109*
Peso do Poder de Compra	Pearson	0,440*
Poder de Compra per capita	Pearson	0,517*
Valor de Oferta de Venda por m ²	Pearson	0,552*
Valor de Renda Pedida por m ²	Pearson	0,530*

Notas: n=4103 imóveis; *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Verifica-se que, na maioria dos casos, as correlações são fracas ou muito fracas (menor que 0,30, em valor absoluto). Contudo, algumas variáveis preditoras apresentam relações moderadas ou fortes. Destas, destacam-se a freguesia (0,685), não se sabendo o sentido da relação pois o eta não o dá, a área do imóvel (0,607) e o número de casas de banho (0,595) e, com uma relação negativa a população com o 1ª ciclo e com o 2ª ciclo (-0,365 em ambos) e a população sem escolaridade (-0,361).

Além disso, a análise das correlações entre variáveis preditoras quantitativas² permite destacar correlações muito fortes entre a população da freguesia e as restantes variáveis da população (exceto a da população com o ensino superior), assim como com a variável dos alojamentos, pelo que se considera, para os modelos, apenas a variável da população da freguesia, eliminando as demais. Por sua vez, as três variáveis relacionadas com as escolas também se relacionam fortemente (0,949 com as escolas privadas e 0,862 com as escolas públicas), decidindo manter apenas a variável do número de escolas na freguesia. Também as variáveis dos estabelecimentos policiais e dos estabelecimentos da PSP apresentam uma forte correlação (0,961), mantendo-se apenas a primeira. Por fim, as quatro variáveis dos indicadores económicos também se correlacionam fortemente (com valores superiores a 0,880), o que faz com que permaneça apenas a do poder de compra *per capita*.

3.3. Preparação dos Dados

A terceira etapa diz respeito à preparação dos dados, ou seja, às tarefas de preparação, de “limpeza” e de aplicação de técnicas de análise da base de dados inicial de modo a ajustá-la para a construção de um *dataset* final preparado para a etapa seguinte (Chapman *et al.*, 1999).

Quando identificados *outliers*, ao invés de se alterarem as variáveis originais, optou-se por se criar variáveis. Para não terem tanto impacto nos resultados, estes *outliers* são substituídos pela média mais ou menos três desvios padrão, em que começam por “S_O_” antes do nome das variáveis.

Por sua vez, são criadas variáveis tendo por base as existentes que podem ser muito importantes na explicação da variável alvo, sendo, por exemplo, rácios como percentagem de um estrato da população, quartos por metro quadrado, logaritmo do valor do imóvel, a idade do imóvel.

São também feitas recodificações a variáveis existentes, dando origem a novas variáveis. Em variáveis quantitativas são realizadas recodificações como, por exemplo, nas variáveis de crescimento em vez de devolver o valor, devolver o tipo de crescimento (nulo, negativo ou positivo) ou criar atributos que devolvam a existência ou não de hospitais, estabelecimentos policiais, entre outros, ao invés do número. Estas alterações podem dar maior fiabilidade na interpretação de resultados. No que diz respeito às variáveis qualitativas, são efetuadas recodificações em apenas duas, a Eficiência Energética e os Lugares de Estacionamento. Na primeira, existiam muitos escalões, pelo que, nos casos

² A tabela de correlações está presente no Anexo B: Correlação de variáveis.

em que a eficiência era “A+” e “A”, passaram a ser todos “A”, assim como para o “B” e o “B-“. Além disso, decidi criar-se um escalão em que estivessem compreendidos os últimos dois níveis. Na segunda, os registos que contêm “não” são alterados para “0”, pois significa que não existem lugares adjacentes ao imóvel. Contudo, para variáveis qualitativas, foram criadas variáveis *dummy* para poderem ser consideradas em modelos como a regressão linear. As variáveis encontram-se expostas na Tabela 3.7 e as estatísticas descritivas na Tabela 3.8 e na Tabela 3.9.

Tabela 3.7 - Dicionário de dados e procedimento das novas variáveis relativas ao imóvel

Variável	Definição	Procedimento (Tipo)
Log_10_Valor	Logaritmo do Valor	Valor: Logaritmo de base 10 (QC)
WC_por_m2	Número de casas de banho por m ²	Casas_de_Banho/ Area_Imovel (QC)
Quartos_por_m2	Número de quartos por m ²	Quartos/ Area_Imovel (QC)
Idade_Imovel	Idade do imóvel	2022-Ano_de_Construcao (QD)
Nova_Eficiencia_Energetica	Eficiência Energética do imóvel	Criação de escalões da variável Eficiencia_Energetica (QO)
Novo_Lug_Estacionamento	Número de lugares de estacionamento do imóvel	Alteração de “não” para “0” da variável Lug_Estacionamento (QO)
Tem_lugares_estacionamento	Existência de lugares de estacionamento	Variável com “sim” ou “não” (QN)
Percentagem_Populacao_ens_sup	Percentagem da população residente na freguesia com o ensino superior	100 x Populacao_ens_sup/ Populacao (QC)
S_O_Crescimento_Populacao	Crescimento_Populacao sem outliers	Crescimento_Populacao: Média ± 3 DP (QC)
Dummy_Crescimento_Populacao	Tipo de crescimento da população	Dummy com “positivo” ou “negativo” (QN)
Densidade_Populacional	Número de habitantes da freguesia por quilómetro quadrado	Populacao/Area_Freguesia (QC)
Tipo_Crescimento_Alojamentos	Tipo de crescimento dos alojamentos	Variável com “nulo”, “positivo” e “negativo” (QC)
Escolas_por_habitante	Número de escolas por habitante da freguesia	Escolas/Populacao (QC)
Tem_Hospital	Existência de hospitais	Variável com “sim” ou “não” (QN)
Hospitais_por_Habitante	Número de hospitais por habitante da freguesia	Hospitais/Populacao (QC)
Tem_Farmacia	Existência de farmácias	Variável com “sim” ou “não” (QN)
Farmacias_por_habitante	Número de farmácias por habitante da freguesia	Farmacias/Populacao (QC)
Tem_Estacao_Transporte	Existência de estações de transporte	Variável com “sim” ou “não” (QN)
Tem_Estacao_Metro	Existência de estações de metro	Variável com “sim” ou “não” (QN)
Tem_Estacao_Comboio	Existência de estações de comboio	Variável com “sim” ou “não” (QN)
Estacoes_Transporte_por_Habitante	Número de estações de transportes por habitante da freguesia	Estacoes_Transportes/Populacao (QC)
Estacoes_Metro_por_habitante	Estacoes_Metro_por_habitante	Número de estações de metro por habitante da freguesia (QC)
Estacoes_Comboio_por_habitante	Número de estações de comboio por habitante da freguesia	Estacoes_Comboio/Populacao (QC)
Tem_Estabelecimento_Policia	Existência de estabelecimentos policiais	Variável com “sim” ou “não” (QN)
Estabelecimentos_Policiais_por_Habitante	Número de estabelecimentos policiais por habitante da freguesia	Estabelecimentos_Policiais/Populacao (QC)
Praias_por_habitante	Número de praias por habitante da freguesia	Praias/Populacao (QC)

Notas: QC: quantitativa contínua; QD: quantitativa discreta; QO: quantitativa ordinal; e QN; quantitativa nominal

Tabela 3.8 - Estatísticas descritivas das novas variáveis qualitativas

Variável	Categoria mais frequente		Categoria menos frequente		Número de categorias
	Categoria	%	Categoria	%	
Nova_Eficiencia_Energetica	D	39,9	Menos que é	2,8	6
Novo_Lug_Estacionamento	0	59,9	3 ou mais	2,8	4
Tem_lugares_estacionamento	Não tem	59,9	Tem	40,1	2
Dummy_Crescimento_Populacao	Positivo	59,9	Negativo	40,1	2
Tipo_Crescimento_Alojamentos	Positivo	53,7	Nulo	4,2	3
Tem_Hospital	Não	63,1	Sim	36,9	2
Tem_Farmacia	Sim	98	Não	2	2
Tem_Estacao_Transporte	Sim	73,5	Não	26,5	2
Tem_Estacao_Metro	Não	68	Sim	32	2
Tem_Estacao_Comboio	Sim	56,3	Não	43,7	2
Tem_Estabelecimento_Policial	Sim	83,7	Não	16,3	2
Tem_Praia	Não	88,9	Sim	11,1	2

Nota: n=4103 imóveis.

Tabela 3.9 - Estatísticas descritivas das novas variáveis quantitativas

Variável	Média	Desvio Padrão	Mínimo	Máximo	Percentis		
					25	50	75
S_O_Valor (x1000€)	335,70	211,72	49,00	972,11	170,00	269,00	445,00
Log_10_Valor	5,45	0,26	4,69	6,00	5,23	5,43	5,65
S_O_Area_Imovel	101,07	41,10	25,00	232,00	72,00	92,00	121,00
S_O_Casas_de_Banho	1,70	0,77	1	4	1	2	2
S_O_Quartos	2,35	0,93	0	5	2	2	3
Idade_Imovel	27,82	20,83	-2	72	11	27	42
WC_por_m2	0,02	0,01	0	0,05	0,01	0,02	0,02
S_O_WC_por_m2	0,02	0,01	0	0,03	0,01	0,02	0,02
Quartos_por_m2	0,02	0,01	0	0,05	0,02	0,02	0,03
S_O_Crescimento_Populacao	0,48	6,20	-19,74	20,45	-1,30	0,50	3,60
Porcentagem_Populacao_ens_sup	25,94	11,53	6,89	51,33	16,22	21,52	35,44
Densidade_Populacional	5368,19	3919,04	51,27	17019,00	2527,63	4414,78	7052,97
S_O_Crescimento_Alojamentos	0,09	3,70	-15,23	14,50	-0,80	0,20	1,90
S_O_Area_Freguesia	12,72	14,68	1,49	82,52	3,09	7,18	16,74
Escolas_por_habitante (x1000)	0,70	0,25	0,09	2,20	0,60	0,70	0,80
Hospitais_por_Habitante (x1000)	0,02	0,03	0	0,20	0	0	0,03
Farmacias_por_habitante (x1000)	0,30	0,30	0	1,50	0,20	0,30	0,40
Estacoes_Transporte_por_habitante (x1000)	0,10	0,08	0	0,50	0	0,04	0,10
Estacoes_Metro_por_habitante (x1000)	0,04	0,07	0	0,30	0	0	0,05
Estacoes_Comboio_por_habitante (x1000)	0,03	0,04	0	0,50	0	0,02	0,04
Estabelecimentos_Policiais_por_habitante (x1000)	0,10	0,10	0	0,60	0,02	0,04	0,08
Praias_por_habitante (x1000)	0,02	0,09	0	0,70	0	0	0

Notas: n=4103 imóveis; a variável Idade_Imovel assume valores negativos quando o imóvel ainda se encontra em fase de construção e o número corresponde ao número de anos em que se prevê terminar a construção;

Assim, para a próxima fase, são considerados 4103 apartamentos e 60 variáveis (não estando incluídas as *dummy*), que são conjugadas de modo a obter as melhores métricas possíveis e, conseqüentemente, um modelo capaz de prever o valor dos imóveis e, conseqüentemente, cumprir os objetivos.

3.4. Modelação

Na quarta etapa, seleciona-se a técnica de modelação, dependendo do problema do negócio e dos dados (Chapman *et al.*, 1999) como forma de cumprimento dos objetivos delineados na etapa da compreensão do negócio, da melhor maneira possível.

Como explicado anteriormente na RSL, as técnicas preditivas mais utilizadas na literatura consultada são os modelos de regressão, redes neuronais e árvores de decisão, concretamente *Classification And Regression Trees* (CART) e *Chi-squared Automatic Interaction Detection* (CHAID), pelo que foram estes os algoritmos utilizados, no programa *SPSS Modeler* (v. 18.2).

Para tentar melhorar a qualidade dos modelos baseados em árvores de decisão e redes neuronais recorre-se às técnicas de *ensemble*, *bagging* e *boosting*,

O *boosting* produz uma sucessão de modelos de componentes, construídos em todo o conjunto de dados. Antes de construir cada modelo de componente, os registos são ponderados com base nos resíduos do modelo de componente anterior. Casos com grandes resíduos recebem pesos de análise relativamente mais altos para que o próximo modelo de componente se concentre em prever bem esses registos. Juntos, esses modelos formam um modelo de conjunto (IBM, 2021).

Por sua vez, o *bagging* produz réplicas do conjunto de dados de treino com substituição do conjunto de dados original. Isso cria amostras de *bootstrap* de tamanho igual ao conjunto de dados original. De seguida, um modelo de componente é construído em cada réplica e todos formam um modelo de conjunto (IBM, 2021). A técnica *bagging* melhora a capacidade de generalização aumentando a estabilidade do modelo e a técnica *boosting* melhora a precisão do modelo na fase de treino (Abrantes, 2020). Estes métodos foram usados por vários estudos na literatura como é o caso dos estudos de Sing *et al.* (2021), Bogin *et al.* (2020) ou Su *et al.* (2021).

Para os modelos utiliza-se como variável alvo o logaritmo do valor, atributo que, segundo especialistas e estudos como Belotti (2017), Krause *et al.* (2020) ou Doumpos *et al.* (2021), oferece maior acerto nos modelos.

As variáveis sem *outliers* foram utilizadas na regressão linear em substituição das originais porque este modelo não lida bem com *outliers* e foram criadas *dummies* para todas as variáveis qualitativas para este modelo que só aceita variáveis quantitativas, usando-se todas as variáveis (à exceção do ano de construção, usando-se apenas a idade do imóvel, pois assim um modelo, futuramente, considera a idade ao invés de ser treinado para considerar que o ano mais recente é 2024 quando não vai ser).

Na regressão linear utiliza-se o método *stepwise*, com um valor de entrada de 0,05 e de saída de 0,1 do teste de probabilidade F, em relação a entradas e saídas de variáveis, em que a restante parametrização corresponde à predefinição do nó do *SPSS Modeler*, a inclusão de constante na equação e o cálculo da importância das variáveis preditoras (IBM, 2021). Por sua vez, as diferentes

parametrizações utilizadas nas árvores de decisão encontram-se expostas na Tabela 3.10 e na Tabela 3.11, em que foram tentados dois níveis de profundidade diferentes, com sete níveis, que traz maior complexidade e maior probabilidade de sobreajustamento, e com cinco níveis, modelo mais simples, podendo levar ao subajustamento (IBM, 2021). Por fim, as redes neuronais recorrem ao modelo *Multilayer Perceptron*, fazendo a computação automática das unidades nas camadas ocultas, a regra de paragem é 15 minutos no treino do modelo, com 10 modelos de componentes para *boosting* e *bagging* e a regra de combinação predefinida para a variável alvo é a média (IBM, 2021).

Tabela 3.10 - Parametrizações do algoritmo CART

Algoritmo	CART					
	-	Boosting		Bagging		-
Modelo	A	B	C	D	E	F
Profundidade máxima	6	6	6	5	5	5
Casos do nó Pai	2	2	2	2%	2%	2%
Casos do nó filho	1	1	1	1%	1%	1%

Tabela 3.11 - Parametrizações do algoritmo CHAID

Algoritmo	CHAID					
	-	Boosting		Bagging		-
Modelo	G	H	I	J	K	L
Profundidade máxima	6	6	6	5	5	5
Casos do nó Pai	2	2	2	2%	2%	2%
Casos do nó filho	1	1	1	1%	1%	1%

Para avaliar a qualidade dos modelos, recorre-se a técnicas de teste fora da amostra, nomeadamente, o *holdout*, que divide o conjunto de dados de forma aleatória. Esta divisão resulta em 70% dos dados para o conjunto de treino e 30% dos dados para o conjunto de teste, sendo esta a divisão mais utilizada nos estudos consultados para a quantidade de dados do *dataset* (Su *et al.*, 2021).

Além disso, para avaliar os modelos aplicam-se as diversas métricas do nó *Analysis* do SPSS *Modeler* (v. 18.2), sendo estas o coeficiente de correlação (R), o erro mínimo, o erro máximo, o MAE e o desvio padrão (IBM, 2021), adicionando-se o R^2 .

3.5. Avaliação

Nesta etapa são avaliados os resultados de todo o processo realizado até então, analisando-se, com recurso a métricas, os modelos obtidos na etapa anterior e se estes vão de encontro aos objetivos traçados na primeira etapa (Chapman *et al.*, 1999).

Como explicado anteriormente, analisam-se os modelos com recurso às métricas de problemas de regressão. Na comparação dos modelos e na seleção do melhor modelo analisam-se as métricas na amostra de teste, na Tabela 3.12.

Tabela 3.12 - Descrição das métricas de avaliação dos modelos

Métrica	Expressão	Explicação
Erro mínimo	$\text{Mín.}(v_n - \hat{v}_n)$	Menor diferença entre o valor observado e previsto; quanto mais próximo de 0 melhor.
Erro máximo	$\text{Máx.}(v_n - \hat{v}_n)$	Maior diferença entre o valor observado e previsto; quanto mais próximo de 0 melhor.
MAE	$\frac{1}{N} \sum_{n=1}^N v_n - \hat{v}_n $	Média da soma das diferenças absolutas entre os valores observados e os valores previstos; quanto mais próximo de 0 melhor.
Desvio padrão	$sd(v_n - \hat{v}_n)$	Desvio padrão da diferença entre os valores observados e previstos; quanto mais próximo de 0 melhor.
R	$\frac{COV(v, \hat{v})}{sd_v sd_{\hat{v}}}$	Mede a correlação entre os valores previstos e os valores observados; quanto mais próximo de 1 melhor.
R ²	$1 - \frac{\sum_{n=1}^N (v_n - \hat{v}_n)^2}{\sum_{n=1}^N (v_n - \bar{v})^2}$	Soma dos quadrados da regressão dividida pela soma total dos quadrados; quanto mais próximo de 1 melhor.

Notas: v_n : valores observados dos imóveis correspondem ao v_n ; \hat{v}_n : valores previstos; \bar{v} : média aritmética dos valores observados; n : número total de observações; e sd : desvio padrão; COV : covariância; R: coeficiente de correlação; R²: coeficiente de determinação.

Fonte: Adaptado de Steurer *et al.* (2021).

O melhor modelo é aquele que apresenta as melhores métricas, na generalidade, ou seja, o que tiver o R² mais elevado e as métricas dos erros mais reduzidas. É importante analisar as métricas como um todo de modo a ter um modelo com bom desempenho na generalidade.

Adicionalmente, na avaliação verifica-se a explicabilidade dos modelos, sendo que quando esta não é imediata (como nos casos das árvores de decisão ou do modelo de regressão linear) realizam-se análises de sensibilidade, isto é, analisa-se a relação entre cada uma das variáveis preditoras mais importantes e o valor estimado do imóvel (Dimopoulos & Bakas, 2019).

3.6. Implementação

Por fim, a última etapa do CRISP-DM corresponde à apresentação e ao plano de implementação dos resultados obtidos através do modelo, de monitorização e de manutenção (Chapman *et al.*, 1999). Esta dissertação assume-se relevante para o setor porque possibilita a evolução da construção de novos sistemas de avaliação, os AVM, de modo a prever o valor dos imóveis, recorrendo a técnicas analíticas avançadas. Contudo, os prazos estabelecidos para esta investigação não permitiram a execução desta fase, para além desta dissertação.

4. Resultados e Discussão

Este capítulo pretende dar seguimento à metodologia abordada anteriormente, expondo os resultados das análises efetuadas, interpretando-os, e discutindo a sua relevância para cada um dos objetivos definidos, que consistem na caracterização do valores dos imóveis, procedimento realizado no capítulo 1, na caracterização da relação entre os valores dos imóveis e as suas características, e na criação de modelos preditivos do valor em função das características do imóvel e do mercado onde se insere.

4.1. Caracterização dos valores dos imóveis

Este ponto consiste na execução do primeiro objetivo, a caracterização do valor dos imóveis, processo realizado no capítulo 3 (Metodologia). Da análise, resultaram diversas conclusões, entre as quais se destaca que os imóveis que se situam a norte do Rio Tejo são mais valiosos do que os do sul (média de cerca de 313584 euros). Verificou-se, ainda, que metade dos imóveis têm um valor máximo de 269 mil euros, com dispersão é elevada pois, em média, os valores afastam-se da média em 212 mil euros.

4.2. Relação entre os valores dos imóveis e as suas características

Neste ponto, caracteriza-se a relação entre o valor dos imóveis e as suas características, de modo a perceber o impacto que estas têm na variável dependente.

No que diz respeito à relação com as variáveis quantitativas, recorre-se à correlação de Pearson, uma medida de associação, para analisar se estas variáveis têm capacidade de explicar o valor, obtendo modelos com bom desempenho.

Da Tabela 3.6, destacam-se com os maiores valores, as variáveis predictoras: área do imóvel (0,607), número de casas de banho (0,595), o que significa que têm relações fortes com o valor do imóvel. Assim, conclui-se que à medida que os valores das variáveis aumentam, o valor do imóvel tende a aumentar. Além disso, a população com o 1º ciclo e com o 2º ciclo (-0,365) são as variáveis com maior valor de correlação negativa, ou seja, à medida que os valores destas aumentam, existe uma tendência para a diminuição no valor dos imóveis.

Em relação às variáveis qualitativas caracterizadoras dos imóveis e da sua localização, realiza-se uma análise descritiva, recorrendo a várias medidas de síntese. Na Tabela 4.1 apresenta-se a distribuição do valor do imóvel para cada categoria das suas características.

Tabela 4.1 - Valor do imóvel pelas suas características

Característica	Média	Desvio Padrão	Mínimo	Máximo	Mediana
Piso					
C/V	343417,33	211631,08	67500	1000000	278500
R/C	321523,36	205336,52	68000	980000	253000
1	308570,94	198559,34	57500	999000	246500
2	377804,53	218654,89	50000	998000	327500
3	401287,04	218404,56	60000	990000	350000
4	415487,62	255786,76	55000	990000	359000
5	400101,27	247367,23	115000	1000000	338500
6	502185,82	249329,98	135000	995000	502500
7	387802,85	253319,54	85000	985000	275000
8	503943,79	244139,96	135000	990000	529000
9	230513,56	146715,85	75000	775000	175000
Acima do 10	276508,66	162657,60	49000	1000000	230000
Eficiência Energética					
A	552399,97	220254,26	175000	999000	520000
A+	426483,08	199139,65	220000	997000	339000
B	507576,53	225813,70	110000	1000000	455000
B-	415462,09	206240,51	95000	980000	381000
C	307059,80	190237,84	49000	995000	240000
D	321800,38	206089,04	50000	1000000	250000
E	252578,13	163725,82	55000	960000	190000
F	255160,09	170561,91	70000	902500	190000
G	180500,00	53294,33	103000	280000	181000
A	552399,97	220254,26	175000	999000	520000
A+	426483,08	199139,65	220000	997000	339000
Elevador					
Não	225200,22	135423,27	49000	995000	180000
Sim	408584,22	221758,62	50000	1000000	350000
Lugares de Estacionamento					
Não	250056,95	161127,17	49000	1000000	190000
1	406136,04	192876,04	110000	998000	352500
2	536436,69	214491,50	140000	1000000	515000
3 ou mais	623241,39	210715,34	139000	995000	610000

Nota: n=4103 imóveis.

Analisando a Tabela 4.1, verifica-se que o valor sofre alterações aquando da variação do piso pois, em média, um imóvel é mais valioso se se situar no sexto andar (cerca de 502186 euros) e no oitavo andar (cerca de 503944 euros) e, a partir do nono andar, o valor tende a diminuir. Apesar do primeiro andar ser a categoria mais frequente (19,90%), tem o terceiro valor médio mais baixo (cerca de 308571 euros) e metade destes imóveis tem um valor de, no máximo, 246500 euros. Por sua vez, constata-se que existe uma tendência para os imóveis terem um valor maior entre o terceiro e o sexto andar.

Além disso, analisa-se o impacto do escalão da eficiência energética no valor do imóvel, onde se conclui que quanto melhor for o escalão (tendência para o A), maior tende a ser o valor. Por exemplo, metade dos imóveis com o escalão A vale mais de 520 mil euros, enquanto que metade dos imóveis com o último escalão (G) vale até 181 mil euros. Ainda assim, à exceção do escalão G, verifica-se que todos os escalões estão associados a todo o tipo de imóveis (sejam mais ou menos valorizados), pois a amplitude entre o máximo e o mínimo não regista diferenças acentuadas.

Por sua vez, verifica-se que o facto do prédio do imóvel ter elevador tem muita influência no valor do imóvel. Em média, a existência de elevador aumenta muito o valor de um apartamento (cerca de 408584 euros) em relação a não existir (cerca de 225200 euros). Além disso, metade dos imóveis sem elevador valem até 180 mil euros, enquanto que, com elevador, valem até 350 mil euros.

Além disso, analisou-se a relação dos lugares de estacionamento com o valor, constatando-se que um imóvel, em média, vale mais quando tem (a partir de 406 mil euros) do que quando não tem (cerca de 250 mil euros). Por exemplo, metade dos imóveis com três lugares associados tem um valor mínimo de 610 mil euros e sem lugares valem até 190 mil euros. Imóveis com lugares associados estão relacionados com maiores condições, o que aumenta o valor dos mesmos.

Posteriormente, na Tabela 4.2, analisa-se a distribuição do valor pelas características da macro localização como o distrito, o concelho e a freguesia, sendo que, devido à extensão da informação desta (mais de 100 freguesias), analisam-se as 10 mais representadas no *dataset*.

Tabela 4.2 - Valor do imóvel por distrito, concelho e freguesia (*top 10*)

Característica	Média	Desvio Padrão	Mínimo	Máximo	Mediana
Distrito					
Lisboa	361820,94	218125,90	55000	1000000	290000
Setúbal	234670,28	148528,63	49000	913000	185000
Concelho					
Alcochete	404075,71	272068,19	99900	913000	298000
Almada	220166,89	93149,28	65000	570000	196750
Amadora	221892,96	94192,62	80000	600000	190000
Barreiro	254859,24	152944,89	49000	750000	215000
Cascais	441203,53	210734,93	100000	995000	382500
Lisboa	477315,87	222110,97	80000	1000000	440000
Loures	275355,78	161520,23	70000	849000	225000
Mafra	303543,70	182845,46	112500	967659	249000
Moita	115944,68	41272,91	50000	322000	111000
Montijo	304953,13	152463,95	101000	760000	290000
Odivelas	265239,80	115702,02	87000	650000	230000
Oeiras	450819,27	214950,81	109000	980000	420000
Palmela	182333,04	54090,89	95000	315000	176250
Seixal	205288,51	116459,39	80000	682500	164900
Sesimbra	378753,96	217951,86	100000	805000	295000
Setúbal	173642,24	80717,43	57500	495000	155000
Sintra	192778,80	93501,08	55000	830000	165000
Vila Franca de Xira	194101,08	86105,69	57000	560000	170000
Freguesia					
Algés, Linda-a-Velha e Cruz Quebrada-Dafundo	521433,32	224819,60	109000	980000	557500
Avenidas Novas	631639,65	219253,61	175000	998000	650000
Queluz e Belas	213065,71	123667,81	75000	830000	170000
Odivelas	272779,06	135028,56	97500	650000	207000
Oeiras e São Julião da Barra, Paço de Arcos e Caxias	430467,37	198202,79	145000	890000	399000
Carcavelos e Parede	505536,81	226949,50	100000	975000	490000
Arroios	403541,75	178989,60	99000	850000	370000
Seixal, Arrentela e Aldeia de Paio Pires	233590,80	143467,28	90000	682500	159500
Montijo e Afonsoeiro	314116,45	152891,99	104000	760000	297000
Santo António	596558,11	169866,46	299000	1000000	550000

Nota: n=4103 imóveis.

A Tabela 4.2 demonstra que a localização tem importância. Um imóvel tem maior valor, em média, se pertencer ao distrito de Lisboa (cerca de 361821 euros). No entanto, os imóveis alteram o seu valor se pertencerem a concelhos diferentes. Por exemplo, um imóvel no município de Lisboa, Cascais, Oeiras ou Alcochete tem, em média, um valor superior a 400 mil euros, valor que é superior aos máximos de Palmela (315 mil euros) e da Moita (322 mil euros).

Continuando a análise na macro localização, verifica-se que, por exemplo, um imóvel que se situe na freguesia de Avenidas Novas é muito valorizado (média de cerca de 631 mil euros), onde metade dos imóveis tem um valor superior a 650 mil euros, contrastando com a freguesia de Queluz e Belas, em que metade dos imóveis vale até 170 mil euros. Assim, é possível que dois imóveis com as mesmas características, mas com localizações diferentes tenham valores muito díspares, fator que importa na decisão do comprador/vendedor e é incorporado na avaliação dos imóveis.

4.3. Modelos preditivos dos valores dos imóveis

Na construção dos modelos preditivos opta-se por se utilizar sempre a mesma variável dependente, o logaritmo do valor do imóvel, onde foi comprovado por autores como Belotti (2017), Krause *et al.* (2020) ou Doumpos *et al.* (2021) que produz maior eficácia nestes modelos do que a variável do valor.

No entanto, para os restantes modelos, as variáveis *dummy* não são utilizadas, pois as árvores de decisão aceitam variáveis qualitativas e as redes neuronais criam variáveis *dummy* automaticamente. Também não são usadas variáveis sem *outliers* pois os restantes modelos lidam bem com os mesmos (IBM, 2021). Por sua vez, utilizam-se sempre as mesmas variáveis para os modelos de modo a perceber os comportamentos diferentes dos vários algoritmos quando sujeitos aos mesmos atributos.

4.3.1. Qualidade dos modelos preditivos

Como referido no ponto 3.5, os modelos são avaliados segundo várias métricas, expostas entre a Tabela 4.3 e a Tabela 4.6, aplicadas aos conjuntos de treino e teste. Como nos modelos se considera como variável alvo o logaritmo do valor, as métricas, à exceção do R e do R², estão expressas em logaritmo de base 10. Assim, para se converter os valores dos erros e os valores previstos pelo modelo usa-se a função inversa do logaritmo de base 10 (conhecido por antilog), em que \hat{v} é o valor previsto.

$$\text{Antilog} = 10^{\hat{v}} \quad (8)$$

O modelo de regressão linear múltipla³, inclui, assim, 62 variáveis significativas, destacando-se como a mais importante a *dummy* sobre a pertença à freguesia de Santiago (Sesimbra), o que confirma o valor da relação forte com o valor, e a menos importante o poder de compra *per capita*.

³ A equação do modelo encontra-se no Anexo C: Equação da Regressão.

Tabela 4.3 - Métricas de avaliação do modelo de regressão linear múltipla

Conjunto de dados	Erro mínimo	Erro máximo	MAE	Desvio Padrão	R	R ²
Treino	-0,46	0,39	0,07	0,09	0,94	0,88
Teste	-0,44	0,37	0,07	0,09	0,94	0,88

Notas: MAE: erro médio absoluto; R: coeficiente de correlação; R²: coeficiente de determinação; variável dependente: log₁₀valor.

Em relação aos modelos do algoritmo CART, como é possível verificar na Tabela 4.4 destaca-se como o modelo E (CART com o método *boosting*), pois, apesar de não ter o melhor R², tem melhor combinação de métricas. Este teve, igualmente, como variável alvo o logaritmo do valor e as variáveis predictoras com maior valor de importância para a explicação do modelo são: a freguesia, o concelho, a área do imóvel, a percentagem de população com o ensino superior e o número de casas de banho.

Para a visualização do modelo, realiza-se um modelo CART sem *ensembles*, em que o valor do logaritmo previsto é a variável alvo, mantendo as variáveis independentes. Daqui, resulta que o modelo tem cinco níveis de profundidade e 48 nós, dos quais 25 são terminais. Da análise das regras da árvore, destacam-se, por exemplo, as referentes aos nós 42 (valor estimado sobrevalorizado), 34 (valor com boa qualidade de estimação) e 23 (valor estimado subvalorizado)⁴.

Tabela 4.4 - Métricas de avaliação dos modelos CART

Conjunto de dados	Modelos	Erro mínimo	Erro máximo	MAE	DP	R	R ²
Treino	A	-0,47	0,48	0,08	0,11	0,91	0,83
	B	-0,26	0,28	0,07	0,08	0,95	0,90
	C	-0,40	0,36	0,07	0,09	0,94	0,88
	D	-0,47	0,51	0,09	0,12	0,90	0,80
	E	-0,33	0,30	0,07	0,09	0,94	0,88
	F	-0,41	0,47	0,08	0,10	0,93	0,87
Teste	A	-0,48	0,51	0,09	0,12	0,89	0,79
	B	-0,43	0,50	0,08	0,10	0,93	0,87
	C	-0,41	0,54	0,08	0,10	0,92	0,85
	D	-0,45	0,51	0,09	0,12	0,88	0,77
	E	-0,40	0,46	0,08	0,10	0,92	0,85
	F	-0,41	0,58	0,08	0,11	0,91	0,83

Notas: MAE: erro médio absoluto; DP: desvio padrão; R: coeficiente de correlação; R²: coeficiente de determinação; variável dependente: log₁₀valor.

No que diz respeito aos modelos do algoritmo CHAID, constata-se que, através da Tabela 4.5 o melhor é o I (CHAID com o método *bagging*), pois mesmo não tendo o melhor erro máximo, tem o melhor R² (0,87) e a melhor combinação de métricas. À semelhança do modelo E do CART, este modelo tem como variável alvo o logaritmo do valor e como variáveis predictoras, com maior valor de importância para a explicação do modelo, a freguesia, o concelho, a área do imóvel, a percentagem de população com o ensino superior e o número de casas de banho.

⁴ As regras dos nós estão presentes no Anexo D: Regras das árvores.

Para melhor visualização do modelo, realiza-se um modelo CHAID sem *ensembles*, em que o valor do logaritmo previsto é a variável alvo, mantendo as variáveis preditoras. Resulta daí um modelo com seis níveis de profundidade e 68 nós, dos quais 34 são terminais. Da análise das regras da árvore, destacam-se, por exemplo, as referentes aos nós 22 (valor estimado sobrevalorizado), 34 (valor com boa qualidade de estimação) e 44 (valor estimado subvalorizado)⁵.

Tabela 4.5 - Métricas de avaliação de treino dos modelos CHAID

Conjunto de dados	Modelos	Erro mínimo	Erro máximo	MAE	DP	R	R ²
Treino	G	-0,40	0,33	0,05	0,07	0,96	0,92
	H	-0,15	0,12	0,02	0,03	0,99	0,98
	I	-0,26	0,28	0,04	0,05	0,98	0,96
	J	-0,46	0,49	0,09	0,12	0,90	0,80
	K	-0,36	0,29	0,08	0,09	0,94	0,88
	L	-0,44	0,43	0,08	0,10	0,93	0,87
Teste	G	-0,54	0,57	0,09	0,12	0,89	0,79
	H	-0,45	0,44	0,08	0,10	0,92	0,85
	I	-0,43	0,53	0,07	0,10	0,93	0,87
	J	-0,60	0,47	0,10	0,13	0,87	0,76
	K	-0,64	0,42	0,09	0,11	0,91	0,83
	L	-0,55	0,53	0,09	0,11	0,90	0,80

Notas: MAE: erro médio absoluto; DP: desvio padrão; R: coeficiente de correlação; R²: coeficiente de determinação; variável dependente: log₁₀ valor.

No que diz respeito aos modelos das redes neuronais, verifica-se na Tabela 4.6 que o melhor é o que recorre ao método *bagging*, pois é o melhor em todas as métricas. O modelo tem como variável alvo o logaritmo do valor e como preditoras mais importantes: a freguesia, o concelho, a área do imóvel, os lugares de estacionamento e o poder de compra *per capita*.

Tabela 4.6 - Métricas de avaliação dos modelos das redes neuronais

Ensembles	Conjunto de dados	Erro mínimo	Erro máximo	MAE	DP	R	R ²
-	Treino	-0,56	0,49	0,07	0,09	0,93	0,87
	Teste	-0,47	0,55	0,08	0,10	0,93	0,87
<i>Boosting</i>	Treino	-0,15	0,28	0,02	0,03	0,99	0,98
	Teste	-0,46	0,50	0,07	0,09	0,94	0,88
<i>Bagging</i>	Treino	-0,28	0,28	0,04	0,05	0,98	0,96
	Teste	-0,37	0,47	0,06	0,08	0,95	0,90

Notas: MAE: erro médio absoluto; DP: desvio padrão; R: coeficiente de correlação; R²: coeficiente de determinação; variável dependente: log₁₀ valor.

Através da análise das Tabela 4.3 a Tabela 4.6, conclui-se que se obtêm modelos com bom desempenho, dentro dos valores do R² dos estudos consultados (entre 0,75 e 0,90), apesar de existirem quatro modelos que não cumprem o mínimo definido de 0,80 no R² (os modelos sem *ensembles* do CART e os do CHAID). O modelo com melhores resultados no R² é a rede neuronal com

⁵ As regras dos nós estão presentes no Anexo D: Regras das árvores.

o método *bagging*, (0,87), sendo o melhor no MAE (0,06), no desvio padrão (0,08) e no erro mínimo (-0,37), valorizando menos os imóveis do que o resto dos modelos, apesar de não ser dos melhores no erro máximo (0,47).

Em relação ao erro mínimo, as redes neuronais apresentam menor valor e, no que diz respeito ao erro máximo, a regressão linear obteve a melhor qualidade. Em relação ao MAE, todos apresentam valores muito próximos (entre 0,06 e 0,10). Da análise das métricas resultou que o modelo das redes neuronais com o método *bagging* é o melhor. Em relação às variáveis preditoras mais importantes nos diferentes modelos, a Tabela 4.7 apresenta o *top 5* para os melhores modelos de cada algoritmo, no caso de cada variável ter uma importância relativa superior a 5%.

Tabela 4.7 - Variáveis preditoras mais importantes

Regressão Linear	CART	CHAID	Redes Neuronais
<i>Dummy</i> da freguesia de Santiago (Sesimbra)	Freguesia	Freguesia	Freguesia
Área do imóvel sem <i>outliers</i>	Percentagem da população com o ensino superior	Percentagem da população com o ensino superior	Concelho
<i>Dummy</i> do concelho de Odivelas	Concelho	Concelho	Área do imóvel
Quartos sem <i>outliers</i>	Área do imóvel	Área do imóvel	Lugares de estacionamento (variável recodificada)
<i>Dummy</i> da inexistência de lugares de estacionamento	Casas de banho	Casas de banho	Poder de compra <i>per capita</i>

Através da análise da tabela, verifica-se que as variáveis da localização (freguesia e concelho) e área do imóvel são muito valorizadas por todos modelos, sendo cruciais na determinação do valor dos imóveis. Além disso, verifica-se que a percentagem de população com o ensino superior também é importante para os algoritmos das árvores de decisão, e os restantes algoritmos substituem esta variável pelo número de lugares de estacionamento. Contudo, as casas de banho são uma componente importante na obtenção do valor, sendo das mais importantes em dois dos algoritmos.

Como não é possível avaliar o impacto de todas as variáveis na variável alvo, devido à extensão, foi realizada, às cinco variáveis preditoras mais importantes do melhor modelo (redes neuronais com *bagging*), uma análise de sensibilidade, tal como nos modelos do estudo de Sing *et al.* (2021).

Como consta na Figura 4.1, o logaritmo do valor tende a ser maior quanto maior é a área. Esta relação faz sentido pois, apartamentos com maiores áreas, são maiores, podem ter mais quartos, mais casas de banho, mais condições e, assim, terem maior valor. Analisando a figura, percebe-se que imóveis com área superior a 200 m² têm, no mínimo, um logaritmo do valor de cerca de 5,6, o que perfaz um valor de cerca de 400 mil euros. Observando a Figura 4.2, verifica-se que a relação existente entre a previsão do modelo e a área do imóvel é muito semelhante ao que se verifica na realidade.

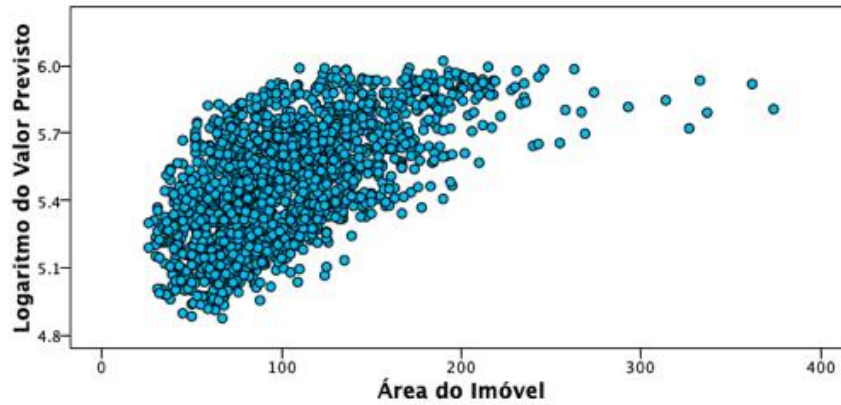


Figura 4.1 - Relação entre a área do imóvel e o logaritmo do valor previsto

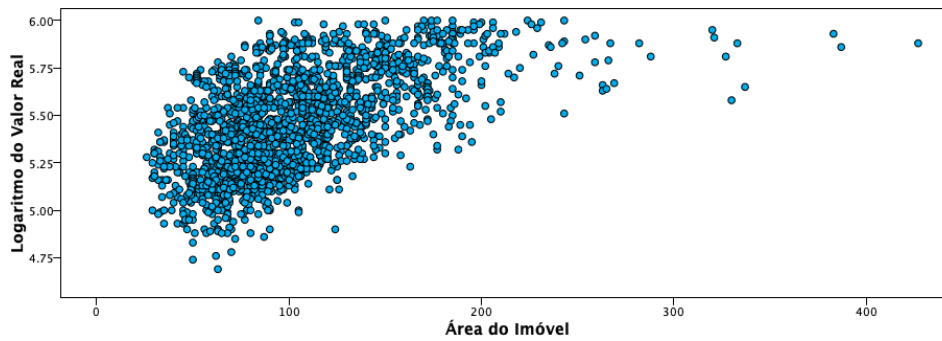


Figura 4.2 - Relação entre a área do imóvel e o logaritmo do valor observado

Também os lugares de estacionamento, outra característica do imóvel, foram analisados (Figura 4.3). Quanto mais lugares um imóvel tem, maior tende a ser o logaritmo do valor. Tendencialmente, um imóvel com vários lugares é mais caro, pois características como ter três lugares podem ser consideradas luxo, inflacionando o valor do imóvel. Por exemplo, um imóvel com pelo menos dois lugares tem, em média, um logaritmo do valor mínimo de 5,7 (cerca de 500 mil euros). Destaca-se a similitude entre o logaritmo do valor previsto e o real, onde a maior diferença reside nos imóveis com, pelo menos, três lugares de estacionamento, sendo uma diferença muito reduzida (cerca de 0,01).

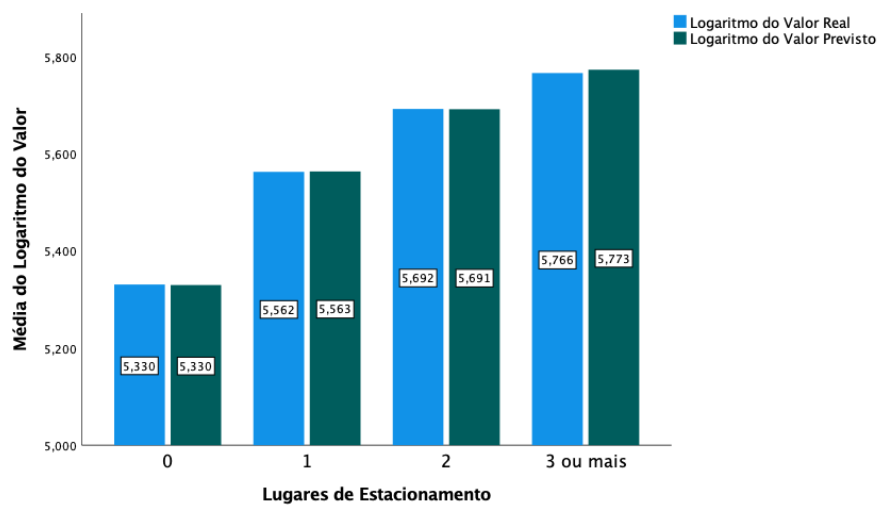


Figura 4.3 - Logaritmo médio do valor por lugares de estacionamento

No que diz respeito a indicadores económicos, avaliou-se o poder de compra *per capita* na Figura 4.4. Verifica-se uma tendência para que à medida que o poder de compra aumenta, o logaritmo do valor também aumenta. Esta conjugação faz todo o sentido, pois zonas com maior riqueza têm tendência a melhores condições, logo maior acesso a imóveis mais valorizados. Como se observa na figura, um valor de 120, na variável poder de compra *per capita*, tem um logaritmo do valor mínimo de cerca de 5,5 (cerca de 316 mil euros), enquanto que para valores inferiores a 120, o valor máximo do logaritmo do valor previsto é 5,6 (cerca de 398 mil euros). Os valores previstos são muito semelhantes aos valores reais, como comprovado na Figura 4.5.

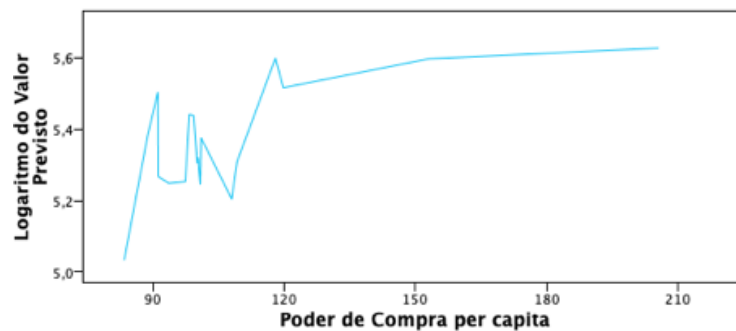


Figura 4.4 - Relação entre o poder de compra per capita e o logaritmo do valor previsto

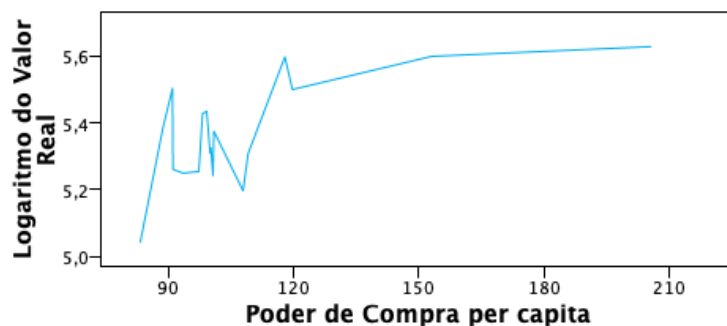


Figura 4.5 - Relação entre o poder de compra per capita e o logaritmo do valor real

Já quanto à macro localização analisaram-se os concelhos, na Figura 4.6 e Figura 4.7, e as freguesias, na Figura 4.8 e Figura 4.9. Analisando a variável dos concelhos constata-se que os municípios de Alcochete, Cascais, Lisboa, Oeiras e Sesimbra são aqueles que estão associados aos maiores valores. É de realçar que o concelho de Lisboa é aquele que tem o maior valor (ultrapassando os 5,6, ou seja, mais de 400 mil euros), logo seguido pelos concelhos de Cascais e Oeiras (cerca de 5,6). Por sua vez, quanto aos valores mais baixos, o concelho da Moita destaca-se largamente com um logaritmo do valor que ultrapassa ligeiramente 5 (100 mil euros), sendo que o segundo concelho menos valioso (Setúbal) assume um valor de 5,2 (cerca de 158500 euros). Destaca-se a similitude entre a previsão dos modelos e a realidade.

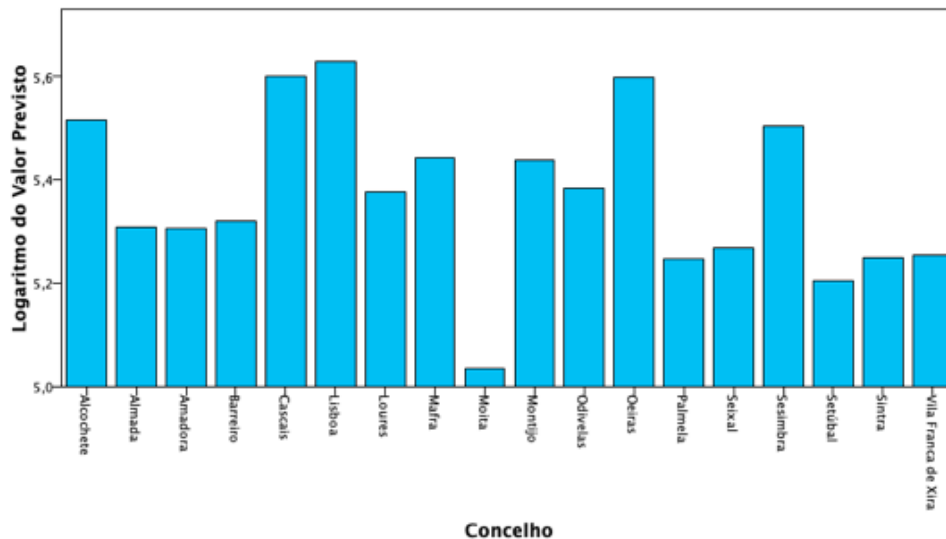


Figura 4.6 - Média do logaritmo do valor previsto por concelho

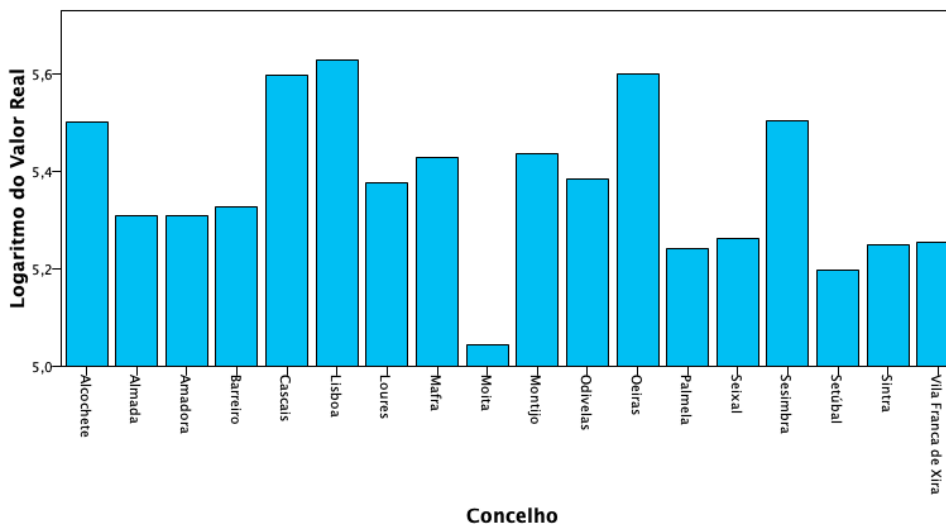


Figura 4.7 - Média do logaritmo do valor real por concelho

Por outro lado, na variável das freguesias analisaram-se, novamente, as 10 mais representadas no *dataset* (há mais de 100 freguesias). Observa-se que, por exemplo, as freguesias de Avenidas Novas e Santo António são as que estão associados aos maiores valores, cerca de 5,75 (cerca de 560 mil euros). Estas freguesias pertencem ao concelho de Lisboa que, como consta na Figura 4.6, tem um dos maiores valores, logo confirma a veracidade desta informação. Já as freguesias de Queluz e Belas, e Seixal, Arrentela e Aldeia de Paio Pires são aquelas que apresentam a média do logaritmo do valor mais reduzida, rondando os 5,3 (cerca de 200 mil euros). Destaca-se a semelhança entre a previsão e a realidade.

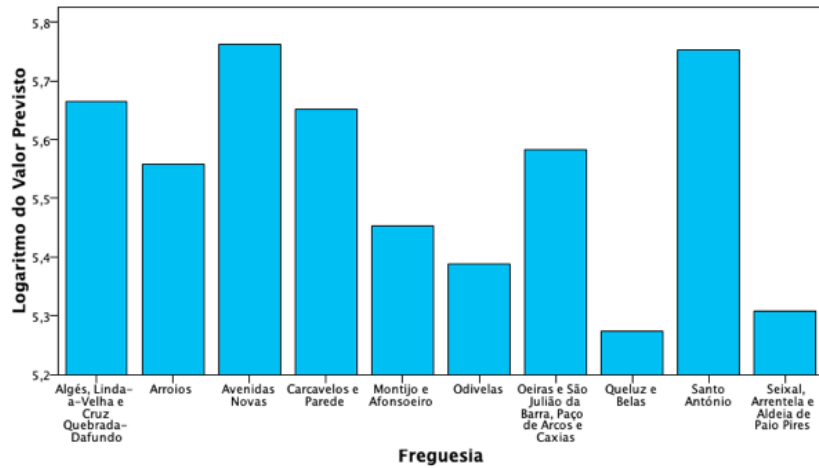


Figura 4.8 - Média do logaritmo do valor previsto por freguesia

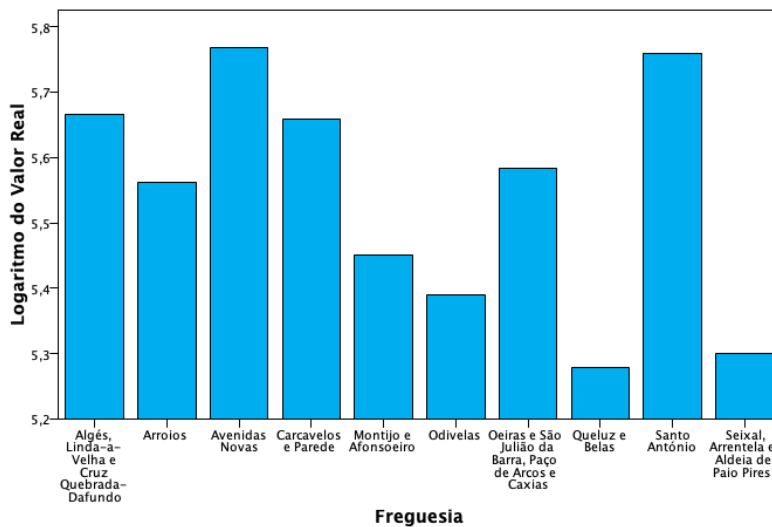


Figura 4.9 - Média do logaritmo do valor real por freguesia

Em relação às características dos imóveis, o modelo confirma a área útil como a mais importante, como defendem Sing *et al.* (2021) e Hinrichs *et al.* (2021). Por outro lado, Hinrichs *et al.* (2021) também referem as casas de banho como um dos atributos mais importantes. Por sua vez, Chmielewska *et al.* (2020) atribuem maior importância aos quartos, variável que não é muito relevante no modelo (12ª variável mais importante com um valor de 2%). Contudo, estes autores atribuem a mesma relevância aos lugares de estacionamento.

No que diz respeito à localização, a investigação confirma a importância para os modelos (concelho e freguesia foram das mais importantes) sendo chave nas previsões, tal como para Tchuente e Nyawa (2022), sendo que para Oust *et al.* (2020), os modelos perdem qualidade aquando da sua inserção.

Posteriormente, recorre-se ao CART para realizar a árvore do melhor modelo (que não é uma árvore), em que a variável alvo é a previsão da rede neuronal, mantendo-se as variáveis. A árvore tem

como intuito aprimorar a explicação do modelo com a descrição de algumas regras formadas pelo mesmo. Isto é importante para o *stakeholder* perceber a razão das previsões⁶.

A árvore tem seis níveis de profundidade e 70 nós, em que 35 são nós terminais. O concelho é a primeira ramificação, em que o nó 1 corresponde aos concelhos de Alcochete, Cascais, Lisboa, Oeiras e Sesimbra e o nó 2 aos restantes. No nó 1 prevê-se, em média, que o logaritmo do valor seja 5,613 e no nó 2 que seja 5,296. O nó 3 e o nó 4 dizem respeito à área do imóvel, que tem muita importância na estimativa do modelo, onde maior área equivale a maior logaritmo do valor. O quinto e sexto nós estão relacionados com os lugares de estacionamento, em que um apartamento com lugares associados vale mais do que um sem lugares. A variável *Area_Imovel* é a que está presente em mais ramificações (12), correspondendo ao facto de ser das variáveis com maior importância.

Da análise das regras da árvore, distinguem-se as dos nós 75 (imóvel com valor sobrevalorizado), 38 (imóvel com valor com boa qualidade de estimação) e 45 (com valor subvalorizado).

- Se (Concelho="Loures" ou Concelho="Mafra") e (Novo_Lug_Estacionamento="1" ou Novo_Lug_Estacionamento="2" ou Novo_Lug_Estacionamento="3 ou mais"), e (*Area_Imovel* > 88,500 m² e *Area_Imovel* <= 122 m²), e *Idade_Imovel* <= 12,500, e (Freguesia="Ericeira" ou Freguesia="Moscavide e Portela" ou Freguesia="Santo Isidoro"), então *Log_10_Valor* = 5,815 (Nó 75: suporte = 4 imóveis);
- Se (Concelho="Alcochete" ou Concelho="Cascais" ou Concelho="Lisboa" ou Concelho="Oeiras" ou Concelho="Sesimbra"), e (*Area_Imovel* > 58,500 m² e *Area_Imovel* <= 100,500 m²), e *Idade_Imovel* > 19,500, e *Tem_Estacao_Transporte*="Sim", então *Log_10_Valor* = 5,493 (Nó 38: suporte = 173 imóveis);
- Se *Novo_Lug_Estacionamento*=0, e *Area_Imovel* <= 83,500, e (Concelho="Barreiro" ou Concelho="Moita" ou Concelho="Montijo" ou Concelho="Setúbal"), e *Percentagem_Populacao_ens_sup* <= 16,945 então *Log_10_Valor* = 4,987 (Nó 45: suporte = 42 imóveis).

Na literatura não foram usadas variáveis que contextualizem a população como neste estudo, constituindo um fator inovador, pois são das mais importantes. É de salientar que não foram encontrados casos que utilizem o poder de compra para medir a riqueza, utilizando, por exemplo, taxas de juro (crédito à habitação) ou benefícios fiscais como consta no estudo de Hinrichs *et al.* (2021).

Contudo, quando se compara com a literatura analisada, não havendo nenhum modelo com a dados de Portugal, é necessário compreender que os países têm contextos diferentes, quer seja ao nível do imóvel, quer seja da população e, principalmente, nos fatores económicos, peça chave na valorização dos imóveis. Por exemplo, se o poder de compra for mais baixo, o valor dos imóveis desce.

⁶ O *rule set* está presente no Anexo E: *Rule set* do melhor modelo.

Apesar disso, existem elementos comuns na valorização de um imóvel na maioria dos modelos, como as condições económicas, as variáveis da área do imóvel, das casas de banho ou do estacionamento.

Os objetivos de negócio e de DM, definidos no ponto 3.1, como o aumento da qualidade da estimativa dos valores dos imóveis e, assim, melhorar as estimativas do valor do imóvel fornecidas aos clientes e a criação de um modelo preditivo capaz de prever o valor dos imóveis foram cumpridos. A qualidade do modelo comprova-se, também, pelo valor do R^2 , cuja meta era a obtenção de um R^2 superior a 0,8 e alcançou-se um valor de 0,90, atingindo-se esse objetivo, conseguindo-se, assim, um modelo preditivo capaz de estimar o logaritmo do valor e, posteriormente, o valor.

4.3.2. Perfis de imóveis com avaliações certas e erradas

Para melhorar a compreensão do modelo, realiza-se uma análise de perfis de imóveis para os quais o modelo tem estimativas ajustadas, sobrestimadas e subestimadas, sendo útil para os *stakeholders*.

Assim, ao conjunto de treino, procede-se ao cálculo do resíduo (diferença entre o valor observado e o previsto) e calcula-se o seu peso no valor observado. Daí realiza-se a recodificação do resíduo em três escalões: sobrestimado, quando o imóvel tem um resíduo inferior a -10%, subestimado, quando o resíduo é superior a 10%, e ajustado, caso contrário. Após estas etapas, realiza-se uma análise descritiva das variáveis do valor, do valor previsto, do resíduo e do peso do resíduo (Tabela 4.8).

Tabela 4.8 - Estatísticas descritivas das variáveis do processo de análise de resíduos

Variável	Média	Desvio Padrão	Mínimo	Máximo	Mediana
Valor (€)	336025,81	212531,97	49000,00	1000000,00	268000,00
Valor_Previsto (€)	332040,67	203388,98	67620,16	988524,41	267573,04
Resíduo (€)	3985,14	40169,30	-166898,23	342083,98	348,67
Peso_Resíduo (%)	-0,52	11,23	-88,72	48,18	0,17

Nota: n=2834 imóveis.

Da Tabela 4.8, verifica-se que os valores reais e previstos são idênticos, à exceção dos mínimos que distam cerca de 18600 euros, o que pode indiciar que o modelo é pior para imóveis de baixo valor.

Além disso, o resíduo apresenta uma média positiva (cerca de 3895 euros), ou seja, na generalidade o modelo subestima os valores (valor observado superior ao previsto). Apesar de haver um resíduo máximo de cerca de 342 mil euros, verifica-se que metade dos imóveis apresenta um resíduo até 348,67 euros, o que é positivo para o modelo. Contudo, no que toca ao peso do resíduo, existem imóveis cujo peso é cerca de 88%, ou seja, a estimativa está muito errada.

Apesar de alguns imóveis terem uma previsão errada (cerca de 29%), como consta na Figura 4.10, na generalidade, o modelo acerta na previsão do valor dos imóveis (70% de valores ajustados na estimativa), sendo importante perceber quais os valores que o modelo consegue estimar acertadamente (com uma margem de erro máxima de 10%) e quais os que não consegue.

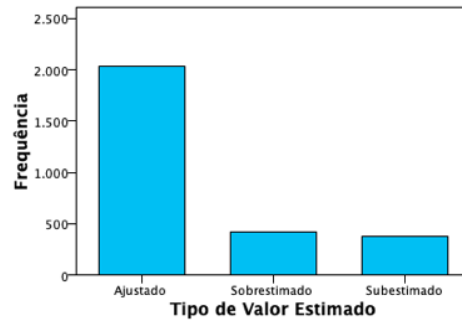


Figura 4.10 - Frequência do tipo de estimativa do modelo

Assim, elabora-se uma árvore de decisão (CART), com as mesmas variáveis, (a variável alvo é o Tipo_Erro_Avaliação), para traçar perfis em que o modelo subestima, sobrestima ou acerta na previsão. Com os perfis, os utilizadores do modelo conseguem perceber quais os imóveis em que há fiabilidade para prever o valor. Estes podem ser usados para muitos cenários (descrito no ponto 2.1.3). Além disso, tem a parametrização igual ao modelo A, à exceção do uso de custos de classificação. A árvore tem cinco níveis de profundidade, 92 nós (47 são terminais), onde as regras foram usadas para caracterizar perfis ajustado, subestimado e sobrestimado⁷. Com estas regras⁸, é possível conhecer as limitações e as potencialidades do modelo e, assim, entender em que circunstâncias o usar.

Por exemplo, imóveis em construção ou construídos no ano corrente, na cave, entre o 2º e o 5º andar ou acima do 10º andar, que pertençam a Almada, Amadora, Barreiro, Cascais, Lisboa, Loures, Odivelas, Oeiras, Palmela, Seixal, Sintra ou Vila Franca de Xira, e que a área do imóvel seja, no máximo, 116 m², o modelo tem boa qualidade de estimação (confiança=1; suporte=52 imóveis).

Em imóveis contruídos até ao ano anterior ao ano corrente, nos concelhos de Alcochete, Almada, Cascais, Mafra, Odivelas, Oeiras, Sesimbra ou Setúbal, cujo escalão da eficiência energética seja A, B ou abaixo de E, na freguesia de Almada, Cova da Piedade, Pragal e Cacilhas, Carnaxide e Queijas, Cascais e Estoril, Laranjeiro e Feijó, Odivelas, Oeiras e São Julião da Barra, Paço de Arcos e Caxias, Pontinha e Famões, Porto Salvo, Póvoa de Santo Adrião e Olival Basto, Santiago (Sesimbra), São Domingos de Rana, São Francisco, São Sebastião, ou Venda do Pinheiro e Santo Estêvão das Galés, no 2º, 5º, 7º andares, cave ou rés do chão e tenham, pelo menos, dois quartos, o modelo subestima-os, ou seja, atribui-lhe um valor inferior ao valor real (confiança=0,75; suporte=20).

Por fim, imóveis contruídos até ao ano anterior ao ano corrente e que tenham, no máximo, 71 anos, nos concelhos de Lisboa e Seixal, cuja área não seja superior a 36 m², da freguesia de Corroios, Estrela, Misericórdia, Penha de França ou Santa Maria Maior, o modelo sobrestima-os, ou seja, atribui um valor maior do que o real (confiança=1; suporte=7 imóveis).

⁷ O *rule set* está presente no Anexo F: *Rule set* do modelo para perfis ajustado, subestimado e sobrestimado.

⁸ As regras estão presentes no Anexo G: Regras de perfis de imóveis ajustados, subestimados e sobrestimados.

5. Conclusões e recomendações

5.1. Síntese

Nesta investigação, o objetivo principal é a criação de um modelo preditivo para prever o valor dos imóveis, baseado em técnicas de DM. Este modelo aumenta a qualidade das previsões e, assim, cria uma alternativa para definir o valor de um imóvel. Com um modelo preditivo com boa *performance*, pretende-se resolver o problema da resistência à mudança por parte dos peritos avaliadores e o reconhecimento de que estes modelos podem ser muito úteis nos processos de avaliação.

Assim, definiu-se como objetivo de negócio o aumento da qualidade da estimativa dos valores dos imóveis e a melhoria das estimativas fornecidas aos *stakeholders* e como objetivo de DM a criação de um modelo preditivo com boas métricas na previsão do valor dos imóveis.

Para tal, depois da RL sobre os temas da valorização de imóveis, realizou-se uma análise descritiva aos dados para a construção do modelo, tendo sido realizadas alterações aos mesmos e, até, criações de variáveis, que poderiam ser importantes na explicação do valor dos imóveis (*asking price*).

Após esta fase, para este problema de regressão, construíram-se modelos de algoritmos como a regressão linear, CART, CHAID e as redes neuronais artificiais, em que se utilizaram os métodos *bagging* e *boosting* em todos modelos, exceto na regressão linear. Para validar os dados, recorreu-se ao *holdout* dividindo aleatoriamente os dados em 70% para o conjunto de treino e 30% para o conjunto de teste. Como tal, aplicaram-se, aos modelos obtidos, diversas métricas para os avaliar. No geral obtiveram-se bons modelos, com R^2 entre 0,76 e 0,90. O melhor modelo foi o da rede neuronal com o método *bagging*, com um R^2 igual a 0,90, sendo também o modelo que apresenta melhores métricas quanto aos erros. Depois da seleção do melhor modelo, analisaram-se as variáveis com maior importância na explicação do modelo, onde atributos como o concelho, a freguesia, a área do imóvel, os lugares de estacionamento e o poder de compra *per capita* têm os maiores valores.

Posteriormente, analisou-se o modelo (a árvore de decisão), extraíndo-se regras e conseguindo, assim, dar melhor explicabilidade ao mesmo. Selecionaram-se ainda as cinco variáveis com maior importância na explicação do valor do imóvel, de várias categorias como as características do apartamento, a localização e indicadores económicos, para perceber qual era o seu impacto no valor. Depois deste procedimento, analisaram-se perfis de imóveis em que o modelo acerta, subestima e sobrestima o valor para um melhor entendimento do utilizador.

Após a análise, conclui-se que os objetivos delineados previamente foram alcançados e, assim, obteve-se um modelo com bom desempenho na previsão do valor dos imóveis, assumindo-se como uma alternativa válida na temática da formulação do valor de um imóvel. Em suma, o estudo permite responder à questão de investigação formulada (como prever o valor de venda de imóveis?), já que, com base numa amostra de apartamentos localizados na AML, com valores de venda inferiores a um

milhão de euros, estimou-se o valor do imóvel com elevada qualidade ($R^2=0,87$), recorrendo-se a uma rede neuronal artificial e tendo por variáveis preditoras as características dos imóveis e da sua localização, sendo as características freguesia, concelho, área útil, lugares de estacionamento e poder de compra no concelho as mais relevantes para a explicação do valor do imóvel.

5.2. Contributos

Esta investigação permitiu avanços no estudo da eficácia dos AVM e na sua afirmação como uma ferramenta útil para os peritos em avaliação do valor dos imóveis. Foi alcançado um modelo com boa eficácia na previsão, contribuindo, assim, para o aumento da qualidade da estimativa destes modelos, o que pode ser útil para *stakeholders*, como sejam vendedores, compradores, mediadores, Estado, peritos e bancos. Estes têm nos modelos ferramentas independentes e objetivas para a obtenção dos valores dos imóveis, características que têm sido criticadas no que toca ao desempenho dos peritos.

Além do mais, este estudo é inovador ao utilizar o CRISP-DM, que é, para muitos, a melhor metodologia em problemas de DM. A utilização do CRISP-DM possibilitou a obtenção de vários modelos com bons desempenhos que cumprissem os objetivos delineados na metodologia.

O melhor modelo obtido pode ser integrado num sistema cuja finalidade é o apoio a profissionais que necessitem de avaliar imóveis. Por exemplo, pode ser uma ferramenta que permite a um perito comparar o valor obtido por si com o valor previsto pelo modelo, pode ser uma ferramenta utilizada num banco para determinar o valor do crédito à habitação, pode ser utilizada para investidores com a previsão de valores de carteiras de imóveis, entre muitos outros cenários.

Por outro lado, introduziram-se novos atributos no modelo que não estavam nos modelos reportados na literatura consultada, como sejam as variáveis que descrevem o contexto da população em termos gerais, em escalões etários e em níveis de habilitação, sendo das que mais importância têm na explicação do valor do imóvel. Por exemplo, freguesias que têm uma percentagem de população analfabeta elevada provocam uma desvalorização do imóvel, pois pode indicar que o emprego nessa zona não é tão bem remunerado, provocando uma redução do poder de compra, aplicando-se o raciocínio contrário para uma freguesia em que população com o ensino superior é elevada.

Por fim, esta investigação constitui um avanço no estudo dos modelos e da temática em Portugal, onde não existe, na literatura consultada, um estudo nestes moldes e, assim, servir como uma confirmação da aplicabilidade dos AVM neste país e incentivo para a continuidade do estudo do tema.

5.3. Limitações

Apesar do sucesso dos modelos obtidos, este estudo tem algumas limitações. O facto das leis da proteção de dados terem sido reforçadas nos últimos anos inibe os bancos e as empresas de avaliação

de poderem partilhar dados dos imóveis, o que faz com que apenas se possa extrair informação que é pública, como é o caso do *asking price*, que não corresponde ao valor de avaliação, mas sim o que um vendedor pretende, o que, por vezes, não corresponde ao valor da transação nem de avaliação.

Por se tratar de uma plataforma de venda de imóveis, as pessoas encarregues pela submissão do imóvel na mesma podem inflacionar as características do imóvel avaliando, por exemplo, a área maior, um parâmetro melhor do que a realidade e, assim, inflacionar o valor.

Apesar de uma variável da população ter importância no modelo, os dados só são atualizados de 10 em 10 anos, o que pode ser prejudicial para este modelo para prever valores em anos futuros, pois já podem estar desfasados da realidade e, assim, prever um valor que não corresponde à realidade.

Além do mais, este estudo foi aplicado apenas a apartamentos, e pode não ter o mesmo desempenho quando aplicado a outro tipo de imóveis, em que a valorização pode depender de outras características e à zona da AML, que pode não corresponder às restantes regiões do país, ou de outros países. Por fim, como este modelo não foi implementado na realidade, não é possível conhecer como se adequa no contexto real e se, de facto, é uma ferramenta útil para os *stakeholders*.

5.4. Trabalho Futuro

Esta dissertação abre muitas perspetivas no que diz respeito a investigações futuras que complementem e melhorem este estudo, tentando colmatar as limitações anteriormente descritas.

Para obter o valor de avaliação e dados que correspondem à realidade dos factos, pode-se tentar criar parcerias com bancos ou com empresas responsáveis pela avaliação de imóveis, onde cedem as suas bases de dados de imóveis avaliados para a construção dos modelos. Se tal não for possível, como aconteceu nesta investigação, devem-se tentar extrair dados de outras agências imobiliárias de modo a perceber se um imóvel tem um valor maior por estar presente numa plataforma em concreto.

No que diz respeito a variáveis que possam explicar o valor do imóvel, pode treinar-se o modelo para conseguir avaliar o estado do imóvel através de imagens. Também se pode efetuar uma análise mais aprofundada com a pesquisa de mais estudos onde se explique a obtenção de outras variáveis passíveis de explicar a variável alvo como, por exemplo, a medição de distância a pontos de interesse.

Em investigações futuras podem utilizar-se outros algoritmos com outras parametrizações e, assim, obterem-se modelos com, possivelmente, maior qualidade. Para avaliação dos modelos podem utilizar-se métricas diferentes daquelas que foram usadas nesta investigação. Já no que respeita à localização, podem investigar-se outras regiões do país que não a AML, por exemplo, regiões com valores mais baixos, longe dos grandes centros urbanos, nomeadamente o interior.

Por fim, na implementação do modelo pode realizar-se um *dashboard* onde se introduzem dados sobre o imóvel e obter o valor do mesmo. Esta ferramenta visual é uma opção interessante, dando maior explicabilidade ao modelo, verificando a variação do valor em cada atributo.

6. Bibliografia

- Anysz, H., Podwórna, M., Ibadov, N., Lennerts, K., & Dikarev, K. (2021). Hybrid predictions of the homogenous properties' market value with the use of ann. *Archives of Civil Engineering*, 67(1), 285–301. <https://doi.org/10.24425/ace.2021.136474>
- Arcuri, N., de Ruggiero, M., Salvo, F., & Zinno, R. (2020). Automated valuation methods through the cost approach in a BIM and GIS integration framework for smart city appraisals. *Sustainability (Switzerland)*, 12(18). <https://doi.org/10.3390/su12187546>
- Banco de Portugal. (2006). Aviso nº5/2006. *Diário da República*. <https://www.bportugal.pt/sites/default/files/anexos/cartas-circulares/5-2006a.pdf>
- Banco de Portugal, Comissão do Mercado de Valores Mobiliários, & Instituto de Seguros de Portugal. (2013). *A Avaliação e Valorização - Uma Abordagem Integrada para o Sistema Financeiro Português*. Consulta Pública do CNSF n.º 1/2013. https://www.bportugal.pt/sites/default/files/anexos/documentos-relacionados/consultapublicacnsf_avaliacaovalorizaca_imoveis.pdf
- Bellotti, A. (2017). Reliable region predictions for automated valuation models. *Annals of Mathematics and Artificial Intelligence*, 81(1), 71–84. <https://doi.org/10.1007/s10472-016-9534-6>
- Bogin, A. N., & Shui, J. (2020). Appraisal Accuracy and Automated Valuation Models in Rural Areas. *Journal of Real Estate Finance and Economics*, 60(9), 40–52. <https://doi.org/10.1007/s11146-019-09712-0>
- Caetano, N. (2013). Previsão de tempos de internamento de pacientes via técnicas de Data Mining [Dissertação de Mestrado, Iscte - Instituto Universitário de Lisboa]. Repositório do Iscte.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). *CRISP-DM 1.0 Step-by-step data mining guide*. The CRISP-DM Consortium. <https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf>
- Chmielewska, A., Adamiczka, J., & Romanowski, M. (2020). Genetic Algorithm as Automated Valuation Model Component in Real Estate Investment Decisions System. *Real Estate Management and Valuation*, 28(4), 1–14. <https://doi.org/10.1515/remav-2020-0027>
- Crosby, H., Damoulas, T., Caton, A., Davis, P., Albuquerque, J., & Jarvis, S. (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>
- Das, S., Ali, M., Li, Y., Kang, Y., & 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>
- Demirci, O. (2021). *Automated Valuation Models (AVMs): Machine Learning, namely Mass (Advanced) Valuation Methods and Algorithms*. Property Elite. <https://doi.org/10.13140/RG.2.2.12649.42080>

- Dimopoulos, T., & Bakas, N. (2019). Sensitivity analysis of machine learning models for the mass appraisal of real estate. Case study of residential units in Nicosia, Cyprus. *Remote Sensing*, 11(24), 3047. <https://doi.org/10.3390/rs11243047>
- Doumpos, M., Papastamos, D., Andritsos, D., & Zopounidis, C. (2021). Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches. *Annals of Operations Research*, 306(2), 415–433. <https://doi.org/10.1007/s10479-020-03556-1>
- Epley, D. (2017). The need to reference automatic valuation models to the valuation process. *Journal of Real Estate Literature*, 25(1), 237–251. <https://doi.org/10.1080/10835547.2017.12090450>
- Vaz, J., & Anjos, M. (2022). Normas Internacionais de Avaliação (TEGOVA, RICS e IVSC). *ANAI- Associação Nacional de Avaliadores Imobiliários*. <http://hdl.handle.net/10198/24860>
- Fu, J., Sheel, A., & Lang, J. (2013). A reexamination of current hotel valuation techniques – which approach is more realistic? *Journal of Hospitality Financial Management*, 21(1), 17–30. <https://doi.org/10.1080/10913211.2013.820074>
- Füss, R., D'Silva, A., Oust, A. & Birkeland, K. (2021). The Predictability of House Prices: “Human Against Machine”. *International Real Estate Review*, 24(2), 139–183. <https://doi.org/10.53383/100319>
- Gallin, J., Molloy, R., Nielsen, E., Smith, P., & Sommer, K. (2021). Measuring Aggregate Housing Wealth: New Insights from an Automated Valuation Model. *Journal of Housing Economics*, 51(1), 101734. <https://doi.org/10.17016/FEDS.2018.064>
- Gates, S. (2002). Review of methodology of quantitative reviews using meta-analysis in ecology. *Journal of Animal Ecology*, 71, 547-557. <https://doi.org/10.1046/j.1365-2656.2002.00634.x>
- Glumac, B., & Rosiers, F. (2018). Real estate and land property automated valuation systems: A taxonomy and conceptual model. *Luxembourg Institute of Socio-Economic Research (LISER) Working Paper Series 2018-09*. <https://dx.doi.org/10.2139/ssrn.3183132>
- Gnat, S. (2021). Property mass valuation on small markets. *Land*, 10(4), 388. <https://doi.org/10.3390/land10040388>
- Hinrichs, N., Kolbe, J., & Werwatz, A. (2021). Using shrinkage for data-driven automated valuation model specification—a case study from Berlin. *Journal of Property Research*, 38(2), 130–153. <https://doi.org/10.1080/09599916.2021.1905690>
- Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, 79, 403–408. <https://doi.org/10.1016/j.procir.2019.02.106>
- IAAO. (2018). *Standard on Automated Valuation Models (AVMs)*. IAAO Technical Standards. https://www.iaao.org/media/standards/Standard_on_Automated_Valuation_Models.pdf
- IBM. (2021). *IBM SPSS Modeler 18.2.2 User's Guide*. IBM Support. <https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.2.2/en/ModelerUsersGuide.pdf>

- Kitchenham, B. (2004). *Procedures for performing systematic reviews*. Software Engineering Group, Department of Computer Science, Keele University. <https://www.inf.ufsc.br/~aldo.vw/kitchenham.pdf>
- Kok, N., Koponen, E., & Martínez-Barbosa, C. (2017). Big data in real estate? from manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6), 202–211. <https://doi.org/10.3905/jpm.2017.43.6.202>
- Kolbe, J., Schulz, R., Wersing, M., & Werwatz, A. (2021). Real estate listings and their usefulness for hedonic regressions. *Empirical Economics*, 61(6), 3239–3269. <https://doi.org/10.1007/s00181-020-01992-3>
- Krause, A., Martin, A., & Fix, M. (2020). Uncertainty in automated valuation models: Error-based versus model-based approaches. *Journal of Property Research*, 37(4), 308–339. <https://doi.org/10.1080/09599916.2020.1807587>
- Kruger, S., & Maturana, G. (2021). Collateral misreporting in the residential mortgage-backed security market. *Management Science*, 67(5), 2729–2750. <https://doi.org/10.1287/mnsc.2019.3569>
- Larraz, B., Alfaro-Navarro, J., Cano, E., Alfaro-Cortes, E., Garcia, N., & Gámez, M. (2021). A computer-assisted expert algorithm for real estate valuation in Spanish cities. *Environment and Planning B: Urban Analytics and City Science*, 48(6), 1712–1727. <https://doi.org/10.1177/2399808320947729>
- Laureano, R., Caetano, N. & Cortez, P. (2014). Previsão de tempos de internamento num hospital português: Aplicação da metodologia CRISP-DM. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, 13, 83-98. Doi: <http://dx.doi.org/10.4304/risti.13.83-98>
- Lim, Z., & Bellotti, A. (2021). Normalized nonconformity measures for automated valuation models. *Expert Systems with Applications*, 180, 115165. <https://doi.org/10.1016/j.eswa.2021.115165>
- Locurcio, M., Morano, P., Tajani, F., & Liddo, F. (2020). An innovative GIS-based territorial information tool for the evaluation of corporate properties: An application to the Italian context. *Sustainability (Switzerland)*, 12(14), 5836. <https://doi.org/10.3390/su12145836>
- Manjula, R., Jain, S., Srivastava, S., & Kher, P. (2017). Real estate value prediction using multivariate regression models. *IOP Conference Series: Materials Science and Engineering*, 263(4), 042098. <https://doi.org/10.1088/1757-899X/263/4/042098>
- Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048–3061. <https://doi.org/10.1109/TKDE.2019.2962680>
- Morano, P., Rosato, P., Tajani, F., Manganelli, B., & Liddo, F. (2019). Contextualized property market models vs. Generalized mass appraisals: An innovative approach. *Sustainability (Switzerland)*, 11(18), 4896. <https://doi.org/10.3390/su11184896>
- Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, 1(2), 161–169. <https://doi.org/10.1016/j.joi.2006.12.001>

- Oust, A., Hansen, S., & Pettrem, T. (2020). Combining Property Price Predictions from Repeat Sales and Spatially Enhanced Hedonic Regressions. *Journal of Real Estate Finance and Economics*, 61(2), 183–207. <https://doi.org/10.1007/s11146-019-09723-x>
- Pace, R., & Calabrese, R. (2021). Ignoring Spatial and Spatiotemporal Dependence in the Disturbances Can Make Black Swans Appear Grey. *Journal of Real Estate Finance and Economics*, 65(1), 1-21. <https://doi.org/10.1007/s11146-021-09836-2>
- Pai, P., & Wang, W. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences (Switzerland)*, 10(17), 5832. <https://doi.org/10.3390/app10175832>
- Poursaeed, O., Matera, T., & Belongie, S. (2018). Vision-based real estate price estimation. *Machine Vision and Applications*, 29(4), 667–676. <https://doi.org/10.1007/s00138-018-0922-2>
- Renigier-Biłozor, M., Chmielewska, A., Walacik, M., Janowski, A., & Lepkova, N. (2021). Genetic algorithm application for real estate market analysis in the uncertainty conditions. *Journal of Housing and the Built Environment*, 36(4), 1629–1670. <https://doi.org/10.1007/s10901-020-09815-8>
- Renigier-Biłozor, M., Janowski, A., & d’Amato, M. (2019). Automated Valuation Model based on fuzzy and rough set theory for real estate market with insufficient source data. *Land Use Policy*, 87, 104021. <https://doi.org/10.1016/j.landusepol.2019.104021>
- Renigier-Biłozor, M., Żróbek, S., Walacik, M., Borst, R., Grover, R., & d’Amato, M. (2022). International acceptance of automated modern tools use must-have for sustainable real estate market development. *Land Use Policy*, 113, 105876. <https://doi.org/10.1016/j.landusepol.2021.105876>
- Schröer, C., Kruse, F., & Gómez, J. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>
- Schulz, R., & Wersing, M. (2021). Automated Valuation Services: A case study for Aberdeen in Scotland. *Journal of Property Research*, 38(2), 154–172. <https://doi.org/10.1080/09599916.2020.1861066>
- Schulz, R., Wersing, M., & Werwatz, A. (2014). Automated valuation modelling: a specification exercise. *Journal of Property Research*, 31(2), 131–153. <https://doi.org/10.1080/09599916.2013.846930>
- Shiller, R., & Weiss, A. (1999). Evaluating Real Estate Valuation Systems. *Journal of Real Estate Finance and Economics*, 18(2), 147-161. Kluwer Academic Publishers. <https://doi.org/10.1023/A:1007756607862>
- Sing, T., Yang, J., & Yu, S. (2021). Boosted Tree Ensembles for Artificial Intelligence Based Automated Valuation Models (AI-AVM). *Journal of Real Estate Finance and Economics*, 65, 649–674. <https://doi.org/10.1007/s11146-021-09861-1>
- Steurer, M., Hill, R., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 38(2), 99–129. <https://doi.org/10.1080/09599916.2020.1858937>
- Su, T., Li, H., & An, Y. (2021). A BIM and machine learning integration framework for automated property valuation. *Journal of Building Engineering*, 44, 102636. <https://doi.org/10.1016/j.job.2021.102636>

- Tchuente, D., & Nyawa, S. (2022). Real estate price estimation in French cities using geocoding and machine learning. *Annals of Operations Research*, 308(1–2), 571–608. <https://doi.org/10.1007/s10479-021-03932-5>
- Valier, A. (2020). Who performs better? AVMs vs hedonic models. *Journal of Property Investment and Finance*, 38(3), 213–225. <https://doi.org/10.1108/JPIF-12-2019-0157>
- Walt, K., & Boshoff, D. (2017). An analysis of the use of mass appraisal methods for agricultural properties. *Acta Structilia*, 24(2), 44-76. <https://doi.org/10.18820/24150487/as24i2.2>
- Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, 39(1), 93–112. <https://doi.org/10.1177/0739456X17723971>

7. Anexos

7.1. Anexo A: Dicionário de dados

Tabela 7.1 - Dicionário de dados de variáveis relativas ao imóvel

Variável	Descrição	Tipo
Valor	Valor pretendido pelo imóvel (<i>asking price</i>)	Quantitativa contínua (QC)
Tipo	Tipo de imóvel	Qualitativa nominal (QN)
Area_Imovel	Área, em metros quadrados, do imóvel	QC
Casas_de_Banho	Número de casas de banho do imóvel	Quantitativa discreta (QD)
Quartos	Número de quartos do imóvel	QD
Piso	Piso do imóvel	QN
Distrito	Distrito a que pertence o imóvel	QN
Concelho	Município a que pertence o imóvel	QN
Freguesia	Freguesia a que pertence o imóvel	QN
Ano_Construcao	Ano de construção do imóvel	QO
Eficiencia_Energetica	Eficiência energética do imóvel	Qualitativa ordinal (QO)
Piso	Piso em que se situa o imóvel	QO
Elevador	Elevador presente no imóvel ou no seu prédio	QN
Lug_Estacionamento	Número de lugares de estacionamento afetos ao imóvel	QO

Tabela 7.2 - Dicionário de dados de variáveis relativas à população da freguesia do imóvel

Variável	Descrição	Tipo
Populacao	Número de habitantes da freguesia a que pertence o imóvel, em 2021	QD
Crescimento_Populacao	Varição, do número de habitantes da freguesia a que pertence o imóvel, em 2021, face a 2011	QC
Populacao_0-14	Número de habitantes da freguesia a que pertence o imóvel, em 2021, até aos 14 anos de idade	QD
Populacao_15-24	Número de habitantes da freguesia a que pertence o imóvel, em 2021, entre os 15 e os 24 anos de idade	QD
Populacao_25-64	Número de habitantes da freguesia a que pertence o imóvel, em 2021, entre os 25 e os 64 anos de idade	QD
Populacao_+65	Número de habitantes da freguesia a que pertence o imóvel, em 2021, a partir dos 65 anos de idade	QD
Populacao_sem_escolaridade	Número de habitantes da freguesia a que pertence o imóvel, em 2021, sem qualquer grau de escolaridade	QD
Populacao_1_ciclo	Número de habitantes da freguesia a que pertence o imóvel, em 2021, com o 1º ciclo de escolaridade	QD
Populacao_2_ciclo	Número de habitantes da freguesia a que pertence o imóvel, em 2021, com o 2º ciclo de escolaridade	QD
Populacao_3_ciclo	Número de habitantes da freguesia a que pertence o imóvel, em 2021, com o 3º ciclo de escolaridade	QD
Populacao_sec_e_pos_sec	Número de habitantes da freguesia a que pertence o imóvel, em 2021, com o secundário e pós secundário	QD

Tabela 7.3 - Dicionário de dados de variáveis relativas a características da freguesia do imóvel

Variável	Descrição	Tipo
Alojamentos	Número de alojamentos da freguesia a que pertence o imóvel	QD
Crescimento_Alojamentos	Varição, do número de alojamentos da freguesia a que pertence o imóvel, em 2021, face a 2011	QC
Area_Freguesia	Área, em quilómetros quadrados, da freguesia a que pertence o imóvel	QC
Escolas	Número de escolas da freguesia a que pertence o imóvel	QD
Escolas_Privadas	Número de escolas privadas da freguesia a que pertence o imóvel	QD
Escolas_Publicas	Número de escolas públicas da freguesia a que pertence o imóvel	QD
Hospitais	Número de hospitais da freguesia a que pertence o imóvel	QD
Farmacias	Número de farmácias da freguesia a que pertence o imóvel	QD
Estacoes_Transportes	Número de estações de metro e comboio da freguesia a que pertence o imóvel	QD
Estacoes_Metro	Número de estações de metro da freguesia a que pertence o imóvel	QD
Estacoes_Comboio	Número de estações de comboio da freguesia a que pertence o imóvel	QD
Estabelecimentos_Policiais	Número de estabelecimentos policiais da freguesia a que pertence o imóvel	QD
Estabelecimentos_PSP	Número de estabelecimentos da Polícia de Segurança Pública (PSP) da freguesia a que pertence o imóvel	QD
Estabelecimentos_GNR	Número de estabelecimentos da Guarda Nacional Republicana (GNR) da freguesia a que pertence o imóvel	QD
Praias	Número de praias da freguesia a que pertence o imóvel	QD

Tabela 7.4 - Dicionário de dados de variáveis de indicadores económicos do concelho do imóvel

Variável	Descrição	Tipo
Peso_Poder_Compra	Peso, em percentagem, do poder de compra, em relação ao poder de compra nacional, do município a que pertence o imóvel	QC
Poder_Compra_per_capita	Poder de compra por habitante do município a que pertence o imóvel, medindo o afastamento em relação à média nacional	QC
Valor_Venda_Oferta_m2	Valor de oferta por metro quadrado do município a que pertence o imóvel	QD
Valor_Renda_Pedida_m2	Valor de renda pedida por metro quadrado do município a que pertence o imóvel	QC

7.2. Anexo B: Correlação de variáveis

Tabela 7.5 - Correlação com variáveis do imóvel e da população

Variável	Área do Imóvel	Casas de Banho	Quartos	Ano de Construção	População	Crescimento da População	População até aos 14 anos
Área do Imóvel	1,00	0,738**	0,689**	0,383**	0,082**	0,174**	0,093**
Casas de Banho	0,738**	1,00	0,622**	0,370**	0,048**	0,111**	0,039*
Quartos	0,689**	0,622**	1,00	0,036*	0,152**	0,179**	0,150**
Ano de Construção	0,383**	0,370**	0,036*	1,00	-0,083**	0,036*	-0,056**
População	0,082**	0,048**	0,152**	-0,083**	1,00	0,281**	0,972**
Crescimento da População	0,174**	0,111**	0,179**	0,036*	0,281**	1,00	0,344**
População até aos 14 anos	0,093**	0,039*	0,150**	-0,056**	0,972**	0,344**	1,00
População entre os 15 e os 24 anos	0,071**	0,02	0,140**	-0,078**	0,980**	0,299**	0,980**
População entre os 25 e os 64 anos	0,070**	0,03	0,138**	-0,090**	0,994**	0,288**	0,976**
População a partir dos 65 anos	0,098**	0,101**	0,169**	-0,088**	0,908**	0,175**	0,809**
População Sem Escolaridade	0,01	-0,042**	0,118**	-0,095**	0,909**	0,207**	0,915**
População com o 1º Ciclo	-0,01	-0,049**	0,115**	-0,090**	0,843**	0,161**	0,840**
População com o 2º Ciclo	0,01	-0,050**	0,112**	-0,086**	0,898**	0,233**	0,925**
População com o 3º Ciclo	0,02	-0,03	0,122**	-0,103**	0,937**	0,239**	0,940**
População com o Secundário e o Pós Secundário	0,039*	-0,01	0,123**	-0,106**	0,960**	0,273**	0,956**
População com o Ensino Superior	0,180**	0,198**	0,150**	-0,063**	0,665**	0,230**	0,576**
Alojamentos	0,079**	0,065**	0,134**	-0,098**	0,964**	0,228**	0,893**
Crescimento dos Alojamentos	0,153**	0,084**	0,182**	0,135**	0,322**	0,798**	0,368**
Área da Freguesia	0,083**	0,03	0,096**	0,021	0,01	0,261**	0,053**
Número de Escolas	0,097**	0,083**	0,115**	-0,041**	0,808**	0,264**	0,762**
Escolas Privadas	0,110**	0,107**	0,088**	-0,043**	0,712**	0,242**	0,653**
Escolas Públicas	0,054**	0,03	0,133**	-0,058**	0,786**	0,239**	0,772**
Hospitais	0,064**	0,096**	0,00	0,029	-0,037*	0,108**	-0,111**
Farmácias	0,063**	0,092**	0,01	-0,022	0,404**	0,03	0,317**
Estações de Transportes	0,02	0,060**	-0,035*	0,007	0,173**	-0,090**	0,068**
Estações de Metro	-0,01	0,02	-0,046**	0,013	-0,089**	-0,069**	-0,153**
Estações de Comboio	0,043**	0,066**	0,01	0,008	0,417**	-0,042**	0,343**
Estabelecimentos Policiais	-0,049**	-0,01	-0,090**	-0,048**	0,03	-0,270**	0,02
Estabelecimentos PSP	-0,048**	0,001	-0,079**	-0,069**	0,060**	-0,256**	0,025
Estabelecimentos GNR	-0,006	-0,029	-0,041**	0,062**	-0,108**	-0,060**	-0,033*
Praias	0,03	0,041**	0,008	0,009	0,204**	0,110**	0,152**
Peso do Poder de Compra	-0,063**	-0,001	-0,141**	-0,046**	-0,360**	-0,317**	-0,422**
Poder de Compra per capita	-0,001	0,070**	-0,100**	0,034*	-0,377**	-0,307**	-0,454**
Valor de Oferta de Venda por m ²	-0,001	0,069**	-0,112**	0,03	-0,288**	-0,270**	-0,357**
Valor de Renda Pedida por m ²	-0,009	0,075**	-0,099**	-0,008	-0,190**	-0,272**	-0,274**

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Tabela 7.6 - Correlação com variáveis da população

Variável	População entre os 15 e os 24 anos	População entre os 25 e os 64 anos	População a partir dos 65 anos	População Sem Escolaridade	População com o 1º Ciclo	População com o 2º Ciclo
Área do Imóvel	0,071**	0,070**	0,098**	0,01	-0,01	0,01
Casas de Banho	0,02	0,03	0,101**	-0,042**	-0,049**	-0,050**
Quartos	0,140**	0,138**	0,169**	0,118**	0,115**	0,112**
Ano de Construção	-0,078**	-0,090**	-0,088**	-0,095**	-0,090**	-0,086**
População	0,980**	0,994**	0,908**	0,909**	0,843**	0,898**
Crescimento da População	0,299**	0,288**	0,175**	0,207**	0,161**	0,233**
População até aos 14 anos	0,980**	0,976**	0,809**	0,915**	0,840**	0,925**
População entre os 15 e os 24 anos	1,00	0,985**	0,825**	0,920**	0,833**	0,930**
População entre os 25 e os 64 anos	0,985**	1,00	0,862**	0,913**	0,839**	0,909**
População a partir dos 65 anos	0,825**	0,862**	1,00	0,769**	0,745**	0,718**
População Sem Escolaridade	0,920**	0,913**	0,769**	1,00	0,974**	0,992**
População com o 1º Ciclo	0,833**	0,839**	0,745**	0,974**	1,00	0,952**
População com o 2º Ciclo	0,930**	0,909**	0,718**	0,992**	0,952**	1,00
População com o 3º Ciclo	0,957**	0,943**	0,783**	0,984**	0,922**	0,984**
População com o Secundário e o Pós Secundário	0,973**	0,970**	0,805**	0,946**	0,864**	0,946**
População com o Ensino Superior	0,582**	0,633**	0,752**	0,306**	0,225**	0,281**
Alojamentos	0,908**	0,945**	0,953**	0,813**	0,754**	0,785**
Crescimento dos Alojamentos	0,331**	0,309**	0,273**	0,307**	0,303**	0,314**
Área da Freguesia	0,034*	0,01	-0,03	0,048**	0,049**	0,083**
Número de Escolas	0,757**	0,779**	0,823**	0,637**	0,576**	0,622**
Escolas Privadas	0,650**	0,683**	0,752**	0,458**	0,378**	0,445**
Escolas Públicas	0,765**	0,762**	0,758**	0,785**	0,767**	0,771**
Hospitais	-0,095**	-0,054**	0,086**	-0,225**	-0,227**	-0,247**
Farmácias	0,314**	0,391**	0,484**	0,160**	0,132**	0,129**
Estações de Transportes	0,117**	0,148**	0,305**	0,01	0,00	-0,01
Estações de Metro	-0,130**	-0,094**	0,00	-0,189**	-0,152**	-0,202**
Estações de Comboio	0,388**	0,384**	0,500**	0,305**	0,236**	0,286**
Estabelecimentos Policiais	0,02	0,03	0,042**	-0,063**	-0,100**	-0,053**
Estabelecimentos PSP	0,039*	0,055**	0,099**	-0,061**	-0,099**	-0,060**
Estabelecimentos GNR	-0,070**	-0,092**	-0,200**	-0,008	-0,008	0,024
Praias	0,160**	0,164**	0,327**	0,056**	0,003	0,047**
Peso do Poder de Compra	-0,380**	-0,347**	-0,291**	-0,499**	-0,487**	-0,496**
Poder de Compra per capita	-0,434**	-0,385**	-0,228**	-0,564**	-0,538**	-0,575**
Valor de Oferta de Venda por m²	-0,343**	-0,299**	-0,150**	-0,511**	-0,498**	-0,516**
Valor de Renda Pedida por m²	-0,257**	-0,210**	-0,029	-0,420**	-0,405**	-0,437**

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Tabela 7.7 - Correlação com variáveis da população e características da freguesia

Variável	População com o 3º Ciclo	População com o Secundário e o Pós Secundário	População com o Ensino Superior	Alojamentos	Crescimento dos Alojamentos	Área da Freguesia
Área do Imóvel	0,02	0,039*	0,180**	0,079**	0,153**	0,083**
Casas de Banho	-0,03	-0,01	0,198**	0,065**	0,084**	0,025
Quartos	0,122**	0,123**	0,150**	0,134**	0,182**	0,096**
Ano de Construção	-0,103**	-0,106**	-0,063**	-0,098**	0,135**	0,021
População	0,937**	0,960**	0,665**	0,964**	0,322**	0,008
Crescimento da População	0,239**	0,273**	0,230**	0,228**	0,798**	0,261**
População até aos 14 anos	0,940**	0,956**	0,576**	0,893**	0,368**	0,053**
População entre os 15 e os 24 anos	0,957**	0,973**	0,582**	0,908**	0,331**	0,034*
População entre os 25 e os 64 anos	0,943**	0,970**	0,633**	0,945**	0,309**	0,005
População a partir dos 65 anos	0,783**	0,805**	0,752**	0,953**	0,273**	-0,03
População Sem Escolaridade	0,984**	0,946**	0,306**	0,813**	0,307**	0,048**
População com o 1º Ciclo	0,922**	0,864**	0,225**	0,754**	0,303**	0,049**
População com o 2º Ciclo	0,984**	0,946**	0,281**	0,785**	0,314**	0,083**
População com o 3º Ciclo	1,00	0,982**	0,373**	0,845**	0,307**	0,068**
População com o Secundário e o Pós Secundário	0,982**	1,00	0,475**	0,886**	0,300**	0,052**
População com o Ensino Superior	0,373**	0,475**	1	0,765**	0,196**	-0,111**
Alojamentos	0,845**	0,886**	0,765**	1	0,245**	-0,024
Crescimento dos Alojamentos	0,307**	0,300**	0,196**	0,245**	1	0,202**
Área da Freguesia	0,068**	0,052**	-0,111**	-0,024	0,202**	1
Número de Escolas	0,673**	0,714**	0,720**	0,858**	0,230**	0,044**
Escolas Privadas	0,517**	0,592**	0,805**	0,803**	0,149**	-0,025
Escolas Públicas	0,777**	0,755**	0,425**	0,759**	0,308**	0,145**
Hospitais	-0,210**	-0,176**	0,358**	0,088**	0,002	-0,046**
Farmácias	0,184**	0,240**	0,681**	0,516**	-0,111**	-0,192**
Estações de Transportes	0,02	0,047**	0,425**	0,293**	-0,198**	-0,176**
Estações de Metro	-0,214**	-0,209**	0,191**	-0,005	-0,162**	-0,219**
Estações de Comboio	0,360**	0,394**	0,402**	0,485**	-0,077**	0,045**
Estabelecimentos Policiais	-0,037*	-0,02	0,181**	0,075**	-0,459**	-0,052**
Estabelecimentos PSP	-0,035*	-0,008	0,258**	0,114**	-0,412**	-0,133**
Estabelecimentos GNR	-0,009	-0,040*	-0,266**	-0,137**	-0,182**	0,287**
Praias	0,106**	0,164**	0,341**	0,343**	0,039*	0,105**
Peso do Poder de Compra	-0,506**	-0,486**	0,115**	-0,260**	-0,407**	-0,322**
Poder de Compra per capita	-0,568**	-0,545**	0,208**	-0,249**	-0,389**	-0,287**
Valor de Oferta de Venda por m²	-0,507**	-0,459**	0,299**	-0,155**	-0,342**	-0,278**
Valor de Renda Pedida por m²	-0,422**	-0,370**	0,370**	-0,053**	-0,333**	-0,360**

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Tabela 7.8 - Correlação com variáveis de escolas, hospitais, farmácias e transportes

Variável	Escolas	Escolas Privadas	Escolas Públicas	Hospitais	Farmácias	Estações de Transportes
Área do Imóvel	0,097**	0,110**	0,054**	0,064**	0,063**	0,016
Casas de Banho	0,083**	0,107**	0,027	0,096**	0,092**	0,060**
Quartos	0,115**	0,088**	0,133**	0,002	0,01	-0,035*
Ano de Construção	-0,041**	-0,043**	-0,058**	0,029	-0,022	0,007
População	0,808**	0,712**	0,786**	-0,037*	0,404**	0,173**
Crescimento da População	0,264**	0,242**	0,239**	0,108**	0,029	-0,090**
População até aos 14 anos	0,762**	0,653**	0,772**	-0,111**	0,317**	0,068**
População entre os 15 e os 24 anos	0,757**	0,650**	0,765**	-0,095**	0,314**	0,117**
População entre os 25 e os 64 anos	0,779**	0,683**	0,762**	-0,054**	0,391**	0,148**
População a partir dos 65 anos	0,823**	0,752**	0,758**	0,086**	0,484**	0,305**
População Sem Escolaridade	0,637**	0,458**	0,785**	-0,225**	0,160**	0,011
População com o 1º Ciclo	0,576**	0,378**	0,767**	-0,227**	0,132**	0,004
População com o 2º Ciclo	0,622**	0,445**	0,771**	-0,247**	0,129**	-0,013
População com o 3º Ciclo	0,673**	0,517**	0,777**	-0,210**	0,184**	0,022
População com o Secundário e o Pós Secundário	0,714**	0,592**	0,755**	-0,176**	0,240**	0,047**
População com o Ensino Superior	0,720**	0,805**	0,425**	0,358**	0,681**	0,425**
Alojamentos	0,858**	0,803**	0,759**	0,088**	0,516**	0,293**
Crescimento dos Alojamentos	0,230**	0,149**	0,308**	0,002	-0,111**	-0,198**
Área da Freguesia	0,044**	-0,025	0,145**	-0,046**	-0,192**	-0,176**
Número de Escolas	1	0,949**	0,862**	0,188**	0,557**	0,421**
Escolas Privadas	0,949**	1	0,658**	0,278**	0,615**	0,481**
Escolas Públicas	0,862**	0,658**	1	0	0,340**	0,233**
Hospitais	0,188**	0,278**	0	1	0,490**	0,518**
Farmácias	0,557**	0,615**	0,340**	0,490**	1	0,640**
Estações de Transportes	0,421**	0,481**	0,233**	0,518**	0,640**	1
Estações de Metro	0,095**	0,162**	-0,037*	0,548**	0,539**	0,800**
Estações de Comboio	0,543**	0,537**	0,436**	0,013	0,224**	0,415**
Estabelecimentos Policiais	0,155**	0,213**	0,028	0,197**	0,340**	0,226**
Estabelecimentos PSP	0,171**	0,245**	0,013	0,219**	0,386**	0,291**
Estabelecimentos GNR	-0,049**	-0,108**	0,056**	-0,071**	-0,150**	-0,221**
Praias	0,437**	0,497**	0,248**	0,038*	0,061**	0,156**
Peso do Poder de Compra	-0,217**	-0,058**	-0,427**	0,367**	0,283**	0,344**
Poder de Compra per capita	-0,184**	-0,013	-0,420**	0,409**	0,347**	0,359**
Valor de Oferta de Venda por m ²	-0,083**	0,102**	-0,362**	0,352**	0,331**	0,341**
Valor de Renda Pedida por m ²	0,003	0,180**	-0,282**	0,342**	0,375**	0,383**

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Tabela 7.9 - Correlação com variáveis de transportes, estabelecimentos policiais e praias

Variável	Estações de Metro	Estações de Comboio	Estabelecimentos Policiais	Estabelecimentos da PSP	Estabelecimentos da GNR	Praias
Área do Imóvel	-0,011	0,043**	-0,049**	-0,048**	-0,006	0,03
Casas de Banho	0,021	0,066**	-0,007	0,001	-0,029	0,041**
Quartos	-0,046**	0,013	-0,090**	-0,079**	-0,041**	0,008
Ano de Construção	0,013	0,008	-0,048**	-0,069**	0,062**	0,009
População	-0,089**	0,417**	0,029	0,060**	-0,108**	0,204**
Crescimento da População	-0,069**	-0,042**	-0,270**	-0,256**	-0,060**	0,110**
População até aos 14 anos	-0,153**	0,343**	0,015	0,025	-0,033*	0,152**
População entre os 15 e os 24 anos	-0,130**	0,388**	0,019	0,039*	-0,070**	0,160**
População entre os 25 e os 64 anos	-0,094**	0,384**	0,029	0,055**	-0,092**	0,164**
População a partir dos 65 anos	-0,002	0,500**	0,042**	0,099**	-0,200**	0,327**
População Sem Escolaridade	-0,189**	0,305**	-0,063**	-0,061**	-0,008	0,056**
População com o 1º Ciclo	-0,152**	0,236**	-0,100**	-0,099**	-0,008	0,003
População com o 2º Ciclo	-0,202**	0,286**	-0,053**	-0,060**	0,024	0,047**
População com o 3º Ciclo	-0,214**	0,360**	-0,037*	-0,035*	-0,009	0,106**
População com o Secundário e o Pós Secundário	-0,209**	0,394**	-0,019	-0,008	-0,040*	0,164**
População com o Ensino Superior	0,191**	0,402**	0,181**	0,258**	-0,266**	0,341**
Alojamentos	-0,005	0,485**	0,075**	0,114**	-0,137**	0,343**
Crescimento dos Alojamentos	-0,162**	-0,077**	-0,459**	-0,412**	-0,182**	0,039*
Área da Freguesia	-0,219**	0,045**	-0,052**	-0,133**	0,287**	0,105**
Número de Escolas	0,095**	0,543**	0,155**	0,171**	-0,049**	0,437**
Escolas Privadas	0,162**	0,537**	0,213**	0,245**	-0,108**	0,497**
Escolas Públicas	-0,037*	0,436**	0,028	0,013	0,056**	0,248**
Hospitais	0,548**	0,013	0,197**	0,219**	-0,071**	0,038*
Farmácias	0,539**	0,224**	0,340**	0,386**	-0,150**	0,061**
Estações de Transportes	0,800**	0,415**	0,226**	0,291**	-0,221**	0,156**
Estações de Metro	1	-0,214**	0,090**	0,139**	-0,171**	-
Estações de Comboio	-0,214**	1	0,232**	0,263**	-0,101**	0,139**
Estabelecimentos Policiais	0,090**	0,232**	1	0,961**	0,173**	0,465**
Estabelecimentos PSP	0,139**	0,263**	0,961**	1	-0,107**	0,134**
Estabelecimentos GNR	-0,171**	-0,101**	0,173**	-0,107**	1	0,025
Praias	-0,139**	0,465**	0,140**	0,134**	0,025	1
Peso do Poder de Compra	0,434**	-0,098**	0,283**	0,341**	-0,194**	-
Poder de Compra per capita	0,424**	-0,058**	0,276**	0,331**	-0,185**	-
Valor de Oferta de Venda por m ²	0,374**	-0,012	0,297**	0,349**	-0,172**	0,064**
Valor de Renda Pedida por m ²	0,386**	0,039*	0,306**	0,378**	-0,243**	0,114**

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

Tabela 7.10 - Correlação com variáveis económicas

Variável	Peso do Poder de Compra	Poder de Compra per capita	Valor de Oferta de Venda por m ²	Valor de Renda Pedida por m ²
Área do Imóvel	-0,063**	-0,001	-0,001	-0,009
Casas de Banho	-0,001	0,070**	0,069**	0,075**
Quartos	-0,141**	-0,100**	-0,112**	-0,099**
Ano de Construção	-0,046**	0,034*	0,03	-0,008
População	-0,360**	-0,377**	-0,288**	-0,190**
Crescimento da População	-0,317**	-0,307**	-0,270**	-0,272**
População até aos 14 anos	-0,422**	-0,454**	-0,357**	-0,274**
População entre os 15 e os 24 anos	-0,380**	-0,434**	-0,343**	-0,257**
População entre os 25 e os 64 anos	-0,347**	-0,385**	-0,299**	-0,210**
População a partir dos 65 anos	-0,291**	-0,228**	-0,150**	-0,029
População Sem Escolaridade	-0,499**	-0,564**	-0,511**	-0,420**
População com o 1º Ciclo	-0,487**	-0,538**	-0,498**	-0,405**
População com o 2º Ciclo	-0,496**	-0,575**	-0,516**	-0,437**
População com o 3º Ciclo	-0,506**	-0,568**	-0,507**	-0,422**
População com o Secundário e o Pós Secundário	-0,486**	-0,545**	-0,459**	-0,370**
População com o Ensino Superior	0,115**	0,208**	0,299**	0,370**
Alojamentos	-0,260**	-0,249**	-0,155**	-0,053**
Crescimento dos Alojamentos	-0,407**	-0,389**	-0,342**	-0,333**
Área da Freguesia	-0,322**	-0,287**	-0,278**	-0,360**
Número de Escolas	-0,217**	-0,184**	-0,083**	0,003
Escolas Privadas	-0,058**	-0,013	0,102**	0,180**
Escolas Públicas	-0,427**	-0,420**	-0,362**	-0,282**
Hospitais	0,367**	0,409**	0,352**	0,342**
Farmácias	0,283**	0,347**	0,331**	0,375**
Estações de Transportes	0,344**	0,359**	0,341**	0,383**
Estações de Metro	0,434**	0,424**	0,374**	0,386**
Estações de Comboio	-0,098**	-0,058**	-0,012	0,039*
Estabelecimentos Policiais	0,283**	0,276**	0,297**	0,306**
Estabelecimentos PSP	0,341**	0,331**	0,349**	0,378**
Estabelecimentos GNR	-0,194**	-0,185**	-0,172**	-0,243**
Praias	-0,166**	-0,096**	0,064**	0,114**
Peso do Poder de Compra	1	0,941**	0,899**	0,841**
Poder de Compra per capita	0,941**	1	0,937**	0,882**
Valor de Oferta de Venda por m ²	0,899**	0,937**	1	0,965**
Valor de Renda Pedida por m ²	0,841**	0,882**	0,965**	1

Notas: *: a correlação de Pearson ou de Spearman é significativa no nível 0,01 (2 extremidades).

7.3. Anexo C: Equação da Regressão

$S_O_Area_Imovel \times 0,001693 + S_O_Casas_de_Banho \times 0,05594 + S_O_Quartos \times 0,02496 +$
 $Idade_Imovel \times (-0,001732) + S_O_Crescimento_Alojamentos \times (-0,006969) +$
 $Estabelecimentos_Policiais \times 0,006211 + Praias \times 0,005232 + Poder_Compra_per_capita \times 0,001698 +$
 $S_O_WC_por_m2 \times (-2,102) + Percentagem_Populacao_ens_sup \times 0,005254 + Concelho_Cascais \times$
 $0,1048 + Concelho_Loures \times 0,06358 + Concelho_Moita \times (-0,1047) + Concelho_Montijo \times (-0,033) +$
 $Concelho_Odivelas \times 0,09922 + Concelho_Palmela \times (-0,08326) + Concelho_Setúbal \times (-0,1042) +$
 $Concelho_Vila Franca de Xira \times (-0,05165) + Nova_Eficiencia_Energetica_A \times 0,04471 +$
 $Nova_Eficiencia_Energetica_B \times 0,02605 + Nova_Eficiencia_Energetica_D \times 0,01091 +$
 $Novo_Lug_Estacionamento_0 \times (-0,06784) + Novo_Lug_Estacionamento_2 \times 0,01124 + Piso_1 \times$
 $0,01325 + Piso_5 \times 0,01342 + Piso_C/V \times (-0,03626) + Piso_R/C \times (-0,01451) + Freguesia_Alcochete \times$
 $0,08106 + Freguesia_Alcântara \times 0,04873 + Freguesia_Algueirão-Mem Martins \times (-0,04752) +$
 $Freguesia_Avenidas Novas \times 0,05572 + Freguesia_Barreiro e Lavradio \times (-0,09194) + Freguesia_Benfica$
 $\times (-0,05534) + Freguesia_Camarate, Unhos e Apelação \times (-0,08717) + Freguesia_Campo de Ourique \times$
 $0,08183 + Freguesia_Campolide \times 0,1061 + Freguesia_Caparica e Trafaria \times (-0,09068) +$
 $Freguesia_Carnide \times (-0,0839) + Freguesia_Carvoeira \times 0,1579 + Freguesia_Castanheira do Ribatejo e$
 $Cachoeiras \times (-0,05121) + Freguesia_Charneca de Caparica e Sobreda \times 0,04326 + Freguesia_Encosta$
 $do Sol \times 0,05618 + Freguesia_Ericeira \times 0,1191 + Freguesia_Estrela \times 0,03186 + Freguesia_Lumiar \times (-$
 $0,1094) + Freguesia_Marvila \times 0,07191 + Freguesia_Misericórdia \times 0,04591 + Freguesia_Penha de$
 $França \times 0,02322 + Freguesia_Póvoa de Santa Iria e Forte da Casa \times 0,04694 + Freguesia_Rio de Mouro$
 $\times (-0,02941) + Freguesia_Santa Clara \times (-0,06311) + Freguesia_Santiago (Sesimbra) \times 0,3681 +$
 $Freguesia_Santo António \times 0,07831 + Freguesia_Santo António da Charneca \times (-0,1846) +$
 $Freguesia_Santo António dos Cavaleiros e Frielas \times (-0,05489) + Freguesia_Santo Isidoro \times 0,4401 +$
 $Freguesia_Sintra (Santa Maria e São Miguel, São Martinho e São Pedro de Penaferrim) \times 0,1065 +$
 $Freguesia_São Domingos de Benfica \times (-0,0431) + Freguesia_São João das Lampas e Terrugem \times 0,2989$
 $+ Freguesia_Venteira \times 0,03077 + Elevador_Flag \times 0,03457 + Tem_Farmacia_Flag \times 0,03015 + 4,788$

7.4. Anexo D: Regras das árvores

Modelo CART:

Imóveis com valor sobrestimado:

- Se (Concelho = "Alcochete" ou Concelho = "Cascais" ou Concelho = "Lisboa" ou Concelho = "Mafra" ou Concelho = "Montijo" ou Concelho = "Oeiras" ou Concelho = "Sesimbra"), e Percentagem_Populacao_ens_sup > 30,600, e Area_Imovel > 120,500, e Casas_de_Banho > 2,500, então Log_10_Valor = 5,846 (Nó 42: suporte = 38 imóveis)

Imóveis com boa qualidade de estimativa:

- Se (Concelho = "Alcochete" ou Concelho = "Cascais" ou Concelho = "Lisboa" ou Concelho = "Mafra" ou Concelho = "Montijo" ou Concelho = "Oeiras" ou Concelho = "Sesimbra"), e Area_Imovel <= 96,500, e Idade_Imovel > 19,500, e Area_Imovel > 47,500, e Percentagem_Populacao_ens_sup > 26,620, então Log_10_Valor = 5,462 (Nó 34: suporte = 225 imóveis)

Imóveis com valor subestimado:

- Se (Concelho = "Barreiro" Concelho = "Moita"), e Novo_Lug_Estacionamento = 0, e Area_Imovel <= 82,500, então Log_10_Valor = 4,980 (Nó 23: suporte = 32 imóveis)

Modelo CHAID:

Imóveis com valor sobrestimado:

- Se (Concelho="Alcochete" ou Concelho="Cascais" ou Concelho="Lisboa" ou Concelho="Mafra" ou Concelho="Oeiras" ou Concelho="Sesimbra"), e Casas_de_Banho > 2,500, e Area_Imovel > 155,500 então Log_10_Valor = 5,878 (Nó 22: suporte = 97 imóveis)

Imóveis com boa qualidade de estimativa:

- Se (Concelho="Alcochete" ou Concelho="Cascais" ou Concelho="Lisboa" ou Concelho="Mafra" ou Concelho="Oeiras" ou Concelho="Sesimbra"), e (Area_Imovel > 58,500 e Area_Imovel <= 96,500), e Idade_Imovel <= 19,500, e Populacao > 23455,400, então Log_10_Valor = 5,575 (Nó 34: suporte = 39 imóveis)

Imóveis com valor subestimado:

- Se (Concelho="Barreiro" ou Concelho="Moita" ou Concelho="Setúbal" ou Concelho="Vila Franca de Xira"), e Area_Imovel <= 83,500, e Percentagem_Populacao_ens_sup <= 15,105, e (Freguesia="Alhos Vedros" ou Freguesia=" Alto do Seixalinho, Santo André e Verderena" Freguesia=" Baixa da Banheira e Vale da Amoreira" ou Freguesia=" Moita" ou Freguesia=" São Sebastião" ou Freguesia=" Vialonga"), então Log_10_Valor = 4,998 (Nó 44: suporte = 48 imóveis)

7.5. Anexo E: Rule set do melhor modelo

```

Concelho in ["Alcochete" "Cascais" "Lisboa" "Oeiras" "Sesimbra"] [ Ave: 5,613, Effect: 0,17 ] (932)
  Area_Imovel <= 100,500 [Ave: 5,492, Effect: -0,121] (526)
    Idade_Imovel <= 19,500 [Ave: 5,621, Effect: 0,128] (191)
      Area_Imovel <= 58,500 [Ave: 5,477, Effect: -0,144] (55)
        Freguesia in ["Santo António"] [ Ave: 5,679, Effect: 0,202] => 5,679 (8)
          Freguesia in ["Alcântara" "Alvalade" "Arroios" "Belém" "Campolide"
            "Carcavelos e Parede" "Estrela" "Lumiar" "Marvila" "Misericórdia"
            "Oeiras e São Julião da Barra, Paço de Arcos e Caxias" "Olivais" "Santa
            Maria Maior" "Santiago (Sesimbra)" "São Domingos de Benfica" "São
            Vicente"] [Ave: 5,443, Effect: -0,034] => 5,443 (47)
        Area_Imovel > 58,500 [Ave: 5,679, Effect: 0,058] (136)
          Populacao <= 23455,500 [Ave: 5,724, Effect: 0,046] (94)
            Freguesia in ["Ajuda" "Beato" "Porto Salvo"] [Ave: 5,438,
              Effect: -0,286] => 5,438 (5)
            Freguesia in ["Alcochete" "Alcântara" "Avenidas Novas"
              "Belém" "Campo de Ourique" "Campolide" "Carnide" "Estrela"
              "Misericórdia" "Parque das Nações" "Santa Maria Maior"
              "Santiago (Sesimbra)" "Santo António" "São Vicente"] [Ave:
              5,74, Effect: 0,016] => 5,74 (89)
          Populacao > 23455,500 [Ave: 5,577, Effect: -0,102 ] (42)
            Freguesia in ["Quinta do Conde" "Santa Clara"] [Ave: 5,195,
              Effect: -0,382] => 5,195 (2)
            Freguesia in ["Algés, Linda-a-Velha e Cruz Quebrada-Dafundo"
              "Alvalade" "Arroios" "Carcavelos e Parede" "Carnaxide e
              Queijas" "Cascais e Estoril" "Marvila" "Olivais" "Penha de
              França" "São Domingos de Rana"] [Ave: 5,596, Effect: 0,019]
              => 5,596 (40)
    Idade_Imovel > 19,500 [Ave: 5,419, Effect: -0,073] (335)
      Area_Imovel <= 58,500 [Ave: 5,313, Effect: -0,106] => 5,313 (82)
      Area_Imovel > 58,500 [Ave: 5,453, Effect: 0,034] (253)
        Tem_Estacao_Transporte in [ "Não" ] [Ave: 5,367, Effect: -0,086 ] (80)
  
```

Crescimento_Populacao <= 5,450 [Ave: 5,389, Effect: 0,022]
 => 5,389 (73)
 Crescimento_Populacao > 5,450 [Ave: 5,139, Effect: -0,229]
 => 5,139 (7)
 Tem_Estacao_Transporte in ["Sim"] [Ave: 5,493, Effect: 0,04] => 5,493
 (173)
 Area_Imovel > 100,500 [Ave: 5,77, Effect: 0,157] (406)
 Area_Imovel <= 135,500 [Ave: 5,693, Effect: -0,077] (194)
 Idade_Imovel <= 11 [Ave: 5,803, Effect: 0,11] (73)
 Freguesia in ["Areeiro" "Avenidas Novas" "Beato" "Campo de Ourique"
 "Campolide" "Estrela" "Marvila" "Misericórdia" "Parque das Nações"
 "Santa Maria Maior" "Santo António"] [Ave: 5,882, Effect: 0,079] =>
 5,882 (41)
 Freguesia in ["Alcochete" "Alcântara" "Algés, Linda-a-Velha e Cruz
 Quebrada-Dafundo" "Arroios" "Carcavelos e Parede" "Carnide"
 "Lumiar" "Oeiras e São Julião da Barra, Paço de Arcos e Caxias"
 "Olivais" "Porto Salvo" "São Domingos de Benfica" "São Domingos de
 Rana" "São Vicente"] [Ave: 5,701, Effect: -0,102] => 5,701 (32)
 Idade_Imovel > 11 [Ave: 5,627, Effect: -0,066] (121)
 Percentagem_Populacao_ens_sup <= 31,125 [Ave: 5,511, Effect:
 -0,116] => 5,511 (40)
 Percentagem_Populacao_ens_sup > 31,125 [Ave: 5,684, Effect:
 0,057] (81)
 Freguesia in ["Alcântara" "Alvalade" "Areeiro" "Avenidas
 Novas" "Campolide" "Cascais e Estoril" "Estrela" "Parque das
 Nações" "Santo António"] [Ave: 5,752, Effect: 0,068] => 5,752
 (31)
 Freguesia in ["Algés, Linda-a-Velha e Cruz Quebrada-Dafundo"
 "Arroios" "Belém" "Benfica" "Campo de Ourique" "Carcavelos
 e Parede" "Carnide" "Lumiar" "Misericórdia" "Oeiras e São
 Julião da Barra, Paço de Arcos e Caxias" "São Domingos de
 Benfica"] [Ave: 5,642, Effect: -0,042] => 5,642 (50)
 Area_Imovel > 135,500 [Ave: 5,84, Effect: 0,07] (212)
 Area_Imovel <= 168,500 [Ave: 5,807, Effect: -0,033] (118)
 Percentagem_Populacao_ens_sup <= 30,600 [Ave: 5,704, Effect:
 -0,103] => 5,704 (17)
 Percentagem_Populacao_ens_sup > 30,600 [Ave: 5,824, Effect:
 0,017] (101)
 Freguesia in ["Algés, Linda-a-Velha e Cruz Quebrada-Dafundo"
 "Arroios" "Benfica" "Campo de Ourique" "Carnide" "Cascais e
 Estoril" "Lumiar" "Oeiras e São Julião da Barra, Paço de Arcos
 e Caxias" "Parque das Nações" "São Domingos de Benfica"]
 [Ave: 5,788, Effect: -0,037] => 5,788 (62)
 Freguesia in ["Alvalade" "Areeiro" "Avenidas Novas" "Belém"
 "Carcavelos e Parede" "Estrela" "Santo António" "São
 Vicente"] [Ave: 5,883, Effect: 0,058] => 5,883 (39)
 Area_Imovel > 168,500 [Ave: 5,881, Effect: 0,041] => 5,881 (94)
 Concelho in ["Almada" "Amadora" "Barreiro" "Loures" "Mafra" "Moita" "Montijo" "Odivelas"
 "Palmela" "Seixal" "Setúbal" "Sintra" "Vila Franca de Xira"] [Ave: 5,296, Effect: -0,147] (1.078)
 Novo_Lug_Estacionamento in ["0"] [Ave: 5,2, Effect: -0,096] (711)
 Area_Imovel <= 83,500 [Ave: 5,137, Effect: -0,063] (389)

Concelho in ["Barreiro" "Moita" "Montijo" "Setúbal"] [Ave: 5,034, Effect: -0,103] (64)
 Percentagem_Populacao_ens_sup <= 16,945 [Ave: 4,987, Effect: -0,047] => 4,987 (42)
 Percentagem_Populacao_ens_sup > 16,945 [Ave: 5,123, Effect: 0,089] => 5,123 (22)
 Concelho in ["Almada" "Amadora" "Loures" "Mafra" "Odivelas" "Palmela" "Seixal" "Sintra" "Vila Franca de Xira"] [Ave: 5,157, Effect: 0,02] (325)
 Area_Imovel <= 68,500 [Ave: 5,111, Effect: -0,046] (155)
 Quartos <= 1,500 [Ave: 5,069, Effect: -0,043] => 5,069 (73)
 Quartos > 1,500 [Ave: 5,149, Effect: 0,038] => 5,149 (82)
 Area_Imovel > 68,500 [Ave: 5,199, Effect: 0,042] => 5,199 (170)
 Area_Imovel > 83,500 [Ave: 5,276, Effect: 0,076] (322)
 Area_Imovel <= 127,500 [Ave: 5,256, Effect: -0,02] (285)
 Percentagem_Populacao_ens_sup <= 20,320 [Ave: 5,227, Effect: -0,029] (204)
 Concelho in ["Moita"] [Ave: 5,062, Effect: -0,165] => 5,062 (10)
 Concelho in ["Almada" "Amadora" "Barreiro" "Loures" "Mafra" "Montijo" "Odivelas" "Palmela" "Seixal" "Setúbal" "Sintra" "Vila Franca de Xira"] [Ave: 5,236, Effect: 0,009] => 5,236 (194)
 Percentagem_Populacao_ens_sup > 20,320 [Ave: 5,33, Effect: 0,074] => 5,33 (81)
 Area_Imovel > 127,500 [Ave: 5,43, Effect: 0,154] => 5,43 (37)
 Novo_Lug_Estacionamento in ["1" "2" "3 ou mais"] [Ave: 5,481, Effect: 0,185] (367)
 Area_Imovel <= 122 [Ave: 5,409, Effect: -0,072] (201)
 Area_Imovel <= 88,500 [Ave: 5,266, Effect: -0,144] => 5,266 (37)
 Area_Imovel > 88,500 [Ave: 5,442, Effect: 0,032] (164)
 Idade_Imovel <= 12,500 [Ave: 5,514, Effect: 0,072] (69)
 Freguesia in ["Ericeira" "Moscatavide e Portela" "Santo Isidoro"] [Ave: 5,815, Effect: 0,301] => 5,815 (4)
 Freguesia in ["Aigualva e Mira-Sintra" "Almada, Cova da Piedade, Pragal e Cacilhas" "Alto do Seixalinho, Santo André e Verderena" "Alverca do Ribatejo e Sobralinho" "Caparica e Trafaria" "Laranjeiro e Feijó" "Loures" "Mina de Água" "Montijo e Afonsoeiro" "Pinhal Novo" "Póvoa de Santo Adrião e Olival Basto" "Ramada e Caneças" "Sacavém e Prior Velho" "Seixal, Arrentela e Aldeia de Paio Pires" "Setúbal (São Julião, Nossa Senhora da Anunciada e Santa Maria da Graça)" "Venteira" "Vialonga" "Águas Livres"] [Ave: 5,495, Effect: -0,019] => 5,495 (65)
 Idade_Imovel > 12,500 [Ave: 5,389, Effect: -0,052] (95)
 Concelho in ["Amadora" "Loures" "Odivelas"] [Ave: 5,481, Effect: 0,092] => 5,481 (35)
 Concelho in ["Almada" "Barreiro" "Mafra" "Montijo" "Palmela" "Seixal" "Setúbal" "Sintra" "Vila Franca de Xira"] [Ave: 5,336, Effect: -0,054] => 5,336 (60)
 Area_Imovel > 122 [Ave: 5,568, Effect: 0,087] (166)
 Populacao_ens_sup <= 2.958 [Ave: 5,379, Effect: -0,188] => 5,379 (16)
 Populacao_ens_sup > 2.958 [Ave: 5,588, Effect: 0,02] (150)
 Area_Imovel <= 204,500 [Ave: 5,569, Effect: -0,019] => 5,569 (135)

Area_Imovel > 204,500 [Ave: 5,759, Effect: 0,171] => 5,759 (15)

7.6. Anexo F: *Rule set* do modelo para perfis ajustado, subestimado e sobrestimado

Idade_Imovel <= 0,500 [Mode: Ajustado] (242)

Piso in ["2" "3" "4" "5" "Acima do 10" "C/V"] [Mode: Ajustado] (129)

Concelho in ["Montijo" "Sesimbra"] [Mode: Ajustado] (8)

Piso in ["2" "3" "4"] [Mode: Ajustado] (7)

Piso in ["2"] [Mode: Ajustado] => Ajustado (2; 1,0)

Piso in ["3" "4"] [Mode: Ajustado] (5)

Area_Imovel <= 142 [Mode: Sobrestimado] => Sobrestimado (4; 0,5)

Area_Imovel > 142 [Mode: Ajustado] => Ajustado (1; 1,0)

Piso in ["5"] [Mode: Subestimado] => Subestimado (1; 1,0)

Concelho in ["Almada" "Amadora" "Barreiro" "Cascais" "Lisboa" "Loures" "Odivelas" "Oeiras" "Palmela" "Seixal" "Sintra" "Vila Franca de Xira"] [Mode: Ajustado] (121)

Area_Imovel <= 116,500 [Mode: Ajustado] => Ajustado (52; 1,0)

Area_Imovel > 116,500 [Mode: Ajustado] (69)

Freguesia in ["Alcântara" "Avenidas Novas" "Belém"] [Mode: Ajustado] (7)

Idade_Imovel <= -0,500 [Mode: Subestimado] => Subestimado (2; 1,0)

Idade_Imovel > -0,500 [Mode: Ajustado] => Sobrestimado (5; 0,4)

Freguesia in ["Algueirão-Mem Martins" "Algés, Linda-a-Velha e Cruz Quebrada-Dafundo" "Alto do Seixalinho, Santo André e Verderena" "Barcarena" "Beato" "Carcavelos e Parede" "Cascais e Estoril" "Estrela" "Laranjeiro e Feijó" "Lumiar" "Marvila" "Olivais" "Penha de França" "Póvoa de Santa Iria e Forte da Casa" "Sacavém e Prior Velho" "Santo António" "Seixal, Arrentela e Aldeia de Paio Pires" "São Domingos de Rana" "Venteira" "Águas Livres"] [Mode: Ajustado] (62)

Freguesia in ["Algés, Linda-a-Velha e Cruz Quebrada-Dafundo" "Alto do Seixalinho, Santo André e Verderena" "Carcavelos e Parede"] [Mode: Ajustado] => Ajustado (29; 0,897)

Freguesia in ["Algueirão-Mem Martins" "Barcarena" "Beato" "Cascais e Estoril" "Estrela" "Laranjeiro e Feijó" "Lumiar" "Marvila" "Olivais" "Penha de França" "Póvoa de Santa Iria e Forte da Casa" "Sacavém e Prior Velho" "Santo António" "Seixal, Arrentela e Aldeia de Paio Pires" "São Domingos de Rana" "Venteira" "Águas Livres"] [Mode: Ajustado] => Ajustado (33; 1,0)

Piso in ["1" "6" "7" "8" "9" "R/C"] [Mode: Ajustado] (113)

Concelho in ["Almada" "Amadora" "Mafra" "Seixal" "Sintra"] [Mode: Ajustado] (14)

Area_Imovel <= 110,500 [Mode: Ajustado] (4)

Area_Imovel <= 101 [Mode: Ajustado] (3)

Area_Imovel <= 96 [Mode: Ajustado] => Ajustado (2; 1,0)

Area_Imovel > 96 [Mode: Sobrestimado] => Sobrestimado (1; 1,0)

Area_Imovel > 101 [Mode: Subestimado] => Subestimado (1; 1,0)

Area_Imovel > 110,500 [Mode: Sobrestimado] (10)

Freguesia in ["Algueirão-Mem Martins" "Laranjeiro e Feijó" "Venteira"] [Mode: Ajustado] (7)
 Area_Imovel <= 122 [Mode: Ajustado] => Ajustado (2; 1,0)
 Area_Imovel > 122 [Mode: Ajustado] => Sobrestimado (5; 0,4)

Freguesia in ["Enxara do Bispo, Gradil e Vila Franca do Rosário" "Seixal, Arrentela e Aldeia de Paio Pires" "Águas Livres"] [Mode: Sobrestimado] => Sobrestimado (3; 1,0)

Concelho in ["Barreiro" "Cascais" "Lisboa" "Montijo" "Odivelas" "Oeiras" "Palmela" "Vila Franca de Xira"] [Mode: Ajustado] (99)
 Novo_Lug_Estacionamento in ["2"] [Mode: Ajustado] (32)
 Piso in ["9"] [Mode: Sobrestimado] (2)
 Area_Imovel <= 154,500 [Mode: Ajustado] => Ajustado (1; 1,0)
 Area_Imovel > 154,500 [Mode: Sobrestimado] => Sobrestimado (1; 1,0)
 Piso in ["1" "6" "7" "8" "R/C"] [Mode: Ajustado] => Ajustado (30; 1,0)

Novo_Lug_Estacionamento in ["0" "1" "3 ou mais"] [Mode: Ajustado] (67)
 Freguesia in ["Algés, Linda-a-Velha e Cruz Quebrada-Dafundo" "Alto do Seixalinho, Santo André e Verderena" "Carcavelos e Parede" "Cascais e Estoril" "Estrela" "Misericórdia" "Montijo e Afonsoeiro" "Olivais" "Parque das Nações"] [Mode: Ajustado] (25)
 Quartos_por_m2 <= 0,025 [Mode: Ajustado] => Subestimado (21; 0,476)
 Quartos_por_m2 > 0,025 [Mode: Ajustado] => Ajustado (4; 1,0)

Freguesia in ["Ajuda" "Alcântara" "Avenidas Novas" "Belém" "Campo de Ourique" "Castanheira do Ribatejo e Cachoeiras" "Lumiar" "Marvila" "Pinhal Novo" "Porto Salvo" "Póvoa de Santa Iria e Forte da Casa" "Póvoa de Santo Adrião e Olival Basto" "Santa Maria Maior" "Santo António" "São Domingos de Rana" "São Vicente"] [Mode: Ajustado] (42)
 Freguesia in ["Avenidas Novas" "Campo de Ourique" "Lumiar" "Marvila"] [Mode: Ajustado] => Sobrestimado (11; 0,455)
 Freguesia in ["Ajuda" "Alcântara" "Belém" "Castanheira do Ribatejo e Cachoeiras" "Pinhal Novo" "Porto Salvo" "Póvoa de Santa Iria e Forte da Casa" "Póvoa de Santo Adrião e Olival Basto" "Santa Maria Maior" "Santo António" "São Domingos de Rana" "São Vicente"] [Mode: Ajustado] => Ajustado (31; 0,935)

Idade_Imovel > 0,500 [Mode: Ajustado] (1.768)
 Concelho in ["Alcochete" "Almada" "Cascais" "Mafra" "Moita" "Odivelas" "Oeiras" "Sesimbra" "Setúbal"] [Mode: Ajustado] (533)
 Nova_Eficiencia_Energetica in ["C" "D" "E"] [Mode: Ajustado] (453)
 Area_Imovel <= 93,500 [Mode: Ajustado] (270)
 Piso in ["1" "3" "8" "Acima do 10" "C/V"] [Mode: Ajustado] (141)
 Farmacias <= 17,500 [Mode: Ajustado] => Subestimado (134; 0,284)
 Farmacias > 17,500 [Mode: Subestimado] => Subestimado (7; 0,857)
 Piso in ["2" "4" "5" "6" "7" "9" "R/C"] [Mode: Ajustado] (129)

Concelho in ["Alcochete" "Sesimbra"] [Mode: Subestimado]
 => Subestimado (5; 0,6)
 Concelho in ["Almada" "Cascais" "Mafra" "Moita" "Odivelas"
 "Oeiras" "Setúbal"] [Mode: Ajustado] => Sobrestimado (124; 0,234)
 Area_Imovel > 93,500 [Mode: Ajustado] (183)
 Piso in ["1" "3" "4" "7" "C/V"] [Mode: Ajustado] (98)
 Densidade_Populacional <= 2440,810 [Mode: Sobrestimado]
 => Sobrestimado (36; 0,583)
 Densidade_Populacional > 2440,810 [Mode: Ajustado] =>
 Sobrestimado (62; 0,226)
 Piso in ["2" "5" "6" "8" "9" "Acima do 10" "R/C"] [Mode: Ajustado] (85)
 Concelho in ["Mafra" "Odivelas" "Oeiras" "Sesimbra"] [Mode:
 Ajustado] => Ajustado (43; 0,837)
 Concelho in ["Alcochete" "Almada" "Cascais" "Moita"
 "Setúbal"] [Mode: Ajustado] => Subestimado (42; 0,286)
 Nova_Eficiencia_Energetica in ["A" "B" "Menos que E"] [Mode: Ajustado] (80)
 Freguesia in ["Alcabideche" "Alcochete" "Algés, Linda-a-Velha e Cruz
 Quebrada-Dafundo" "Baixa da Banheira e Vale da Amoreira" "Carcavelos e
 Parede" "Castelo (Sesimbra)" "Ericeira" "Ramada e Caneças" "Setúbal (São
 Julião, Nossa Senhora da Anunciada e Santa Maria da Graça)"] [Mode:
 Ajustado] (32)
 Freguesia in ["Alcabideche" "Algés, Linda-a-Velha e Cruz Quebrada-
 Dafundo" "Castelo (Sesimbra)" "Ericeira" "Ramada e Caneças"] [Mode:
 Ajustado] => Ajustado (12; 1,0)
 Freguesia in ["Alcochete" "Baixa da Banheira e Vale da Amoreira"
 "Carcavelos e Parede" "Setúbal (São Julião, Nossa Senhora da
 Anunciada e Santa Maria da Graça)"] [Mode: Sobrestimado] (20)
 Piso in ["1" "2" "3" "5"] [Mode: Sobrestimado] =>
 Sobrestimado (15; 0,467)
 Piso in ["4" "R/C"] [Mode: Subestimado] => Subestimado (5;
 0,6)
 Freguesia in ["Almada, Cova da Piedade, Pragal e Cacilhas" "Carnaxide e
 Queijas" "Cascais e Estoril" "Laranjeiro e Feijó" "Odivelas" "Oeiras e São
 Julião da Barra, Paço de Arcos e Caxias" "Pontinha e Famões" "Porto Salvo"
 "Póvoa de Santo Adrião e Olival Basto" "Santiago (Sesimbra)" "São Domingos
 de Rana" "São Francisco" "São Sebastião" "Venda do Pinheiro e Santo
 Estêvão das Galés"] [Mode: Ajustado] (48)
 Piso in ["1" "3" "4" "6" "Acima do 10"] [Mode: Ajustado] (27)
 Freguesia in ["Almada, Cova da Piedade, Pragal e Cacilhas"
 "Laranjeiro e Feijó" "Oeiras e São Julião da Barra, Paço de
 Arcos e Caxias" "Póvoa de Santo Adrião e Olival Basto"]
 [Mode: Ajustado] => Subestimado (11; 0,364)
 Freguesia in ["Carnaxide e Queijas" "Cascais e Estoril"
 "Odivelas" "Pontinha e Famões" "Santiago (Sesimbra)" "São
 Domingos de Rana"] [Mode: Ajustado] => Ajustado (16; 1,0)
 Piso in ["2" "7" "C/V" "R/C"] [Mode: Subestimado] (21)
 Quartos <= 1,500 [Mode: Ajustado] => Ajustado (1; 1,0)
 Quartos > 1,500 [Mode: Subestimado] => Subestimado (20;
 0,75)
 Concelho in ["Amadora" "Barreiro" "Lisboa" "Loures" "Montijo" "Palmela" "Seixal" "Sintra"
 "Vila Franca de Xira"] [Mode: Ajustado] (1.235)
 Area_Imovel <= 36,500 [Mode: Ajustado] (24)

Freguesia in ["Arroios" "Corroios" "Estrela" "Misericórdia" "Penha de França" "Santa Maria Maior"] [Mode: Sobrestimado] (13)

Idade_Imovel <= 71,500 [Mode: Sobrestimado] (10)

Freguesia in ["Arroios"] [Mode: Sobrestimado] => Sobrestimado (3; 0,667)

Freguesia in ["Corroios" "Estrela" "Misericórdia" "Penha de França" "Santa Maria Maior"] [Mode: Sobrestimado] => Sobrestimado (7; 1,0)

Idade_Imovel > 71,500 [Mode: Ajustado] => Ajustado (3; 1,0)

Freguesia in ["Aigualva e Mira-Sintra" "Ajuda" "Avenidas Novas" "Santo António dos Cavaleiros e Frielas" "Seixal, Arrentela e Aldeia de Paio Pires" "São Domingos de Benfica" "São Vicente" "Venteira" "Águas Livres"] [Mode: Ajustado] (11)

Distrito in ["Lisboa"] [Mode: Ajustado] => Ajustado (10; 1,0)

Distrito in ["Setúbal"] [Mode: Subestimado] => Subestimado (1; 1,0)

Area_Imovel > 36,500 [Mode: Ajustado] (1.211)

WC_por_m2 <= 0,045 [Mode: Ajustado] (1.208)

Densidade_Populacional <= 4154,115 [Mode: Ajustado] (457)

Idade_Imovel <= 71,500 [Mode: Ajustado] => Sobrestimado (452; 0,168)

Idade_Imovel > 71,500 [Mode: Subestimado] => Subestimado (5; 0,6)

Densidade_Populacional > 4154,115 [Mode: Ajustado] (751)

Populacao_ens_sup <= 12.969 [Mode: Ajustado] => Ajustado (627; 0,751)

Populacao_ens_sup > 12.969 [Mode: Ajustado] => Sobrestimado (124; 0,169)

WC_por_m2 > 0,045 [Mode: Sobrestimado] => Sobrestimado (3; 1,0)

7.7. Anexo G: Regras de perfis de imóveis ajustados, subestimados e sobrestimados

Para móveis com valor ajustado:

- Se Idade_Imovel <= 0,500, e (Piso = "2" ou Piso = "3" ou Piso = "4" ou Piso = "5" ou Piso = "Acima do 10" ou Piso = "C/V"), e (Concelho = "Almada" ou Concelho = "Amadora" ou Concelho = "Barreiro" ou Concelho = "Cascais" ou Concelho = "Lisboa" ou Concelho = "Loures" ou Concelho = "Odivelas" ou Concelho = "Oeiras" ou Concelho = "Palmela" ou Concelho = "Seixal" ou Concelho = "Sintra" ou Concelho = "Vila Franca de Xira"), e Area_Imovel <= 116,500, então Tipo_Erro_Avaliação = "Ajustado" (Nó 17: suporte=52 imóveis; confiança=1);

Para imóveis com valor subestimado:

- Se Idade_Imovel > 0,500, e (Concelho = "Alcochete" ou Concelho = "Almada" ou Concelho = "Cascais" ou Concelho = "Mafra" ou Concelho = "Odivelas" ou Concelho = "Oeiras" ou Concelho = "Sesimbra" ou Concelho = "Setúbal"), e (Nova_Eficiencia_Energetica = "A" ou Nova_Eficiencia_Energetica = "B" ou Nova_Eficiencia_Energetica = "Menos que E", e (Freguesia = "Almada, Cova da Piedade, Pragal e Cacilhas" ou Freguesia = "Carnaxide e Queijas" ou Freguesia = "Cascais e Estoril" ou Freguesia = "Laranjeiro e Feijó" ou Freguesia = "Odivelas" ou Freguesia = "Oeiras e São Julião da Barra, Paço de Arcos e Caxias" ou Freguesia = "Pontinha e Famões" ou Freguesia = "Porto Salvo" ou Freguesia = "Póvoa de Santo Adrião e Olival Basto" ou Freguesia = "Santiago (Sesimbra)" ou Freguesia = "São Domingos de Rana" ou Freguesia = "São Francisco" ou Freguesia = "São Sebastião" ou Freguesia = "Venda do Pinheiro e Santo

Estêvão das Galés"), e (Piso= "2" ou Piso = "7" ou Piso = "C/V" ou Piso= "R/C"), e Quartos > 1,500, então Tipo_Erro_Avaliação = "Subestimado" (Nó 86: suporte= 20 imóveis; confiança=0,75);

Para imóveis com valor sobrestimado:

- Se (Idade_Imovel > 0,500 e Idade_Imovel <= 71,500), e Concelho = "Lisboa" ou Concelho = "Seixal"), e Area_Imovel <= 36,500, e (Freguesia = "Corroios" ou Freguesia = "Estrela" ou Freguesia = "Misericórdia" ou Freguesia = "Penha de França" ou Freguesia = "Santa Maria Maior"), então Tipo_Erro_Avaliação = "Sobrestimado" (Nó 88: suporte= 7 imóveis; confiança=1);