A Importância das Imagens de Alojamentos Turísticos nas Plataformas de Reservas Online: Uma Análise com Dados Reais da Feels Like Home.

Diana Filipa Oliveira Mendes

Mestrado em Business Analytics

Orientador:
Professor Doutor Rui Francisco Pereira Moital Loureiro da Cruz,
Professor Convidado, ISCTE- Executive Education

Co-Orientador:
Professor Doutor Tomás Gomes da Silva Serpa Brandão,
Professor Auxiliar, ISCTE-IUL

Setembro, 2022

Dedico esta dissertação de mestrado ao ISCTE que acompanhou toda a minha vida académica,
assim como a todos aqueles que estiveram sempre presentes.

Agradecimento

Esta dissertação não teria sido realizada sem o contributo e apoio de várias pessoas, às quais não poderia estar mais grata.

Em primeiro lugar, gostaria de deixar uma palavra de apreço ao Professor Francisco e ao Professor Tomás, pelo apoio, ajuda e disponibilidade constante ao longo do projeto, nunca deixando que os problemas e a desmotivação levassem a melhor.

Não poderia deixar de agradecer ao Professor Raul, que me acompanhou desde o início desta jornada. Desde o início do Mestrado em Business Analytics, que procurou proporcionar experiências enriquecedoras, tanto pessoal como profissionalmente.

Por fim, agradecer a compreensão e paciência da minha família e amigos, os meus pilares, que sempre me incentivaram e motivaram.

Muito obrigada a todos!

Resumo

Nesta dissertação procurou-se entender a importância do conteúdo visual das imagens de alojamentos na sua atratividade e quais as características visuais mais atraentes para os clientes. Este estudo foi desenvolvido com o intuito de ajudar a otimizar a estratégia para a escolha das fotografias dos anúncios da empresa Feels Like Home.

A revisão da literatura seguiu uma adaptação do PRISMA e incidiu sobre a importância das imagens utilizadas em plataformas de reserva de viagens e alojamento online. A recolha da literatura baseou-se em critérios de seleção aplicados no Web of Science e Scopus. Selecionaram-se, analisaram-se e avaliaram-se 23 artigos científicos relacionados com este tema ou que contribuem com informações úteis para aprofundar o conhecimento sobre o mesmo.

Na dissertação seguiu-se a metodologia CRISP-DM. Realizou-se a compreensão do negócio e dos dados, onde foram analisados e definidos o problema, os objetivos e o processo de pesquisa. A preparação dos dados englobou a extração de características visuais das imagens e a produção da tabela para modelação. A modelação e avaliação, permitiram, através de segmentação, analise bi-variada e classificação com árvores de decisão, desenvolver modelos e obter resultados. Por fim, a implementação consistiu na escrita deste documento e de um artigo.

Os resultados permitem concluir que o conteúdo das imagens dos anúncios é realmente importante. Para além disso, permitiram chegar a regras com um bom suporte e confiança, que descrevem os padrões dos alojamentos atrativos e não atrativos. Estas foram transformadas em recomendações práticas que a Feels Like Home pode seguir.

Palavras-chave: Revisão Sistemática da Literatura, Atratividade dos Alojamentos Turísticos, Agências de Reserva Online, Conteúdo Visual da Imagem, Análise de dados, Classificação.

Abstract

This dissertation seeks to understand the importance of visual content of the accommodation images in their attractiveness and which visual characteristics most attract customers. This study was developed with the aim of helping to optimize the strategy for choosing the photographs of the advertisements, already existing in Feels Like Home.

The literature review followed an adaptation of PRISMA and focused on the importance of images used in on online travel agencies and accommodation. The literature review was based on selection criteria applied in Web of Science and Scopus. 23 scientific articles related to this topic or that contribute with useful information to deepen knowledge about it were selected, analysed and evaluated.

In the dissertation, the CRISP-DM methodology was followed. The business and data were understood, the problem was analysed, and the objectives and research process defined. The preparation of the data involved the extraction of visual features from the images and the production of the table for modelling. The modelling and evaluation allowed, through segmentation, bivariate analysis, and classification with decision trees, to arrive at a model and results, and finally, the implementation, consisted of writing of this document and an article.

The results allow to conclude that the content of the advertisement images is important. Furthermore, allowed to arrive at rules with good support and confidence, which describe the standards of attractive and unattractive accommodations. These have been turned into practical recommendations that Feels Like Home can follow.

Keywords: Systematic Review of Literature, Attractiveness of Tourist Accommodation, Online Booking Agencies, Visual Image Content, Data Analysis, Classification.

Índice

Agradecimento iii

Resumo v

Abstract vii

CAPÍTULO 1: Introdução 1

- 1.1. Contexto e motivação 1
- 1.2. Questões de investigação 1
- 1.3. Objetivos e formas de validação 2
- 1.4. Contributos 2
- 1.5. Feels Like Home 3
- 1.6. Estrutura da Dissertação 3

CAPíTULO 2: Revisão da Literatura 5

- 2.1. Protocolo da Revisão sistemática da literatura 5
  - 2.1.1. Objetivos e questões de pesquisa 5
  - 2.1.2. Processo de seleção de artigos 6
  - 2.1.3. Avaliação dos artigos da RSL 7
- 2.2. Sintetização dos conteúdos dos artigos 8
- 2.3. Turismo e Alojamentos Turísticos 13
- 2.4. Análise de imagem, visão computacional e âmbitos de aplicação 13
  - 2.4.1. Destinos e atrações turísticas 14
  - 2.4.2. Imobiliário 16
- 2.5. Atratividade dos alojamentos 17
- 2.6. Avaliação dos artigos da RSL 18
- 2.7. Resposta às perguntas de pesquisa 20

CAPÍTULO 3: Metodologia 21

- 3.1. Compreensão do Negócio 21
- 3.2. Compreensão dos Dados 21
- 3.3. Preparação dos Dados 23
- 3.4. Modelação 28
- 3.5. Avaliação 31
- 3.6. Implementação 33
- 3.7. Testes alternativos 33

CAPíTULO 4: Resultados e Discussão 35

- 4.1. Caracterização dos alojamentos da FLH 35
- 4.2. A atratividade dos alojamentos da FLH 38

4.2.1. Relação entre atratividade e as outras características 38
4.2.2. Perfis da atratividade dos alojamentos 42
CAPÍTULO 5: Conclusões e Recomendações 49
5.1. Conclusão 49
5.2. Limitações 51
5.3. Recomendações 51
5.4. Trabalho futuro 52
Referências Bibliográficas 53
Anexos 57

Índice de Figuras

Figura 1. Fluxo de informação das propriedades. 1
Figura 2. Processo de seleção de artigos. 7
Figura 3. Nuvem de palavras-chave dos artigos da RSL. 12
Figura 4. Zonas de foco ocular pelos australianos à esquerda e chineses à direita. 14
Figura 5. Descrição das variáveis elaboradas manualmente para estimar a qualidade estética. 15
Figura 6. Identificação das zonas importantes na definição de atratividade da imagem. 15
Figura 7. Imagens utilizadas nos estudos de processamento de imagem no imobiliário. 17
Figura 8.Variáveis importantes para prever a atratividade de um imóvel. 17
Figura 9. Pontuação por critério de avaliação. 20
Figura 10. Processo metodológico. 21
Figura 11. Classificação das divisões das imagens. 22
Figura 12. Visualização segmentada por cor, das imagens da FLH. 22
Figura 13. Correspondência entre os valores de H (Hue) e as cores. 24
Figura 14. Entropia de uma imagem de exemplo. 24
Figura 15. Modelo de dados com integração de informação das imagens. 24
Figura 16. Outliers das variáveis andar, ocupação máxima e valor médio por noite. 25
Figura 17. Lista de alojamentos numa primeira pesquisa. 26
Figura 18. Página de Airbnb de um alojamento. 26
Figura 19. Exemplo de segmentação de elementos. 29
Figura 20. Exemplo de árvore de decisão binária. 30
Figura 21. Exemplo de árvore de decisão não necessariamente binária. 30
Figura 22. Resultado do modelo de segmentação. 35
Figura 23. Importância das variáveis para a segmentação. 36
Figura 24. Relação entre atratividade e os clusters. 36
Figura 25. Relação entre a região e o alvo. 38
Figura 26. Relação entre presença de elevador e ar condicionado e o alvo. 39
Figura 27. Valor por noite médio, por alvo. 39
Figura 28. Número de fotografias do anúncio, por alvo. 40
Figura 29. Percentagem média das cores presentes nas fotografias, por alvo. 40
Figura 30. Entropia e luminosidade médias. 41
Figura 31. Relação entre a divisão, da primeira à quinta fotografia do anúncio, e o alvo. 41
Figura 32. Exemplos de divisões categorizadas como Outros. 42
Figura 33. Gráfico de importância – modelo apenas com caraterísticas das casas. 44
Figura 34. Gráfico de importância – modelo com caraterísticas das casas e da 1ª imagem. 44
Figura 35. Gráfico de ganhos do modelo apenas com caraterísticas das casas. 45
Figura 36. Gráfico de ganhos do modelo com caraterísticas das casas e da 1ª imagem. 45

Índice de Tabelas

Tabela 1. Critérios de exclusão e inclusão de literatura. 6
Tabela 2. Critérios de avaliação dos artigos da RSL. 7
Tabela 3. Artigos incluídos na RSL, publicados entre 2015 e 2018. 8
Tabela 4. Artigos incluídos na RSL, publicados entre 2019 e 2021. 9
Tabela 5. Sumarização do contexto do estudo dos artigos da RSL. 10
Tabela 6. Sumarização das técnicas e variáveis utilizadas artigos da RSL. 11
Tabela 7. Avaliação dos artigos da RSL. 19
Tabela 8. Informação selecionada para a fase da modelação. 28
Tabela 9. Matriz de Confusão. 31
Tabela 10. Resultados dos modelos para cada conjunto de dados. 43

Lista de Acrónimos

DL – *Deep Learning*

FLH – Feels Like Home

GCV – *Google Cloud Vision*

HLS – *Hue* (escala de cor), lightness (luminosidade) e *saturation* (saturação)

HSV – *Hue* (escala de cor), *saturation* (saturação) e *value* (valor/brilho)

ML - *Machine Learning*

NUTS - Nomenclatura de Unidades Territoriais para Fins Estatísticos

PRISMA - *Preferred Reporting Items for Systematic reviews and Meta-Analyses* (Metodologia para revisões sistemáticas e meta-análises)

RGB – *Red* (vermelho), *green* (verde) e *blue* (azul)

ROC - *Receiver operating characteristic*

RSL – Revisão Sistemática da Literatura

CAPÍTULO 1: Introdução

1.1. Contexto e motivação

Na área do turismo, as plataformas de reservas online têm vindo a ter cada vez mais atenção da comunidade de ciência de dados. Devido ao aumento do número de turistas que reservam os hotéis através das plataformas de reservas online, sobretudo pela conveniência, praticidade e relação tempo versus custo em reservar nestas plataformas (Lien *et al.*, 2015), tem-se verificado um grande aumento da informação recolhida destas plataformas.

O trabalho desenvolvido ao longo desta dissertação baseia-se em dados fornecidos pela Feels Like Home (FLH), uma empresa de gestão de propriedades para aluguer de alojamento local. As suas propriedades são anunciadas não só na sua própria plataforma, mas também em outras agências de reservas online. A Figura 1 mostra o fluxo da informação das propriedades. Esta passa da FLH para as agências de reservas online através de um sistema de gestão de conteúdos, sendo a FLH que define qual a informação que deve ser apresentada e de que forma. Esta definição de conteúdos é simples para as características da casa, mas complexa e subjetiva quando se trata da escolha e disposição das fotografias da casa associadas ao anúncio.

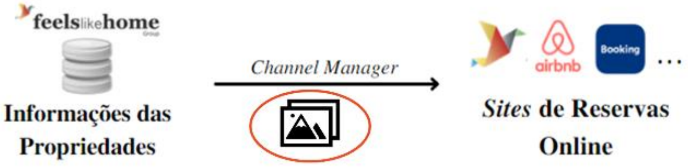


Figura 1. Fluxo de informação das propriedades.
Fonte: Elaboração própria, com recurso à aplicação *canva*.

O problema desta investigação prende-se precisamente com a escolha e disposição das fotografias no anúncio, ou seja, pretende-se compreender se e como é que as imagens das casas influenciam a procura dos clientes, que tipo de imagens devem ser apresentadas no anúncio, e em que ordem, e quais as características que devem estar presentes nas imagens. A resolução deste problema visa ajudar na definição de uma estratégia para a escolha das imagens a colocar nos anúncios de forma a otimizar a sua rentabilidade.

1.2. Questões de investigação

Tendo em conta a problemática e de forma a orientar a dissertação, surgem duas questões de investigação.

- Qual será a importância relativa do conteúdo visual na atratividade dos alojamentos da FLH, quando comparadas com as restantes características dos mesmos?
- Quais as características visuais das imagens dos alojamentos da FLH que mais atraem os potenciais clientes?

1.3. Objetivos e formas de validação

Os objetivos do estudo devem ser divididos em objetivos do negócio e analíticos. O principal objetivo de negócio é a geração de conhecimento sobre formas de proporcionar o aumento dos cliques nos anúncios dos alojamentos da FLH, o que deverá ter um impacto positivo nas reservas efetuadas.

Os objetivos analíticos contribuem naturalmente para cumprir os objetivos de negócio, correspondendo a cada um o seu critério de sucesso. Neste caso os objetivos e critérios de sucesso são:

1) Extrair as características visuais das imagens dos alojamentos da FLH.
    a) Este objetivo está cumprido aquando da elaboração da tabela com todas as variáveis relativas a cada imagem dos alojamentos, previamente definidas.
2) Identificar características dos alojamentos da FLH que permitam definir padrões de segmentação dos mesmos.
    a) Este objetivo é cumprido com a interpretação da análise das características dos alojamentos e com a segmentação dos mesmos, validada através do coeficiente de silhueta.
3) Criar os perfis dos alojamentos da FLH mais atrativos, tendo em conta a atratividade do alojamento perceber quais são as características mais importantes e se as características relacionadas com as imagens estão nesse grupo.
    a) Este objetivo está cumprido com a interpretação dos resultados da modelação através de árvores de decisão, validada com as métricas calculadas através da matriz de confusão.

1.4. Contributos

A investigação contribui a nível teórico para colmatar a lacuna no conhecimento existente na comunidade científica portuguesa relativamente a esta temática, culminando na produção da presente dissertação, escrita em português e baseada em dados reais de uma empresa nacional.

Adicionalmente, é um grande contributo para a comunidade científica no geral, visto que parte do trabalho desenvolvido na dissertação resultou na publicação de um artigo (Mendes et al., 2022). Este artigo consiste numa revisão sistemática da literatura acerca da importância do uso de imagens no contexto de turismo, onde se apresentam as principais abordagens que têm vindo a ser seguidas em análises de atratividade turística e de alojamento imobiliário. O artigo será apresentado na conferência "TMS ALGARVE 2022: Sustainability Challenges in Tourism, Hospitality and Managment" e poderá eventualmente ser também publicado num dos *journals* associados à conferência.

Por outro lado, a dissertação contribui a nível prático ajudando a FLH a melhorar a sua estratégia de escolha e disposição de imagens, visto que a introdução das imagens no anúncio é feita de forma manual e sem uma ordem definida.

1.5. Feels Like Home

A Feels Like Home (FLH) é uma empresa de gestão de propriedades, fundada em 2012, que oferece um serviço completo de gestão hoteleira, mais especificamente, no segmento de aluguer de curta duração.

A empresa disponibiliza aos seus clientes oportunidades de investimento, através da gestão integral do aluguer da propriedade, desde a preparação da casa, manutenção, limpeza, serviço de lavandaria e gestão de reserva, sem que os proprietários tenham de se preocupar com a gestão do seu imóvel.

A FLH tem à escolha mais de 550 propriedades em Lisboa, Porto, Algarve ou Madeira, prontos para serem alugados no regime de alojamento local (FeelsLikeHome, 2020). A empresa é responsável não só pela captação das fotografias dos imóveis, mas também pela sua colocação nos anúncios online.

1.6. Estrutura da Dissertação

A estrutura da dissertação segue de perto a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining). Esta é uma metodologia focada no negócio, que tem presente uma interação da atividade analítica com área da gestão e do negócio, o que justifica o permanente apoio e o acompanhamento da Feels Like Home a este estudo.

Após esta Introdução, no Capítulo 2 é apresentada a revisão sistemática da literatura (RSL) onde são descritos o protocolo e o processo de pesquisa da RSL, tal como a esquematização da extração da informação relevante para a RSL. É sumarizado o conteúdo de cada artigo e sistematizadas as ideias relevantes a reter para o presente estudo. Por fim, é feita a avaliação dos artigos presentes na RSL.

No Capítulo 3 é descrita toda a metodologia utilizada, o CRISP-DM, e em cada secção são descritas todas as tarefas executas em cada uma das fases: a compreensão do negócio, a compreensão de dados, a preparação de dados, a modelação, a avaliação e a implementação. É dado um maior foco às fases de compreensão de dados, preparação de dados e modelação.

No Capítulo 4 são resumidos todos os resultados da investigação, é feita a avaliação e discussão dos mesmos, assim como a comparação com os critérios de sucesso analíticos e com os resultados presentes na literatura.

No Capítulo 5 são apresentadas as principais conclusões do estudo e recomendações à FLH, este capítulo inclui também as limitações sentidas ao longo da execução da investigação e sugestões de pesquisas futuras.

Por último, são apresentados as referências e os anexos.

CAPíTULO 2: Revisão da Literatura

2.1. Protocolo da Revisão sistemática da literatura

Uma revisão sistemática da literatura (RSL) deve identificar, avaliar e sintetizar a literatura publicada pelos investigadores da área em estudo (Kitchenham & Brereton, 2013). Para tal, a estratégia de pesquisa, recolha e avaliação desta RSL está assente num protocolo adaptado da lista do PRISMA e do processo proposto por Barbara Kitchenham (2004).

A metodologia PRISMA (*Preferred Reporting Items for Systematic reviews and Meta-Analyses*) que foi desenvolvida para facilitar redação transparente e completa de revisões sistemáticas, uma metodologia bastante atual uma vez que foi atualizada em 2020 (Page *et al.*, 2021).

O protocolo detalha todo o processo da revisão sistemática. Inicialmente, para a pesquisa e recolha de literatura, são definidos os objetivos iniciais e as questões de partida, para os quais é criada uma *query* de pesquisa e critérios de inclusão e exclusão, que selecionam os artigos pertinentes para a resposta às questões iniciais. Depois, todos os artigos selecionados são analisados e o seu conteúdo é sumarizado e, por fim, avaliados individualmente com o objetivo de identificar quais os mais adequados ao tema em estudo e quais os critérios mais e menos abordados na literatura.

2.1.1. Objetivos e questões de pesquisa

O principal objetivo da RSL que foi realizada é sintetizar o que já se conhece sobre a importância das imagens nas plataformas de reservas online, para alojamentos ou atrações turísticas, ou no imobiliário, e contribuir para a melhoria e suporte dos objetivos e questões de investigação da dissertação, assim como, para um melhor planeamento da metodologia a adotar.

Visa responder à pergunta inicial: “No contexto do turismo e das agências de viagens online de que forma são analisadas, avaliadas e usadas as imagens?”.

A questão anterior foi dividida em quatro questões mais específicas que permitem esquematizar as conclusões a retirar:

i) “Quais os âmbitos e objetivos de análise onde mais se utiliza conteúdo de imagens?”

ii) “Como é que o conteúdo visual é extraído das imagens e quais as principais técnicas aplicadas?”

iii) “Como é definida a atratividade do alojamento e das imagens turísticas e quais as características com maior impacto na sua atratividade?”

iv) “Quão maduras estão as metodologias de pesquisa seguidas pelos estudos mais relevantes?”

2.1.2. Processo de seleção de artigos

A seleção de artigos para a revisão segue um processo com três fases: a identificação, a triagem e a inclusão. Na fase da identificação foram selecionadas duas bases de dados científicas, a *Web of Science* e a *Scopus*, uma vez que ambas se adequam à área em estudo (Mongeon & Paul-Hus, 2015) tornando a pesquisa mais precisa e abrangente (Meho & Yang, 2007).

Das duas bases de dados escolhidas, foram selecionados os artigos que atendiam à seguinte *query*, validada por dois especialistas: um na área de processamento de imagem e outro na área do alojamento turístico: *(image* or picture or photo* or "computer vision" or “scene classification” or “visual analytics”) and ("real estate" or touris* or "visual marketing" or “indoor-outdoor” or “hotel websites”* or "online booking sites" or airbnb or booking or *“location identification” or “tourist attractions” or “engagement” or “housing market” or “mental imagery”).*

Esta *query* resultou em 4798 estudos provenientes do *Web of Science* e 2930 do *Scopus*. Na segunda fase, foi feita uma triagem segundo os critérios de exclusão e por fim, na última fase, a inclusão, selecionaram-se os artigos a incluir na RSL, segundo os critérios de inclusão. Tanto os critérios de inclusão como os de exclusão estão apresentados na Tabela 1.

Tabela 1. Critérios de exclusão e inclusão de literatura.

Critérios de exclusão
• Documentos que não sejam artigos
• Artigos fora do período de publicação de 2015 a 2022
• Artigos que não estão na língua portuguesa ou inglesa
• Artigos sem livre acesso
• Artigos Duplicados
• Artigos sem fator impacto
Critérios de inclusão
• Artigos que abordam o tema do turismo e alojamentos turísticos
• Artigos que abordam o tema da importância das imagens no turismo.
• Artigos que analisam imagens no âmbito do turismo ou do imobiliário.

Os critérios de exclusão levaram a uma seleção de 241 artigos. De seguida, os critérios de inclusão foram aplicados através da leitura dos títulos e *abstract*, o que resultou numa seleção 35 artigos. Após leitura na integra e nova aplicação dos critérios de inclusão foram considerados 23 artigos para entrarem na RSL e serem analisados. Todo o processo de seleção está sintetizado na Figura 2, a qual mostra o número de documentos que permanecem após aplicado cada critério.

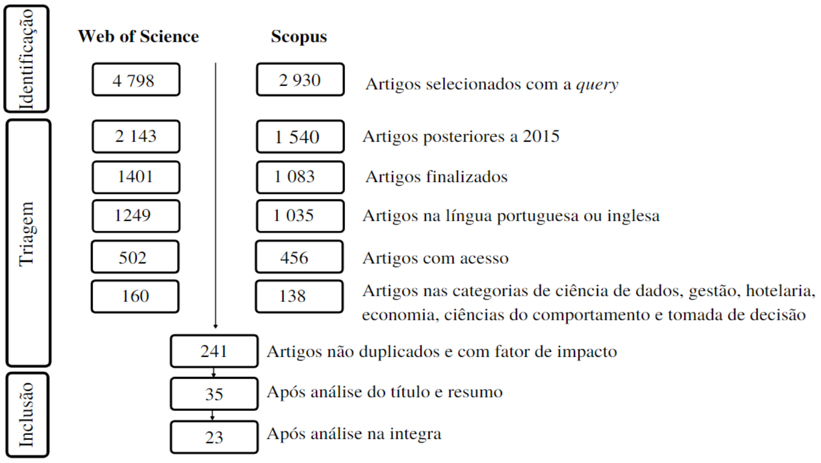

Identificação
Triagem
Inclusão
Web of Science
Scopus
4 798
2 930
Artigos selecionados com a *query*
2 143
1 540
Artigos posteriores a 2015
1401
1 083
Artigos finalizados
1249
1 035
Artigos na língua portuguesa ou inglesa
502
456
Artigos com acesso
160
138
Artigos nas categorias de ciência de dados, gestão, hotelaria, economia, ciências do comportamento e tomada de decisão
241
Artigos não duplicados e com fator de impacto
35
Após análise do título e resumo
23
Após análise na integra

Figura 2. Processo de seleção de artigos.
Fonte: Adaptado do diagrama de seleção do PRISMA.

2.1.3. Avaliação dos artigos da RSL

O último passo da RSL é a avaliação da literatura, onde, para cada questão de pesquisa, se associou um subconjunto de critérios, validados por especialistas na área de hotelaria e imobiliário e de processamento de imagem. Os critérios definidos sintetizam-se na Tabela 2 e permitem avaliar quais os artigos que melhor exploram cada questão. Para cada artigo foram pontuados os nove critérios de avaliação, avaliando-se cada artigo tendo em conta o contexto e objetivo do estudo, as variáveis e técnicas de extração utilizadas, as variáveis relacionadas com a atratividade do alojamento, a metodologia utilizada, e as limitações e contribuições do estudo.

Esta pontuação permitiu identificar o nível de qualidade e importância dos artigos para esta dissertação, contribuindo para a clarificação e confiança dos resultados da RSL (Kitchenham & Brereton, 2013). A pontuação varia entre 0 e 1, de acordo com a resposta às questões associadas: pontuação 1 nos artigos que respondem de forma completa à questão; 0,5 nos que respondem parcialmente; e 0 nos que não respondem de todo à questão.

Tabela 2. Critérios de avaliação dos artigos da RSL.

Questões	Critério para a avaliação dos artigos	
i)	C1	Descreve aprofundadamente o contexto do estudo?
	C2	Descreve, de forma esclarecedora, quais as variáveis a considerar no estudo?
ii)	C3	Extrai informação da imagem através de processamento de imagem? *
	C4	Detalha a técnica de extração da característica da imagem?
iii)	C5	Diz como as características das imagens influenciam os clientes?
	C6	Detalha a relação entre as variáveis e a atratividade de um alojamento?
iv)	C7	A metodologia é rigorosa e replicável?
	C8	Descreve as limitações do estudo?
	C9	Descreve os contributos do estudo?

* 1 – Extrai conteúdo apenas através de processamento de imagem; 0,5 – Extrai conteúdo, mas não com processamento de imagem; 0 – Não extrai conteúdo de imagem.

2.2. Síntese dos conteúdos dos artigos

Nas Tabelas 3 e 4 estão identificados os artigos selecionados, as suas características e objetivo. Para além disso mostram que ainda há poucos estudos apenas sobre as imagens e os alojamentos turísticos, apenas 5 dos 23 artigos selecionados estudam exatamente as imagens de alojamentos (22%).

Tabela 3. Artigos incluídos na RSL, publicados entre 2015 e 2018.

ID	Ano	Título	Autores	Jornal	Q	Objetivo
1	2015	Online hotel booking: the effects of brand image, price, trust and value on purchase intentions.	Lien, C. H., *et al.*	Asia Pacific Management Review	2	Analisar o efeito da marca, preço, confiança e intensão de reserva
2	2016	An eye-tracking study of tourism photo stimuli: image characteristics and ethnicity	Wang, Y., & Sparks, B. A.	Journal of Travel Research	1	Analisar a forma como as imagens turísticas são visualizadas
3	2017	Aggregate consumer ratings and booking intention: the role of brand image	Casado-Díaz, A. B., *et al.*	Service Business	3	Analisar o efeito das classificações globais de hotéis na intenção de compra dos consumidores, considerando marca.
4	2017	Image-based appraisal of real estate properties.	You, Q., *et al.*	IEEE-TM	1	Analisar o impacto do conteúdo visual no preço de uma casa
5	2017	Making sense of tourists' photographs using canonical variate analysis	Balomenou, N., *et al.*	Tourism Management	1	Estudar utilização de variáveis canónicas
6	2017	Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr	Yang, L. Wu, L. Liu, Y. & Kang, C.	ISPRS-IJGI	2	Criar um sistema de recomendação, através da descoberta de padrões de comportamento
7	2018	Asymmetric revelation effect: the influence of an increased number of photos on mental imagery and behavioural responses depending on target market.	Larceneux, F., *et al.*	RAM	2	Estudar a influência do nº de fotos no comportamento dos compradores
8	2018	Characterizing the location of tourist images in cities. Differences in user-generated images (Instagram), official tourist brochures and travel guides	Agustí, D. P.	Annals of Tourism Research	1	Analisar como varia a atratividade de imagem turística tendo em conta o suporte onde é apresentada
9	2018	Effects of the booking.com rating system: bringing hotel class into the picture.	Mariani, M. M., & Borghi, M.	Tourism Management	1	Aprofundar uma análise sobre o efeito da avaliação dos hotéis no booking.com
10	2018	Homeseeker: a visual analytics system of real estate data.	Li, M., *et al.*	JVLC	3	Criar uma aplicação que facilite a procura de casa

Notas: Q – quartil do jornal; RAM - Recherche et Applications En Marketing; JVLC - Journal of Visual Languages and Computing; ISPRS-IJGI - ISPRS-International Journal of Geo-Information; IEEE TM - IEEE Transactions on Multimedia

Tabela 4. Artigos incluídos na RSL, publicados entre 2019 e 2021.

ID	Ano	Título	Autores	Jornal	Q	Objetivo
11	2019	Photographs in tourism research: Prejudice, power, performance and participant-generated images	Balomenou, N. & Garrod, B.	Tourism Management	1	Provar que a utilização de imagens nos estudos de turismo preenche uma lacuna na teoria
12	2019	Predicting image aesthetics for intelligent tourism information systems	Kleinlein, R., *et al.*	Electronics	3	Criar uma aplicação para reconhecer a atração e fornecer informação
13	2020	An integrated picture fuzzy anp-todim multi-criteria decision-making approach for tourism attraction recommendation	Tian, C. & Peng, J.	Technological and Economic Development of Economy	1	Criar um Sistema de recomendação através do *picture fuzzy score*
14	2020	Characterizing tourism destination image using photos' visual content.	Xiao, X., Fang, C., & Lin, H.	ISPRS -IJGI	2	Analisar as características temporais e espaciais de atrações turísticas
15	2020	Is a picture worth a thousand words? An empirical study of image content and social media engagement	Li, Y., & Xie, Y.	Journal of Marketing Research	2	Analisar como é que a imagem e as características das publicações influenciam os utilizadores
16	2020	What image features boost housing market predictions?	Kostic, Z., & Jevremovic, A.	IEEE-TM	1	Analisar a atratividade de uma casa tendo em conta as sua características e imagens
17	2021	Color and engagement in touristic instagram pictures: a machine learning approach.	Yu, J., & Egger, R.	Annals of Tourism Research	1	Analisar a relação entre cor da imagem e interações dos utilizadores
18	2021	How many bedrooms do you need? A real-estate recommender system from architectural floor plan images.	Gan, Y. S., *et al.*	Scientific Programming.	3	Criar um método de processamento de imagem automatizado para calcular o nº quartos através de plantas
19	2021	Location identification and personalized recommendation of tourist attractions based on image processing	Zhang, Q., *et al.*	Traitement du Signal	3	Criar um método de identificação e recomendação de atrações baseado em imagens
20	2021	Research on night tourism recommendation based on intelligent image processing technology	Li, M., & Fan, N.	Scientific Programming.	4	Criação de um sistema de recomendação baseado em imagens turísticas noturnas
21	2021	The role of photograph aesthetics on online review sites: effects of management- versus traveler-generated photos on tourists' decision making	Marder, B., *et al.*	Journal of Travel Research	2	Analisar a influência das fotos amadoras e profissionais têm na intenção de reserva
22	2021	Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos	Kang, Y., *et al.*	ISPRS-IJGI	2	Classificar imagens de atrações turísticas
23	2021	Visualising natural attractions within national parks: preferences of tourists for photographs with different visual characteristics	Zhu, L., *et al.*	PLoS ONE	2	Analisar a relação entre características visuais e a perceção visual de fotografias de atrações na natureza

Notas: Q – quartil do jornal; RAM – Recherche et Applications En Marketing; JVLC – Journal of Visual Languages and Computing; ISPRS-IJGI – ISPRS-International Journal of Geo-Information; IEEE TM – IEEE Transactions on Multimedia

Na Tabela 5 estão identificados o âmbito de cada artigo, o tipo, tamanho, fonte e país dos dados utilizados no estudo, o que clarifica o seu contexto, bem como o tipo de análise.

Tabela 5. Sumarização do contexto do estudo dos artigos da RSL.

ID	Âmbito	Dados	Fonte	País	Análise
1	Reservas Hotel Online	366 inquiridos	Lifewin	China	Descrição
2	Atrações Turísticas	331 inquiridos 16 imagens	Tourism Australia	Austrália e China	Descrição
3	Reservas Hotel Online	15 inquiridos	TripAdvisor	Espanha	Descrição
4	Imobiliário	4 564 imagens	Realtor	EUA	Previsão
5	Atrações Turísticas	278 inquiridos 1 496 imagens	Dados Académicos	N.E.	Descrição
6	Redes Sociais	49 M imagens	Flickr	EUA	Segmentação
7	Imobiliário	3 658 anúncios	meilleursagents.com	França	Previsão
8	Atrações Turísticas	37000 imagens	Brochuras, Guias Turísticos e Instagram	Uruguai	Descrição
9	Reservas Hotel Online	1 228 089 avaliações	Booking.com	Reino Unido	Descrição
10	Imobiliário	1 437 679 anúncios	Realestate.com Google Maps	Austrália	Descrição
11	Atrações Turísticas	N.E.	N.E.	N.E.	Descrição
12	Atrações Turísticas	984 imagens	Flickr	Espanha	Previsão
13	Atrações Turísticas	N.E.	N.E.	N.E.	Recomendação
14	Atrações Turísticas	531 629 imagens	Mafengwo	China	Classificação
15	Redes Sociais	537 imagens	Twitter	N.E.	Previsão
16	Imobiliário	19 942 imagens	MLS Metadata	EUA	Previsão
17	Redes Sociais	4 757 publicações	Instagram	N.E.	Classificação
18	Imobiliário	892 plantas	Várias*	N.E.	Segmentação
19	Atrações Turísticas	100 000 imagens	N.E.	China	Recomendação
20	Atrações Turísticas	32 imagens	N.E.	China	Recomendação
21	Reservas Hotel Online	1 282 inquiridos 10 imagens	Tripadvisor	N.E.	Descrição
22	Redes Sociais	168 216 imagens	Flickr	Coreia	Classificação
23	Atrações Turísticas	30 imagens	N.E.	China	Descrição

Notas: N.E. - não específica; *CVC-FP, SESYD, Robin and Rent3D

A Tabela 5 mostra que o âmbito mais estudado são as atrações turísticas e os objetivos são variados. Mostra que existe uma predominância das análises mais descritivas, principalmente nos artigos mais antigos. Já para os estudos mais recentes observa-se uma evolução das análises realizadas, sendo aplicadas por exemplo análises de recomendação e classificação.

Na Tabela 6 estão identificadas as técnicas de análise e as variáveis utilizadas nos diversos estudos, como o alvo e variáveis relativas a alojamentos, imagens (e a técnica de extração do conteúdo da imagem) e outras variáveis que devem ser tidas em conta.

Tabela 6. Sumarização das técnicas e variáveis utilizadas artigos da RSL.

ID	**Técnicas**	**Variáveis Estruturadas**			
		Alvo	**Das Casas**	**Das imagens**	**Outras**
1	Análise Estatística	N.E.	Marca, preço, valor	N.E.	Confiança, intenção de compra
2	Análise Estatística	N.E.	N.E.	Técnica de rastreio ocular: Dados de rastreamento ocular	Nacionalidade
3	Análise Estatística	N.E.	N.E.	N.E.	Idade, género, educação, rendimento
4	*ML* - supervisionado	Preço	Localização	*DL*: vetor da imagem	N.E.
5	Análise Estatística	N.E.	N.E.	Inquérito: % de céu, nº de pessoas, animais, carros…	N.E.
6	*ML* - não supervisionado	N.E.	N.E.	Localização	N.E.
7	Regressão Múltipla	Cliques; dias no mercado	Preço/m2	M: Nº de fotos do anúncio	Segmento de mercado
8	Análise Estatística	N.E.	N.E.	N.E.	Dimensão do texto, nº de gostos e *hashtags*
9	Análise Estatística	N.E.	Nº de estrelas e avaliação	N.E.	N.E.
10	Análise Estatística	N.E.	Atributos e geo-informações	N.E.	N.E.
11	Análise Qualitativa	N.E.	N.E.	N.E.	N.E.
12	Regressão Logística e *ML* - supervisionado	Estética*	N.E.	A: Intensidade, HSV, entropia, linha do horizonte. DL: variáveis não identificáveis	N.E.
13	*ML* - não supervisionado	N.E.	N.E.	N.E.	N.E.
14	Modelos de NLP	Cenários turísticos	N.E.	DL: cenários turísticos	N.E.
15	Regressão	Interações	N.E.	M: Foto amadora ou não. GCV: cores, objetos, pessoas, qualidade	N.E.
16	Regressão e *ML* - supervisionado	Preço; Dias no mercado	Área, idade, nº de quartos, wc e garagem	A: Entropia, centro de gravidade, HSV. DL: categorias e objetos	N.E.
17	*ML* - supervisionado	Interações	N.E.	GCV: RGB e HSL DL: cenários turísticos	N.E.
18	*ML* - não supervisionado	N.E.	Área, nº de divisões	A: Escalas de cinza DL: posição dos quartos, área, nº portas e objetos	N.E.
19	*ML* - não supervisionado	N.E.	N.E.	DL: variáveis não identificáveis	N.E.
20	*ML* - não supervisionado	N.E.	N.E.	A: Escalas e valores de cinza	N.E.
21	Análise Estatística	N.E.	N.E.	Inquérito: Estética Visual M: foto amadora ou não	N.E.
22	*ML* - supervisionado	Cenários turísticos	N.E.	DL: cenários turísticos	N.E.
23	Análise Estatística	N.E.	N.E.	M: cenários da natureza	N.E.

Notas: N.E. – não específica; ML: *machine learning;* DL – *deep learning*; M – extração manual; A – extração algorítmica simples; GCV – extração através da aplicação *Google Cloud Vision*; *elaborado através de inquérito

A Tabela 6 permite perceber que o alvo, quando considerado é medido de diversas formas. Alguns estudos utilizam questionários para obter a opinião das pessoas sobre atratividade. Outros utilizam o número de cliques no anúncio, os dias de mercado ou o preço, para serem mais precisos.

É também possível entender que as características mais importantes relacionadas com a atratividade dos imóveis são as suas próprias características físicas, tais como a tipologia ou ano de construção, seguidas de perto pelas características das imagens nos anúncios. Já nas imagens de atrações turísticas é mais difícil compreender as características relevantes devido à falta de explicabilidade nos modelos utilizados, mas é bastante claro que o conteúdo da imagem é uma das características mais utilizada.

Observa-se ainda que as características associadas ao conteúdo visual das imagens podem ser extraídas de forma manual ou automática, aplicando técnicas de computação. Exemplos de características extraídas de forma manual incluem informações sobre o número de fotos ou indicação sobre se as imagens foram captadas por amadores ou profissionais.

Quanto a características extraídas de forma automática, estas podem ser obtidas aplicando algoritmos tradicionais de processamento de imagem para extrair, por exemplo, medidas associadas às cores, nos espaços HSV ou RGB, medidas de contraste e luminância em escalas de cinza, medidas entropia, estimação da linha do horizonte e do centro de gravidade da imagem.

Também podem ser utilizadas técnicas computacionais mais complexas de *deep learning*, para extrair categorias, objetos, cenários turísticos, posição das divisões na planta de um imóvel, ou eventualmente recorrendo a ferramentas para desenvolvimento de aplicações de *deep learning*, tais como a Google Cloud Vision API (GCV).

Na Figura 3 encontra-se representada a nuvem de palavras criada a partir das palavras-chave dos artigos analisados na RSL. Através da nuvem de palavras consegue-se apreender que uma das palavras-chave que ocorre mais vezes é *real estate* (mercado imobiliário), apesar do âmbito mais estudado ser a atração turística. Isto pode dever-se ao facto de, nos estudos sobre o mercado imobiliário, as palavras-chave serem mais consistentes, quando comparado com as palavras-chave dos estudos sobre atrações turísticas.

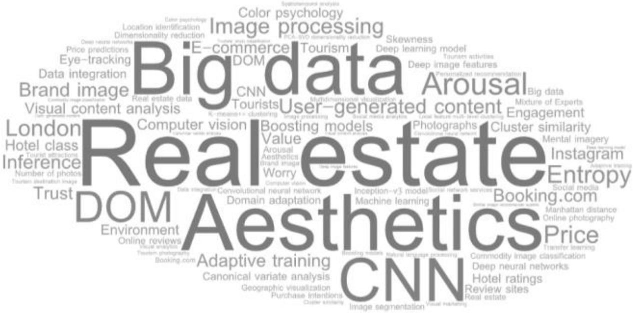

Color psychology
Image processing
Location identification
Dimensionality reduction
Skewness
Price predictions
E−commerce
Tourism
Deep learning model
Eye−tracking
DOM
Deep image features
Data integration
Big data
Arousal
Brand image
CNN
Big data
Real estate data
Tourists
Mixture of Experts
Visual content analysis
User−generated content
Engagement
London
Computer vision
Boosting models
Photographs
Cluster similarity
Value
Mental imagery
Hotel class
Real estate
Instagram
Inference
Arousal
Aesthetics
Brand image
Number of photos
Worry
Entropy
Social media
Trust
Convolutional neural network
Inception−v3 model
Social network services
Booking.com
DOM
Domain adaptation
Machine learning
Manhattan distance
Online photography
Environment
Aesthetics
Price
Online reviews
Transfer learning
Booking.com
Adaptive training
Commodity image classification
Deep neural networks
Canonical variate analysis
CNN
Hotel ratings
Geographic visualization
Review sites
Purchase intentions
Real estate
Image segmentation

Figura 3. Nuvem de palavras-chave dos artigos da RSL.
Fonte: Elaboração própria com.wordclouds.com.

De facto, através da Tabela 5 e da Figura 3, é percetível que existem poucos estudos sobre alojamentos turísticos presentes em *sites* de reserva online, o que comprova que essa área de estudo representa uma oportunidade de investigação que deve ser aproveitada. Por outro lado, no contexto desta dissertação esse facto pode ser considerado uma limitação, pois não existe uma base substancial de trabalhos que permitam uma fácil comparação de resultados entre as diferentes análises realizadas.

Nas próximas secções, o conteúdo dos 23 artigos selecionadas é descrito e analisado de uma forma comparativa e exaustiva. São discutidos temas como estudos existentes sobre as atrações e os alojamentos turísticos, visão computacional e os seus principais âmbitos de aplicação e por fim as abordagens que existem no que toca ao estudo da atratividade dos alojamentos e dos sites de reserva online.

2.3. Turismo e Alojamentos Turísticos

Na literatura existem alguns estudos relacionados com os sites de reserva online e com diferentes objetivos, por exemplo: o estudo da relação entre a avaliação dada pelos hóspedes e classe do hotel (Mariani & Borghi, 2018), a análise dos efeitos diretos ou indiretos da imagem de marca, da confiança, do preço e do valor percebidos nas intenções de reserva do hotel (Lien *et al*., 2015) ou a análise do efeito da classificação global do hotel na intenção de compra dos consumidores (Casado-Díaz *et al*., 2017).

Estes estudos, permitem identificar relações importantes entre as características dos alojamentos, como a classe do hotel (número de estrelas, que representa a qualidade/valor do alojamento) e a avaliação do hotel. A relação entre estas duas caraterísticas mostra que, quanto maior a classe do hotel, mais assimétrica é a distribuição das classificações, isto quer dizer que, quanto melhor é o hotel, maior a dispersão de opiniões (Mariani & Borghi, 2018). Outro estudo mostra que, segundo os inquiridos, quanto mais nítidas e autoexplicativas forem as fotografias maior será a intenção de compra (Lien *et al*., 2015).

Mais recentemente, foi estudado o impacto das fotografias amadoras ou profissionais na intenção de reserva, concluindo-se que a utilização de imagens mais profissionais faz um destino parecer visualmente mais atraente, o que acaba por impulsionar as intenções de reserva (Marder *et al*., 2021).

Estes estudos têm em comum uma abordagem muito simples no que toca ao tipo de análise, recolha de dados e variáveis utilizadas. Ou são assentes em inquéritos (Lien *et al*., 2015), e, por conseguinte, em opiniões, o que torna as conclusões muito subjetivas, ou então, fazem apenas uma pequena análise descritiva e visual dos dados (Marder *et al*., 2021), não aprofundando a análise.

2.4. Análise de imagem, visão computacional e âmbitos de aplicação

A visão computacional, processamento de imagem ou análise de conteúdo visual tem ganhado força nos últimos anos e consequentemente tem sofrido vários avanços e desenvolvimentos (Kang *et al*., 2021; Kleinlein *et al*., 2019; Xiao *et al*., 2020; Yu & Egger, 2021). Estes avanços ocorreram tanto a nível das técnicas utilizadas, como em relação ao tipo de informação extraída das imagens ou vídeos. Têm por isso aumentado os tipos de análises que podem ser feitas baseadas em imagens, o que faz com que o processamento de imagem seja aplicado nos mais variados âmbitos.

2.4.1. Destinos e atrações turísticas

No âmbito do turismo, existem estudos com abordagens menos complexas que utilizam análises muito manuais e também assentes em opiniões recolhidas por inquéritos, como forma de retirar informação sobre a atração turística ou sobre o conteúdo visual das imagens correspondentes (Agustí, 2018; Balomenou *et al.*, 2017; Balomenou & Garrod, 2019; Zhu *et al.*, 2021).

Estes estudos concluem que as imagens de atrações turísticas, sejam elas tiradas por profissionais ou amadores, assim como a sua presença nas redes sociais, são cada vez mais importantes para a atração de turistas. A qualidade visual de uma fotografia é um dos fatores mais importantes que determinam a atratividade das atrações (Zhu *et al.*, 2021), apresentando uma relação positiva com a atratividade das atrações, ou seja, quanto maior a qualidade das suas imagens maior a sua atratividade. Esta conclusão vem corroborar estudos anteriores (Lien *et al.*, 2015; Marder *et al.*, 2021).

Noutro estudo é proposta a utilização de rastreamento ocular, com o objetivo de analisar de que forma é que as imagens são visualizadas por pessoas de diferentes culturas e géneros (Wang & Sparks, 2016). Este estudo afirma que as imagens da natureza e com bastantes pormenores retém maior atenção e, tendo em conta as nacionalidades da amostra, os australianos em comparação com os chineses, mostram uma maior familiaridade com natureza e um processamento ocular mais rápido, observando rapidamente toda a imagem e com maior foco, neste caso na pessoa, como se pode observar na Figura 4. Retém-se então que, por um lado a nacionalidade da pessoa e por outro a quantidade e tipo de conteúdo visual são variáveis que podem explicar a forma como a imagem é percecionada pela audiência.

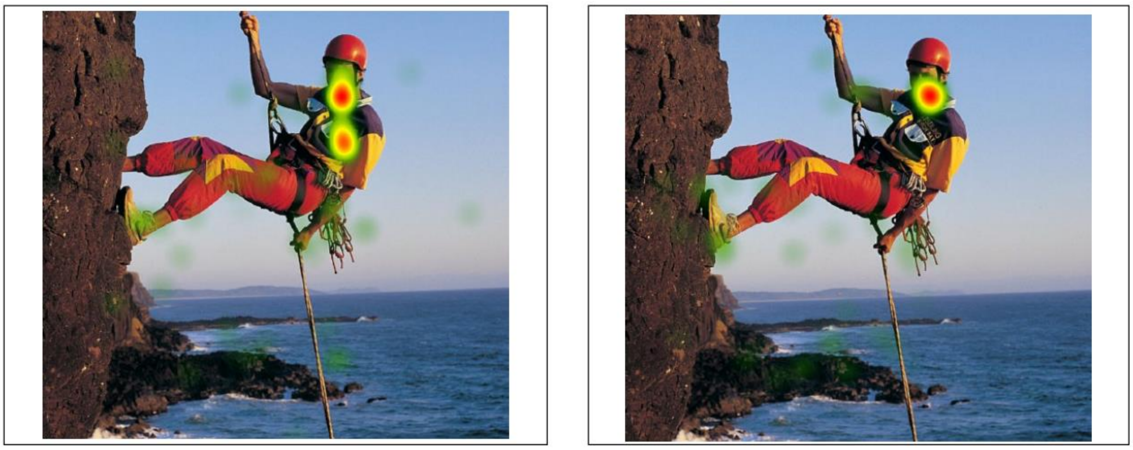


Figura 4. Zonas de foco ocular pelos australianos à esquerda e chineses à direita.
Fonte: Wang & Sparks (2016).

Outros artigos aplicam técnicas de classificação e recomendação, através de modelos de *deep learning*, tanto para identificar e caracterizar destinos turísticos (Kang *et al.*, 2021; Xiao *et al.*, 2020), assim como para criar sistemas de recomendação de atrações turísticas (Tian & Peng, 2020; Yang *et al.*, 2017; Zhang *et al.*, 2021). Estes modelos, apesar de apresentarem bons resultados de classificação e recomendação, devido à sua complexidade não permitem identificar variáveis importantes relativas ao conteúdo imagens. Estes tipos de modelos apenas recebem as imagens como entrada e apresentam o resultado do modelo como saída.

Em contraste, existem estudos que transformam a informação não estruturada das imagens, em informação estruturada, possível de ser analisada. Os objetivos destes estudos tanto são, à semelhança dos anteriores, a identificação (Kleinlein *et al*., 2019) e a recomendação de destinos ou atrações turísticas (M. Li & Fan, 2021), como a análise da relação das principais características de uma publicação com as interações da mesma na rede social (Y. Li & Xie, 2020; Yu & Egger, 2021).

A extração de conteúdo visual das imagens nestes artigos, realiza-se através de algoritmos não profundos e mais tradicionais, programados pelos investigadores, com os quais conseguem extrair, medidas associadas à cor e medidas calculadas com base nos valores de cinza (M. Li & Fan, 2021), como a intensidade média da imagem, a percentagem de cor da imagem, entropia média dos pixéis da imagem, as propriedades da linha do horizonte. Na Figura 5 estão as descrições das variáveis criadas (Kleinlein *et al*., 2019).

Variáveis	**Descrição**
Intensidade	Média de brilho da imagem
Matriz de Cor	Média do valor do H (*hue*) na escala de cores HSV
Saturação	Média do valor do S na escala de cores HSV
Entropia	Valor da entropia nos pixéis da imagem
Coloração	Diferença entre o histograma de cores da imagem e um histograma de cores com distribuição normal
Perfil de cores	Diferença entre o histograma de cores da imagem e um histograma de referência para as 8 cores da escala
Regra dos três	A regra das 3 cores primárias
Linha do horizonte	Presença e propriedades da linha do horizonte na imagem

Figura 5. Descrição das variáveis elaboradas manualmente para estimar a qualidade estética.
Fonte: Adaptado de Kleinlein et al. (2019), pág. 7.

Estas variáveis são frequentemente utilizadas para comparação com o uso de algoritmos de *deep learning* que, à semelhança do estudo de *eye tracking*, permitem a identificação de zonas importantes na imagem tendo em conta a sua atratividade, como ilustrado na Figura 6. Esta comparação permite perceber que a utilização de *deep learning*, em detrimento de variáveis manuais, permite muitas vezes chegar a melhores resultados e resultados mais escaláveis. Porém a utilização de variáveis menos profundas e mais tradicionais, aumenta a explicabilidade e diminui a complexidade dos modelos.

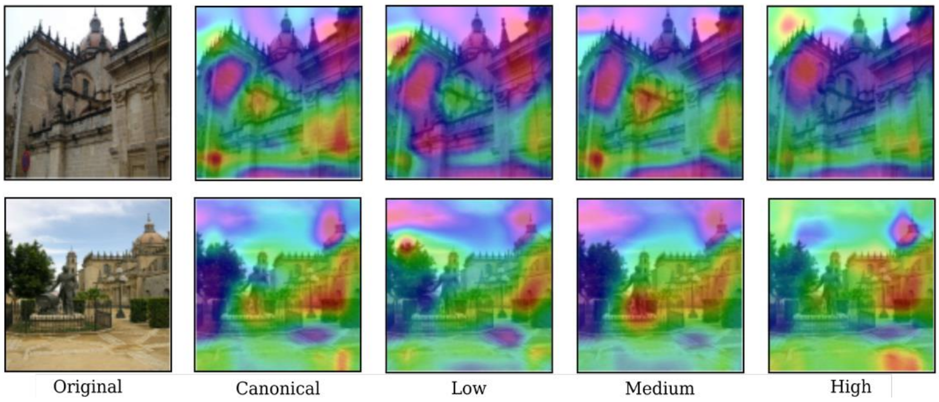


Figura 6. Identificação das zonas importantes na definição de atratividade da imagem.
Fonte: Kleinlein et al. (2019).

Como forma de extração de conteúdo visual é também, muitas vezes utilizada a *Google Cloud Vision API*, uma ligação a um *software* da *Google* integrado com modelos de *deep learning* que extraem informações das imagens automaticamente, tornando assim a informação não estruturada, presente nas imagens em informação estruturada, de uma forma simples e descomplicada. A GCV *API*, em comparação com a extração de informação através de programação sem *API*, tem a vantagem de ser automática e não ser necessário criar um programa de raiz, porém tem a desvantagem de não ser gratuita e de apenas ter como resultado as opções pré-definidas na aplicação.

Alguns estudos utilizam a *Google Cloud Vision API* para analisar as cores e os objetos presentes nas imagens na publicação da rede social (Y. Li & Xie, 2020; Yu & Egger, 2021). Estes estudos salientam que numa publicação de uma rede social não só é importante o conteúdo da imagem, mas também o texto que a ela está associado, os sentimentos, os tópicos e os símbolos presentes na mesma. A junção destes dois tipos de informação texto e imagem é essencial para prever a atratividade da publicação (Y. Li & Xie, 2020). Especificamente, em relação às cores presentes numa imagem da publicação, salienta-se que fotos de comida local devem conter elementos vermelhos, apesar de, quando se trata de alta cozinha, essa relação não estar presente.

A inclusão de um esquema de cores azul pode ser usada para melhorar a atratividade, mais precisamente, nas fotografias de natureza. Já as fotografias relativas a obras de arte devem ter uma elevada variedade de cores. Conclui-se ainda que o violeta, juntamente com algumas cores quentes, pode influenciar positivamente o comportamento do consumidor e otimizar os resultados da atratividade das publicações (Yu & Egger, 2021).

2.4.2. Imobiliário

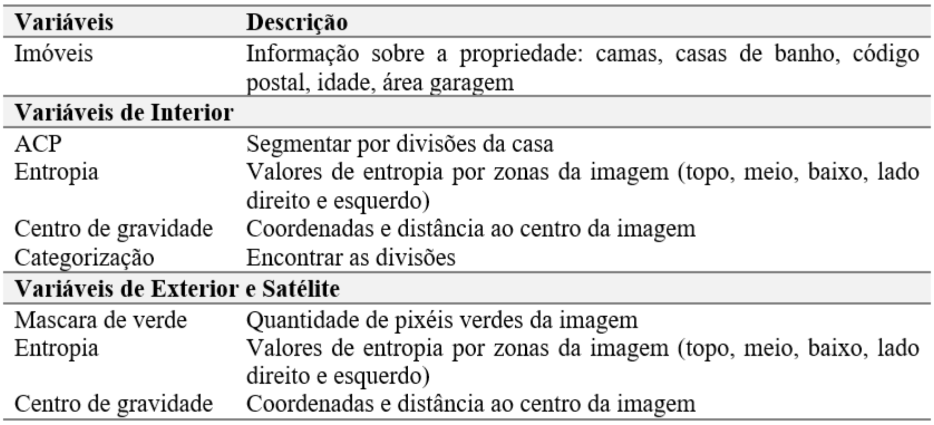
Existem artigos que tentam conjugar visão computacional com a área do imobiliário. Alguns continuam a ser muito descritivos, com o objetivo de descobrir padrões que ajudem na procura de casa apenas com base em características ou localização da mesma (M. Li *et al.*, 2018). No entanto, já começam a aparecer outros estudos que têm em conta não só as características das casas como área, localização, vizinhança, idade do alojamento, número de quartos e casas de banho, mas também as características das imagens (Gan *et al.*, 2021; Kostic & Jevremovic, 2020; You *et al.*, 2017).

As imagens da planta da casa podem ser analisadas para confirmar automaticamente a área construída, o número total de quartos e não-quartos, a posição dos quartos, o número total de portas e objetos, através de modelos de segmentação da imagem (Gan *et al.*, 2021).

As imagens de interior, exterior e de satélite podem ser analisadas com o objetivo de criar modelos de *deep learning*, que identificam automaticamente a divisão da casa, os objetos da imagem, as zonas mais importantes da imagem, as zonas verdes e também os valores calculados através das redes neuronais, para cada imagem (Kostic & Jevremovic, 2020; You *et al.*, 2017). Na Figura 7 podem observar-se algumas das imagens utilizadas em análises do âmbito imobiliário.

Figura 7. Imagens utilizadas nos estudos de processamento de imagem no imobiliário.
Fonte: You et al. (2017).

A partir dos vários tipos de imagens presente nos anúncios imobiliários extraem-se características de forma manual, como o número de fotografias (Larceneux *et al*., 2018) e de forma algorítmica, como a entropia, o centro de gravidade, as medidas associadas à cor como o HSV, entre outras (Kostic & Jevremovic, 2020). Na Figura 8 estão resumidas as características utilizadas num estudo com o objetivo de analisar a atratividade de um alojamento tendo em conta as características do mesmo e das imagens presentes no seu anúncio.



Variáveis	**Descrição**
Imóveis	Informação sobre a propriedade: camas, casas de banho, código postal, idade, área garagem
Variáveis de Interior	
ACP	Segmentar por divisões da casa
Entropia	Valores de entropia por zonas da imagem (topo, meio, baixo, lado direito e esquerdo)
Centro de gravidade	Coordenadas e distância ao centro da imagem
Categorização	Encontrar as divisões
Variáveis de Exterior e Satélite	
Mascara de verde	Quantidade de pixéis verdes da imagem
Entropia	Valores de entropia por zonas da imagem (topo, meio, baixo, lado direito e esquerdo)
Centro de gravidade	Coordenadas e distância ao centro da imagem

Figura 8.Variáveis importantes para prever a atratividade de um imóvel.
Fonte: Adaptado de Kostic & Jevremovic (2020), pág. 1911.

Este estudo conclui que as variáveis com mais importância são as características das casas, havendo, porém, algumas características extraídas das imagens, como a presença de cor verde nas imagens do exterior e de satélite, entropia média e o centro de gravidade em imagens de interior e exterior e também as categorias da divisão da fotografia, que sobem lugares na escala de importância, acabando por apresentar uma importância maior do que características do alojamento, como a existência de garagem ou o número de camas.

2.5. Atratividade dos alojamentos

É importante também ter em atenção outro tipo de características que podem influenciar as escolhas de alguma forma, como o tipo de consumidor e o segmento de mercado em análise (Larceneux *et al.*, 2018). Uma vez que as preferências dos consumidores podem ser subjetivas, tendo em conta a cultura e ambiente envolvente do consumidor (Wang & Sparks, 2016).

Quando o objetivo dos estudos é prever ou identificar relações entre as variáveis anteriormente descritas e a atratividade de um alojamento, é imperativo perceber o que é que os vários estudos entendem por um alojamento atrativo. Ao mesmo tempo é também importante perceber que tipo de técnicas, análises ou modelos são por norma utilizados neste contexto e as formas de avaliação dos modelos.

A definição da atratividade de uma casa varia de estudo para estudo e também consoante o âmbito da análise. Nos artigos menos analíticos e mais empíricos a atratividade é baseada nas opiniões das pessoas, recolhidas através de inquéritos (Lien *et al.*, 2015; Marder *et al.*, 2021). Já nos artigos mais analíticos são propostas diversas estratégias para medir a atratividade: através da taxa/número de cliques no anúncio/imagem e do número de dias no mercado do anúncio (Kostic & Jevremovic, 2020; Larceneux *et al*., 2018), ou também através do preço total ou por metro quadrado do imóvel (Kostic & Jevremovic, 2020; You *et al*., 2017).

Por fim, em relação às análises e técnicas utilizadas para estudar a atratividade de uma casa, são maioritariamente utilizadas previsões, e principalmente com recurso à regressão múltipla (Larceneux *et al*., 2018), à regressão simples e a "árvores impulsionadas" como XGBoost, LightGBM e CATBoost (Kostic & Jevremovic, 2020). Estes tipos de modelos são avaliados através de testes estatísticos, do Coeficiente de Determinação (R^2) e do Erro Médio Absoluto. As Métricas de avaliação utilizadas são poucas e demasiado simples, o que pode enviesar a análise, não permitindo assim perceber o que realmente acontece e se consegue prever bem todas as classes /casos. Como opção pode ser utilizado um alvo binário (atrativo ou não atrativo) para classificar o alojamentos, podendo assim utilizar-se a matriz da confusão e as métricas que dela se conseguem calcular (Sokolova & Lapalme, 2009).

2.6. Avaliação dos artigos da RSL

Na Tabela 7 está sintetizada a avaliação feita aos 23 artigos da RSL. Esta sumariza o quartil do jornal a que pertence, as principais limitações e contributos referidos em cada estudo, assim como a pontuação de cada critério de avaliação.

As limitações podem ser, por exemplo, questões relacionadas com o conjunto de dados, (e.g., conjunto demasiado pequeno ou com dados sintéticos), ou estar vinculadas ao intervalo de tempo dos dados da amostra (e.g., intervalo de tempo demasiado curto ou com uma granularidade demasiado alta). Os estudos podem contribuir para uma abordagem já existente, procurando aprofundar algo que já foi proposto, ou propor uma nova abordagem, que possivelmente poderá vir a ser aprofundada no futuro. As pontuações para cada critério de avaliação são de 0 a 1, conforme definido na secção 2.1.3.

A generalidade destes artigos encontra-se publicada em jornais do primeiro e segundo quartil. O artigo melhor classificado nesta revisão, (Kostic & Jevremovic, 2020), com ID=16, foi publicado na revista IEEE Transactions on Multimedia, classificada no quartil 1 há mais de vinte anos.

Tabela 7. Avaliação dos artigos da RSL.

ID	Q	Limitações	Contributos	C									Total
				1	2	3	4	5	6	7	8	9	
1	2	Dados	A.A.E.	1	1	0	0	0,5	1	0,5	1	1	6
2	1	Dados	Nova abordagem	1	1	0	1	1	0	0,5	1	1	6,5
3	3	Dados	A.A.E.	1	0,5	0	0	0	0,5	0,5	1	1	4,5
4	1	N.E.	Nova abordagem	1	0,5	1	0,5	0	0,5	1	0	1	5,5
5	1	N.E.	A.A.E.	0,5	1	0	0,5	0	0	0	0	0,5	2,5
6	2	Dados	A.A.E.	1	0,5	0	0	0,5	0	0,5	1	1	4,5
7	2	Dados	Nova abordagem	1	1	0	0,5	1	1	1	1	1	7,5
8	1	Extração de dados	A.A.E.	1	1	0	0	0,5	0	0,5	1	1	5
9	1	Dados	A.A.E.	1	0,5	0	0	0	1	0,5	0,5	0,5	4
10	3	Dados e período	A.A.E.	0,5	1	0	0	0	0,5	0,5	1	1	4,5
11	1	N.E.	A.A.E.	1	0	0	0	0	0	0	0	1	2
12	3	N.E.	A.A.E.	1	1	1	0,5	0,5	0	0,5	0	1	5,5
13	1	N.E.	A.A.E.	1	0,5	0	0	0	0	0,5	0	1	3
14	2	N.E.	A.A.E.	0,5	0,5	0,5	1	1	0	1	0	1	5,5
15	2	Dados	A.A.E.	1	1	1	1	0,5	0	1	1	0,5	7
16	1	Dados e período	A.A.E.	1	1	1	1	0,5	0,5	1	1	1	8
17	1	Dados	A.A.E.	1	1	1	0,5	1	0	1	1	0,5	7
18	3	Dados e período	A.A.E.	1	1	1	1	0	0	1	1	1	7
19	3	N.E.	A.A.E.	1	0,5	1	1	0	0	0	0	1	4,5
20	4	N.E.	A.A.E.	0,5	1	1	0,5	0,5	0	0,5	0	1	5
21	2	Dados	Nova abordagem	1	0,5	0	0	0,5	0,5	0	1	1	4,5
22	2	N.E.	A.A.E.	1	1	1	1	0,5	0	1	0	0,5	6
23	2	N.E.	A.A.E.	1	1	0	0	0,5	0	0	0	0,5	3
			Total	20	18	9.5	10	9	5,5	13	12,5	19	

Notas: Q – quartil do jornal; N.E. – Não especifica; A.A.E. – Aprofundamento de uma abordagem já existente

Através da avaliação dos vários artigos da RSL (Tabela 7), consegue-se perceber que os artigos mais relevantes para o presente estudo são os números 18, 16, 7, 15 e 17 (Gan *et al*., 2021; Kostic & Jevremovic, 2020; Larceneux *et al*., 2018; Y. Li & Xie, 2020; Yu & Egger, 2021).

Os artigos 15 e 17 (Y. Li & Xie, 2020; Yu & Egger, 2021), apesar de analisarem imagens turísticas publicadas nas redes sociais, permitem perceber a relação entre o conteúdo visual turístico e as interações dos utilizadores da plataforma. Por outro lado, e de forma complementar, os artigos 18, 16 e 7 (Gan *et al*., 2021; Kostic & Jevremovic, 2020; Larceneux *et al*., 2018), têm um âmbito ligado ao imobiliário, na ótica do cliente que pretende comprar uma casa e não alugar para turismo. Ainda assim, os três mostram quais são as características do alojamento que devem ser tidas em conta para encontrar padrões importantes. Os artigos 16 e 7 (Kostic & Jevremovic, 2020; Larceneux *et al*., 2018) mostram ainda como é que um alojamento pode ser considerado atrativo.

Através da Figura 9 pode-se constatar que as questões com maior pontuação são a questão 1, seguida da 9 e da 2, que estão relacionadas, respetivamente, com o contexto de estudo dos artigos, os seus contributos e a descrição das variáveis utilizadas. Contrariamente, verifica-se que as que têm pontuação mais baixa são as questões 6 e 5, relativas à explicação da relação entre as variáveis e a atratividade do imóvel e as características da imagem que influenciam os consumidores.

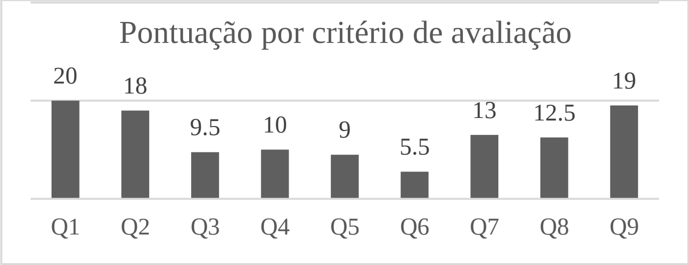


Figura 9. Pontuação por critério de avaliação.
Fonte: Elaboração própria, criado com recurso ao Excel.

2.7. Resposta às perguntas de pesquisa

Com base nos resultados da revisão sistemática realizada, já é possível responder às questões de pesquisa:

i) **"Quais os âmbitos e objetivos de análise onde mais se utiliza conteúdo de imagens?"**

O âmbito onde a análise de imagens é mais utilizada é o das atrações turísticas, seguido pelo imobiliário. O seu principal objetivo é prever a atratividade de uma imagem turística ou de um imóvel tendo em conta as suas características.

ii) **"Como é que o conteúdo visual é extraído das imagens e quais as principais técnicas aplicadas?"**

O conteúdo visual pode ser extraído das imagens através de questionários ou manualmente, extraindo informações como o número de fotos ou se a foto é amadora ou profissional. Por outro lado, as características da imagem podem também ser obtidas através da utilização de técnicas como algoritmos mais tradicionais, por exemplo, extração dos valores de HSV e RGB, entropia, presença de propriedades na linha do horizonte, centro de gravidade ou escalas de cinza ou com recurso a algoritmos mais complexos, como algoritmos de *deep learning*, para extrair categorias, objetos, cenários turísticos ou posição dos quartos na planta do alojamento.

iii) **"Como é definida a atratividade do alojamento e das imagens turísticas e quais as características com maior impacto na sua atratividade?"**

A atratividade é medida de diversas formas, alguns estudos utilizam questionários para obter a opinião das pessoas sobre atratividade, enquanto outros utilizam o número de cliques no site, os dias de mercado ou o preço, para ser mais objetivo e fácil de medir. As características mais importantes relacionadas à atratividade dos imóveis são suas características intrínsecas, seguidas de perto pelas imagens nos anúncios. Nas imagens turísticas é mais difícil compreender as características relevantes devido à falta de explicação nos modelos, mas é também referido que o conteúdo da imagem é muito importante.

iv) **"Quão maduras estão as metodologias de pesquisa seguidas pelos estudos mais relevantes?"**

Todos os estudos descreveram as suas contribuições, mas apenas cerca de metade possui uma metodologia rigorosa e replicável e descreve as limitações de seu estudo.

CAPÍTULO 3: Metodologia

O CRISP-DM (Pete *et al.*, 2000) é uma das metodologias mais utilizadas em projetos de *data mining*, principalmente em investigação nas áreas da saúde e educação. Esta metodologia foca-se principalmente nas primeiras fases, até à construção e avaliação de modelos (Schröer *et al.*, 2021), o que se adequa à persente dissertação, visto que esta se foca nos resultados e conclusões retirados do modelo final e não na implementação, num software, de um modelo em específico.

A Figura 10 permite visualizar todo o processo de um projeto de *data mining* de acordo com o CRISP-DM. Este processo é iterativo, composto por seis fases, sendo normal o recuo a fases anteriores (Pete *et al.*, 2000). As próximas secções sistematizam as principais tarefas de cada fase.

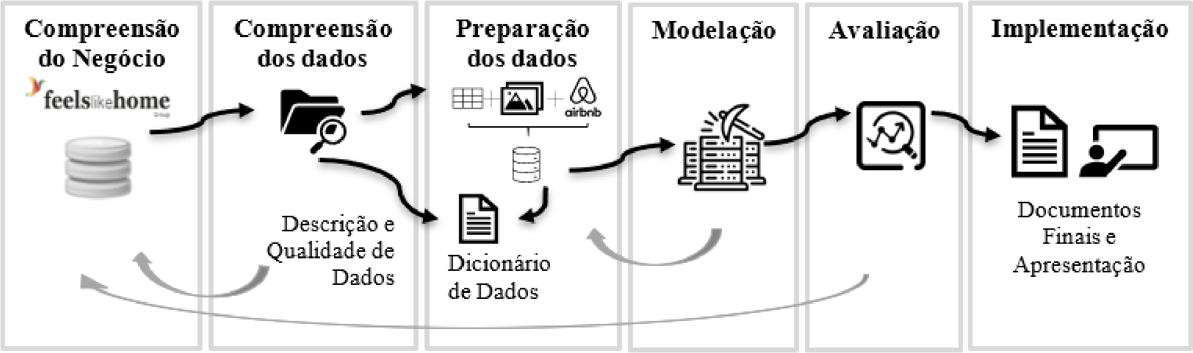


Figura 10. Processo metodológico.
Adaptado de Pete et al. (2000), pág. 12.

3.1. Compreensão do Negócio

A compreensão do negócio é a fase inicial de cada projeto. Visa conhecer e caracterizar a empresa, o problema e definir os objetivos do estudo. É também nesta fase que, através de reuniões com a empresa e da leitura de literatura existente, se define o plano da investigação, para o qual se tem sempre em conta tudo o que possa influenciar a pesquisa, como os recursos, requisitos, restrições, riscos e contingências, e as limitações, ferramentas e técnicas disponíveis. Nesta dissertação, a esta fase corresponde o que foi apresentado nos capítulos de introdução e de revisão da literatura.

3.2. Compreensão dos Dados

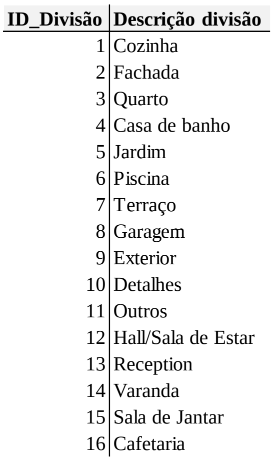
Na segunda etapa ocorre a compreensão dos dados, procedendo não só à recolha, mas também à descrição dos dados disponibilizados. Os dados utilizados na dissertação são considerados secundários (Saunders *et al.*, 2009), visto que a FLH já utilizava esses dados para outros propósitos.

O conjunto de dados utilizado é composto, mais precisamente, pela junção de um conjunto de dados estruturados, na forma de tabelas em ficheiros .csv, com um conjunto de dados não estruturados, na forma de imagens com formato .jpg.

Os dados estruturados, relativos aos alojamentos e suas reservas, são constituídos por:

- tabela *Casas*, com informação relativa a 1072 alojamentos,

- tabela *Reservas*, com informação relativa a 125341 reservas efetuadas,
- tabela *Airbnb_Search_to_listing*, com informação relativa ao número de cliques nas páginas do anúncio de 487 alojamentos, depois de este aparecer nos resultados de pesquisa,
- tabela *Airbnb_Listing_to_booking*, com informação sobre o número de visitantes distintos que visualizaram a página do anúncio e depois reservaram, com informação mensal relativa a 429 alojamentos,
- tabela *Airbnb_Avaliações* com informação da avaliação mensal, para 570 alojamentos,
- tabela das *NUTS*, tendo em conta a localização de cada alojamento,
- tabela *Divisões* da lista com o ID de cada divisão da imagem e a respetiva descrição, apresentada na Figura 11.



ID_Divisão	Descrição divisão
1	Cozinha
2	Fachada
3	Quarto
4	Casa de banho
5	Jardim
6	Piscina
7	Terraço
8	Garagem
9	Exterior
10	Detalhes
11	Outros
12	Hall/Sala de Estar
13	Reception
14	Varanda
15	Sala de Jantar
16	Cafetaria

Figura 11. Classificação das divisões das imagens.
Fonte: Elaboração própria.

Já os dados não estruturados dizem respeito a 13791 imagens de 697 alojamentos. Algumas das imagens estão aleatoriamente representadas na Figura 12, onde é possível notar alguma heterogeneidade de cor.

Figura 12. Visualização segmentada por cor, das imagens da FLH.
Fonte: Elaboração própria, criado com recurso ao programa ImageSorterV4

É também nesta fase que começa a elaboração do dicionário de dados e a análise da qualidade dos dados. A análise da qualidade de dados é uma das tarefas mais importantes, antes da modelação, dado que é através dela que se garante que os dados sejam não só os mais consistentes e precisos, mas também menos redundantes e com menos ruido possível.

Neste caso a qualidade dos dados passa pela correção de erros e verificação da:

- *Consistência*, ou seja, confirma-se se uma informação não está escrita de várias formas (ex. R/C e piso 0) ou se a localidade está correta tendo em conta o código postal.
- *Ambiguidade*, ou seja, verifica-se a existência de registos de reserva com informação de que o alojamento está inativo
- *Completude*, ou seja, se não existem valores omissos e no caso de existirem, qual a sua razão (esquecimento, engano, sem informação, etc.),
- *Conformidade*, ou seja, a facilidade de integração de toda a informação.

3.3. Preparação dos Dados

Na fase de preparação dos dados, procede-se à seleção, limpeza e transformação dos dados. São também criadas variáveis relevantes para a análise e no final procede-se à integração com a informação já existente, sendo criadas as tabelas para a fase seguinte, a modelação.

A literatura mostra que tanto as informações e características dos alojamentos(Gan et al., 2021; Kostic & Jevremovic, 2020; Larceneux et al., 2018; M. Li et al., 2018; Lien et al., 2015; Mariani & Borghi, 2018; You et al., 2017), como as informações das imagens (Gan et al., 2021; Kleinlein et al., 2019; Kostic & Jevremovic, 2020; Larceneux et al., 2018; M. Li & Fan, 2021; Y. Li & Xie, 2020; You et al., 2017; Yu & Egger, 2021), devem ser tidas em conta, por isso, procurou-se integrar ambos os conjuntos de informações.

3.3.1. Conversão de dados não estruturados para estruturados

Dado que se pretendia criar uma tabela com os alojamentos organizados por linha, o primeiro passo foi extrair o conteúdo visual das imagens, de modo a obter informação estruturada por imagem que foi guardada na tabela *Imagens*. A informação das imagens foi extraída através de algoritmos escritos e trabalhados em *python*. Estes permitiram obter as seguintes informações:

- cores presentes em cada imagem e a percentagem de pixéis da imagem associados a uma cor específica. Para tal recorreu-se a máscaras que foram aplicadas no espaço de cores HSV. Na Figura 13 está representada a escala utilizada para fazer a correspondência entre os valores de H (Hue) e as cores (Kostic & Jevremovic, 2020; Y. Li & Xie, 2020; Yu & Egger, 2021);
- luminosidade média, obtida através da componente L (Lightness) do espaço de cores HSL da imagem, (Kleinlein *et al.*, 2019; Yu & Egger, 2021);

- média e desvio padrão da entropia da imagem (Figura 14), obtida através da função de cálculo da entropia da biblioteca *skimage* (Kleinlein *et al*., 2019; Kostic & Jevremovic, 2020);
- cinco clusters por imagem, que representam as 5 cores mais presentes na mesma, através da associação de cor, tendo em conta o espaço de cor HSV (Kostic & Jevremovic, 2020);
- ID da casa a que pertence, divisão associada à fotografia e ordem na qual a imagem aparece no anúncio. (Gan et al., 2021; Kostic & Jevremovic, 2020; Larceneux et al., 2018). Estes dados foram extraídos a partir do nome do ficheiro.

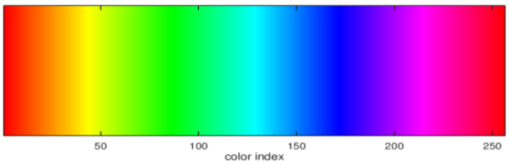


Figura 13. Correspondência entre os valores de H (Hue) e as cores.
Fonte: retirado de (Forge, n.d.)

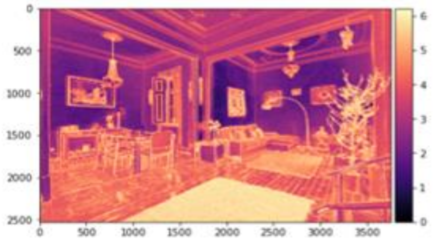
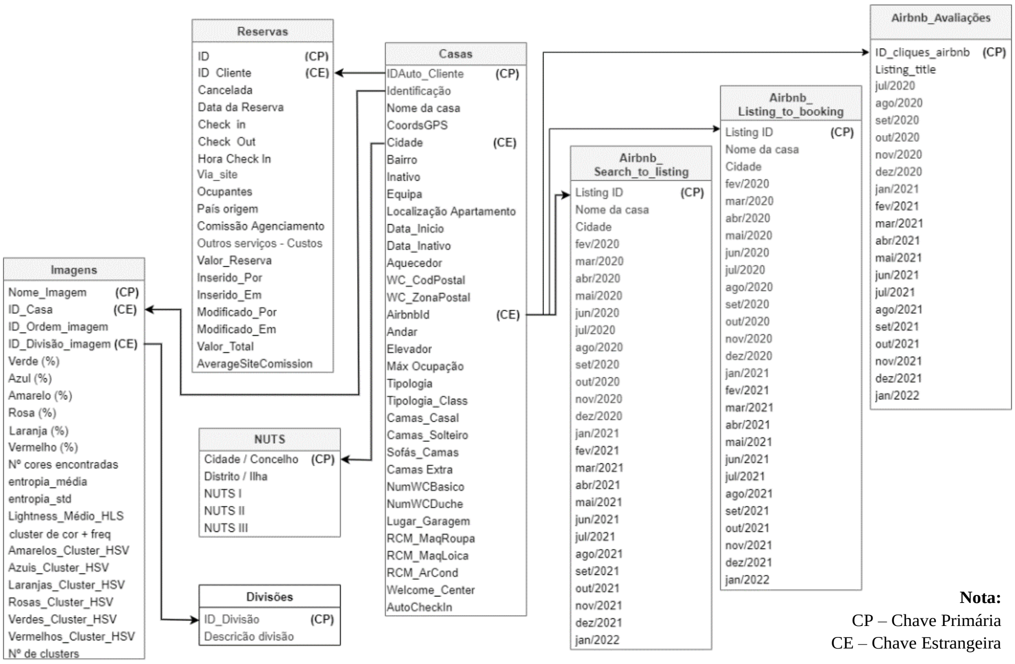


Figura 14. Entropia de uma imagem de exemplo.
Fonte: Elaboração própria, em *python*

Na Figura 15 é possível observar o modelo de dados inicial, contendo a informação estruturada inicial e também a informação extraída das imagens. Todas as variáveis deste modelo de dados, respetivo tipo, descrição e ação tomada estão presentes no Anexo A.


Reservas
ID (CP)
ID Cliente (CE)
Cancelada
Data da Reserva
Check in
Check Out
Hora Check In
Via_site
Ocupantes
País origem
Comissão Agenciamento
Outros serviços - Custos
Valor_Reserva
Inserido_Por
Inserido_Em
Modificado_Por
Modificado_Em
Valor_Total
AverageSiteComission
Casas
IDAuto_Cliente (CP)
Identificação
Nome da casa
CoordsGPS
Cidade (CE)
Bairro
Inativo
Equipa
Localização Apartamento
Data_Inicio
Data_Inativo
Aquecedor
WC_CodPostal
WC_ZonaPostal
AirbnbId (CE)
Andar
Elevador
Máx Ocupação
Tipologia
Tipologia_Class
Camas_Casal
Camas_Solteiro
Sofás_Camas
Camas Extra
NumWCBasico
NumWCDuche
Lugar_Garagem
RCM_MaqRoupa
RCM_MaqLoica
RCM_ArCond
Welcome_Center
AutoCheckIn
Imagens
Nome_Imagem (CP)
ID_Casa (CE)
ID_Ordem_imagem
ID_Divisão_imagem (CE)
Verde (%)
Azul (%)
Amarelo (%)
Rosa (%)
Laranja (%)
Vermelho (%)
Nº cores encontradas
entropia_média
entropia_std
Lightness_Médio_HLS
cluster de cor + freq
Amarelos_Cluster_HSV
Azuis_Cluster_HSV
Laranjas_Cluster_HSV
Rosas_Cluster_HSV
Verdes_Cluster_HSV
Vermelhos_Cluster_HSV
Nº de clusters
NUTS
Cidade / Concelho (CP)
Distrito / Ilha
NUTS I
NUTS II
NUTS III
Divisões
ID_Divisão (CP)
Descrição divisão
Airbnb_Search_to_listing
Listing ID (CP)
Nome da casa
Cidade
fev/2020
mar/2020
abr/2020
mai/2020
jun/2020
jul/2020
ago/2020
set/2020
out/2020
nov/2020
dez/2020
jan/2021
fev/2021
mar/2021
abr/2021
mai/2021
jun/2021
jul/2021
ago/2021
set/2021
out/2021
nov/2021
dez/2021
jan/2022
Airbnb_Listing_to_booking
Listing ID (CP)
Nome da casa
Cidade
fev/2020
mar/2020
abr/2020
mai/2020
jun/2020
jul/2020
ago/2020
set/2020
out/2020
nov/2020
dez/2020
jan/2021
fev/2021
mar/2021
abr/2021
mai/2021
jun/2021
jul/2021
ago/2021
set/2021
out/2021
nov/2021
dez/2021
jan/2022
Airbnb_Avaliações
ID_cliques_airbnb (CP)
Listing_title
jul/2020
ago/2020
set/2020
out/2020
nov/2020
dez/2020
jan/2021
fev/2021
mar/2021
abr/2021
mai/2021
jun/2021
jul/2021
ago/2021
set/2021
out/2021
nov/2021
dez/2021
jan/2022

Nota:
CP – Chave Primária
CE – Chave Estrangeira

Figura 15. Modelo de dados com integração de informação das imagens.
Fonte: Elaboração própria, com recurso ao programa app.diagrams.net

3.3.2. Limpeza, transformação e criação de variáveis

Nesta fase, é realizada a limpeza e correção dos erros e inconsistências encontrados nos registos resultantes da fase anterior. Decidiu-se nãoconsiderar para análise as linhas da tabela *Casas* que não apresentavam valores sobre os cliques na página do alojamento ou que não continham fotografias associadas. Na tabela *Reservas* foram removidas as linhas de reservas que apresentavam valores de reservas nulos ou que correspondiam a reservas canceladas pelos hóspedes.

Foram também realizadas as transformações e criação de variáveis, efetuando-se a limpeza quando necessário, mais precisamente em relação às variáveis. Foram excluídas as variáveis que teoricamente não faziam sentido integrar na análise, como as que não tinham qualquer relação com o alvo ou as que a sua distribuição não permitiria retirar qualquer tipo de informação relevante, como o nome do colaborador e a data de inserção e modificação da reserva, o nome do alojamento e coordenada de GPS, visto que existia essa informação noutras variáveis como *Cidade* e *Bairro*.

Foram ainda analisados os valores nulos e os valores extremos. Os primeiros apenas existiam na tabela *Rating* e foram corrigidos para a média. Já os segundos, nas variáveis que o permitiam foram criadas outras variáveis com categorizações. Na Figura 16 podem-se observar três variáveis com valores extremos que foram por isso transformadas. O *Andar* foi transformado numa variável com quatro categorias: cave, andares baixos, andares médios e andares altos. A *Ocupação Máxima* foi transformada numa variável com três categorias: baixa, média e alta. O *Valor médio por noite*, foi também transformado numa variável com três categorias: baixo, médio e elevado. Dado que os modelos utilizados na modelação são árvores de decisão, os valores extremos não são um problema, logo ambas as variáveis, categóricas ou não, são testadas.

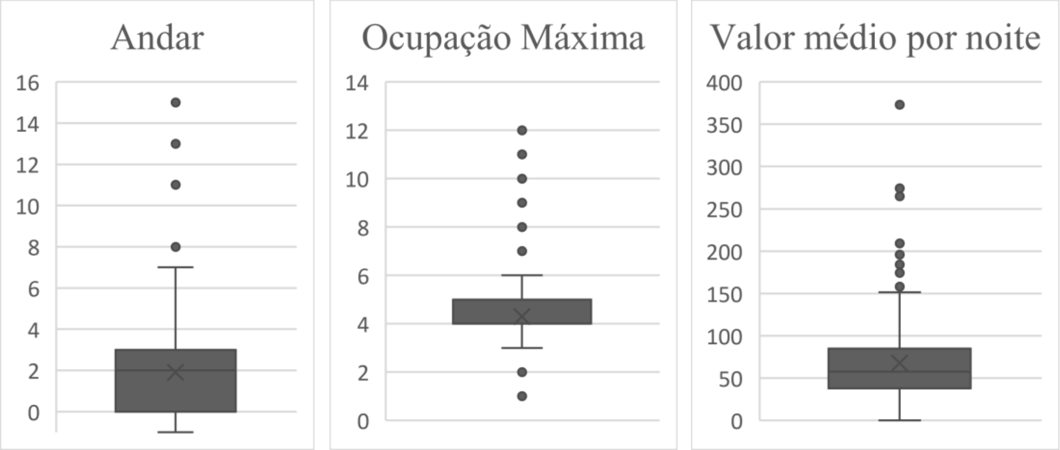


Figura 16. Outliers das variáveis andar, ocupação máxima e valor médio por noite.
Fonte: Elaboração própria, criado com recurso ao Excel

Ainda nesta fase, são efetivamente criadas as variáveis a incluir na tabela de modelação, definidas através de reuniões com a FLH e orientadores. Uma vez que a modelação cai sobre a atratividade percecionada no anúncio do alojamento, esta é medida através da quantidade de cliques sobre a página do alojamento. Como tal, as variáveis que podem influenciar a atratividade devem replicar toda a informação sobre o alojamento presente no anúncio. Para se definir quais as informações presentes que podem influenciar o clique realizaram-se pesquisas no *site* do airbnb[1].

Numa primeira pesquisa por localidade, por exemplo, é apresentada a lista de alojamentos da zona, como se pode observar na Figura 17. A informação de destaque é: a localização, a primeira fotografia, o valor por noite ou o valor da estadia, a avaliação e algumas características, tais como o número de camas.

Quando se clica numa página de um alojamento específica, como exemplificado na Figura 18, a informação de destaque na página do alojamento é bastante semelhante, mas mais pormenorizada, ou seja, são apresentadas 5 fotografias, por norma, e as características do alojamento também são mais detalhadas, apresentando não só o número de camas, mas também o número de quartos, casas de banho e hóspedes.

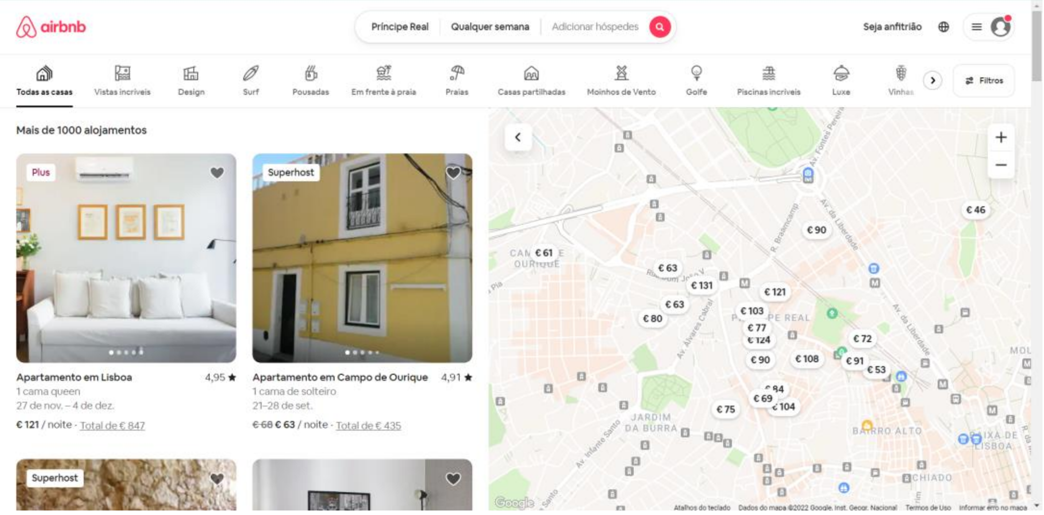

airbnb
Príncipe Real
Qualquer semana
Adicionar hóspedes
Seja anfitrião
Todas as casas
Vistas incríveis
Design
Surf
Pousadas
Em frente à praia
Praias
Casas partilhadas
Moinhos de Vento
Golfe
Piscinas incríveis
Luxe
Vinhas
Filtros
Mais de 1000 alojamentos
Plus
Superhost
Apartamento em Lisboa
4,95 ★
1 cama queen
27 de nov. – 4 de dez.
€ 121 / noite · Total de € 847
Apartamento em Campo de Ourique
4,91 ★
1 cama de solteiro
21–28 de set.
€ 68 € 63 / noite · Total de € 435
Superhost
€ 46
€ 90
€ 61
€ 63
€ 131
€ 121
€ 63
€ 80
€ 103
€ 77
€ 124
€ 72
€ 90
€ 108
€ 91
€ 53
€ 84
€ 69
€ 104
€ 75
JARDIM DA BURRA
BAIRRO ALTO
BAIXA DE LISBOA
CHIADO
Google
Atalhos do teclado
Dados do mapa ©2022 Google, Inst. Geogr. Nacional
Termos de Uso
Informar erro no mapa

Figura 17. Lista de alojamentos numa primeira pesquisa.
Fonte: Retirado de Airbnb.pt.

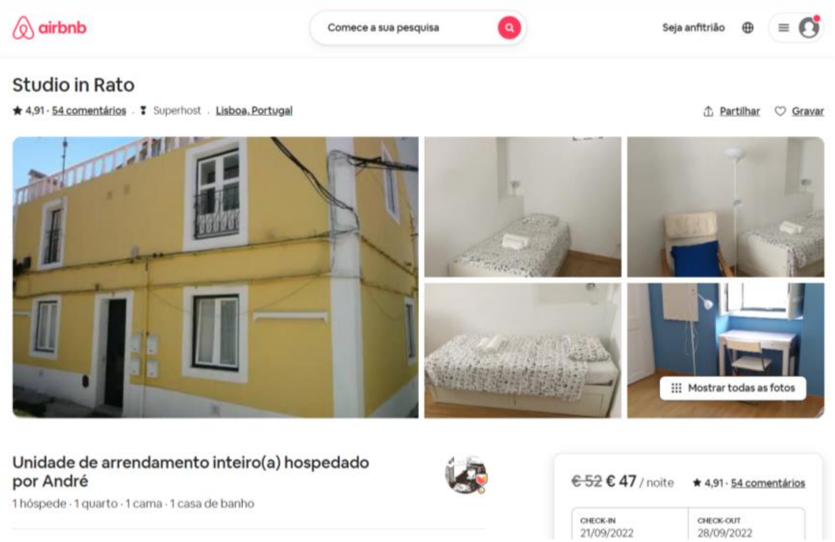

airbnb
Comece a sua pesquisa
Seja anfitrião
Studio in Rato
★ 4,91 · 54 comentários · Superhost · Lisboa, Portugal
Partilhar
Gravar
Mostrar todas as fotos
Unidade de arrendamento inteiro(a) hospedado por André
1 hóspede · 1 quarto · 1 cama · 1 casa de banho
€ 52 € 47 / noite
★ 4,91 · 54 comentários
CHECK-IN
21/09/2022
CHECK-OUT
28/09/2022

Figura 18. Página de Airbnb de um alojamento.
Fonte: Retirado de Airbnb.pt.

---

[1] https://www.airbnb.pt/

3.3.3. Alvo

O principal requisito na definição do alvo passa por representar a atratividade do alojamento da forma mais objetiva possível, ou seja, usar um alvo baseado em algum valor associado ao alojamento ou ao anúncio da mesma, já estudado pela literatura, e não a um inquérito.

Com base nos artigos estudados, algumas das opções utilizadas para alvo são o preço ou o número de dias no mercado (Kostic & Jevremovic, 2020; Larceneux et al., 2018; You et al., 2017). Porém, no presente estudo não foram utilizadas estas opções, visto que não se pretendia estudar características associadas à reserva, mas sim à atratividade. Para além disso, existem vários fatores externos que influenciam a reserva, mas que podem não influenciar a atratividade do alojamento, tais como a disponibilidade financeira do consumidor ou a disponibilidade do alojamento na altura das suas férias.

Concluiu-se então que, de todas as informações disponíveis, os cliques no anúncio, também utilizada por Larceneux et al. (2018), é a informação que melhor representa o interesse ou atração pelo alojamento, mesmo que esta não se venha a transformar posteriormente em intenção de reserva.

Sendo assim, a variável alvo foi construída subtraindo a informação das tabelas *Airbnb_Search_to_listing* e *Airbnb_Listing_to_booking*, procurando garantir que o alvo represente o interesse ou atração pelo alojamento, isto é, pretende diminuir não só o impacto da intenção de compra, como o de outros fatores presentes na página do alojamento que não estejam ligados diretamente à atratividade intrínseca do alojamento, como por exemplo os comentários. Por fim, o alvo foi transformado numa variável binária.

O principal requisito na procura pelo valor do limiar para separação binária do alvo, prendia-se com a criação de dois grupos homogéneos distintos: um para representar os alojamentos mais atrativos e outro para os menos atrativos. A técnica utilizada para encontrar esse limiar foi a segmentação segundo dois grupos da variável criada para o alvo, através do algoritmo *K-means*, o resultado deste modelo encontra-se no Anexo B.

É importante referir que tanto o alvo criado através da subtração como o criado apenas com informação da *Airbnb_Search_to_listing*, seguindo a mesma lógica, foram testados na modelação. Foi então escolhido o alvo criado pela subtração porque, para além de apresentar resultados mais consistentes, isola a influência de outros fatores que condicionam a compra da reserva.

3.3.4. Seleção de variáveis

Com o propósito de selecionar as variáveis que devem ser utilizadas na fase seguinte, estas foram alvo de análises de qualidade, tanto em relação aos registos como às colunas, assim como análises bi-variadas juntamente com a variável alvo. Por fim, para compreender o grau de associação e correlação entre as variáveis (quantitativas e qualitativas nominais) e o alvo (qualitativa nominal) foram realizadas duas análises.

A primeira com o objetivo de medir a associação entre as variáveis qualitativas nominais (ou tratadas como tal) e o alvo, tendo sido utilizada a medida de associação V de Cramer, uma medida baseada no teste da independência do Qui-quadrado. A segunda análise com o objetivo de medir a correlação entre as variáveis quantitativas e o alvo, tendo sido foi utilizada a medida ETA, que permite analisar a relação de dependência entre as variáveis. No caso de as medidas apresentarem o valor de 0 significa que existe ausência de relação ou que as vaiáveis são independentes e, contrariamente, se o valor for 1 existe dependência ou uma relação perfeita entre as duas variáveis. A escala de relação adotada foi a seguinte: entre 0 e 0,2 – relação muito fraca; entre 0,2 e 0,4 – relação fraca; entre 0,4 e 0,7 – relação moderada; entre 0,7 e 0,9 – relação forte; e entre 0,9 e 1 – relação muito forte (Laureano, 2020).

De referir que estas análises são bi-variadas, servem para excluir as variáveis que não têm qualquer ligação ao algo, i.e., são completamente independentes, ou que mostrem uma grande correlação entre outras variáveis, apresentando assim uma redundância de informação. Contudo uma variável que tenha uma fraca relação estatística com o alvo, pode ainda assim, ser inserida nos modelos, visto que através da interação com uma outra variável pode tornar-se importante.

Esta fase acaba com a construção da tabela para modelação, cuja informação está resumida na Tabela 8, e o dicionário de dados. Nos Anexo C e Anexo Dencontram-se as medidas de associação e o dicionário de dados com a descrição e análises descritivas de todas as variáveis selecionadas para a fase da modelação.

Tabela 8. Informação selecionada para a fase da modelação.

Alojamentos		Atratividade (cliques), localização, andar, nº quartos e casas de banho, ocupação máxima, se tem elevador, ar condicionado e/ou garagem, avaliação e valor por noite.
Imagens	Para todas as fotografias do anúncio	Total de fotografias, divisão associada a cada uma das cinco primeiras fotografias, cores, entropia média e desvio padrão da entropia, luminosidade média.
	Para a primeira fotografia do anúncio	Divisão presente na primeira fotografia, cores, entropia média e desvio padrão da entropia, luminosidade média.
	Para as cinco primeiras fotografias do anúncio	Divisão associada a cada uma das cinco primeiras fotografias, cores, entropia média e desvio padrão da entropia, luminosidade média.

3.4. Modelação

Na fase modelação selecionam-se as técnicas de modelação a utilizar tendo em conta os objetivos iniciais do estudo. Desenha-se o procedimento de avaliação, constroem-se vários modelos para que sejam avaliados e revistos, começando assim uma iteração com a fase da avaliação.

No âmbito deste trabalho foram aplicadas técnicas de análise supervisionada e não supervisionada. Dada a natureza iterativa da metodologia utilizada, manteve-se em aberto a possibilidade de refazer algum passo ou até mesmo voltar à fase anterior para melhorar ou criar alguma variável.

Inicialmente, para segmentar os alojamentos da FLH, foram utilizadas técnicas não supervisionadas, de segmentação, com o propósito de identificar se existe algum padrão entre os alojamentos e se sim quais são as principais características dos mesmos.

Os algoritmos de segmentação procuram agrupar os registos de todo o conjunto de dados em grupos (*clusters*), com registos semelhantes entre si e diferentes dos das outras (Larose & Larose, 2015; Quinn, 2020). A Figura 19 mostra de forma muito simplificada a logica de separação dos modelos de segmentação.

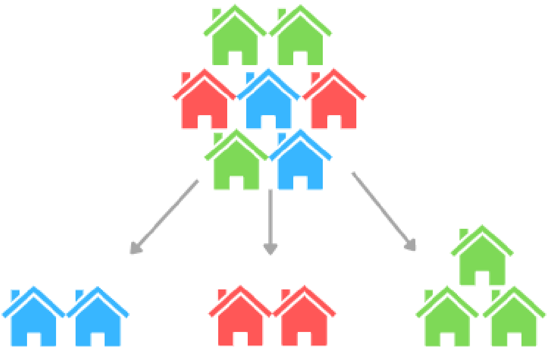


Figura 19. Exemplo de segmentação de elementos.
Fonte: Elaboração própria, com recurso à aplicação canva.

A segunda fase da modelação passou por juntar a informação sobre atratividade, com o objetivo de criar os perfis dos alojamentos da FLH mais atrativas, perceber quais são as características mais importantes, e se informação das imagens está entre essas características. Para tal foram utilizadas técnicas supervisionadas de classificação.

Na classificação utilizou-se, por norma, uma variável categórica alvo, neste caso a atratividade, dividida segundo duas classes predeterminadas: atrativo e não atrativo. O modelo examina um conjunto de registos, contendo informações sobre a variável alvo, bem como um conjunto de variáveis de entrada, as variáveis criadas e/ou transformadas na fase anterior. Desta forma, no treino, o algoritmo aprende qual a combinação de variáveis que está associada, por exemplo, à atratividade. De seguida, no teste, o algoritmo produz classificações automáticas para novos registos, comparando-as com as classificações reais (Larose & Larose, 2015).

As árvores de decisão são algoritmos de classificação, ainda que algumas consigam funcionar com alvos contínuos. São técnicas muito utilizadas porque produzem modelos transparentes e intuitivos que são relativamente fáceis de controlar e entender (Berthold et al., 2020; Larose & Larose, 2015; Quinn, 2020).

Os algoritmos utilizados na fase de modelação foram os seguintes:

- CART, criado por Breiman *et al*. em 1984. As árvores de decisão produzidas pelo CART são estritamente binárias, dividindo em exatamente dois ramos cada nó de decisão (Larose & Larose, 2015).
- C5.0, que é uma atualização do algoritmo C4.5, a extensão de Quinlan de seu próprio algoritmo ID3. Neste as divisões não são necessariamente binárias e para variáveis categóricas produz ramificações para cada categoria, por isso produz uma árvore de formato mais variável. (Larose & Larose, 2015).

- CHAID, que significa Deteção Automática de Interação Qui-Quadrado e foi projetado em 1980 por Gordon V. Kass. É uma técnica especialmente popular porque é baseada em um dos testes de significância estatística mais usados – o teste Qui-Quadrado de Pearson. (Quinn, 2020).
- QUEST, proposto por Loh e Shin em 1997. A velocidade de cálculo neste algoritmo é maior do que em outros métodos; este algoritmo também pode evitar o enviesamento que existe noutros métodos. O algoritmo é mais adequado para variáveis de categoria múltipla e faz partições binárias. (Lin & Fan, 2019)

As Figuras 20 e 21 ilustram exemplos de árvores de decisão, com decisões binárias e não necessariamente binárias, respetivamente. Através das figuras é ainda possível entender a anatomia de uma árvore de decisão. Cada imagem de uma casa representa um nó, que corresponde a um conjunto de elementos com certas características. Os nós que dão origem a outros nós são os nós pai, e consequentemente, os subsequentes são os nós filho. As setas que ligam os nós, são os ramos e o os últimos nós são as folhas ou nós terminais.

Com a diminuição do número mínimo de elementos nos nós pai e/ou filhos, permite-se que a árvore cresça, seja em profundidade, seja em número de regras/nós terminais. É também possível verificar que, nestes exemplos, a profundidade das árvores é de dois níveis. No entanto numa situação real as árvores têm normalmente mais níveis e, à medida que o número de níveis aumenta, a árvore fica mais profunda e com mais regras, e consequentemente mais complexa.

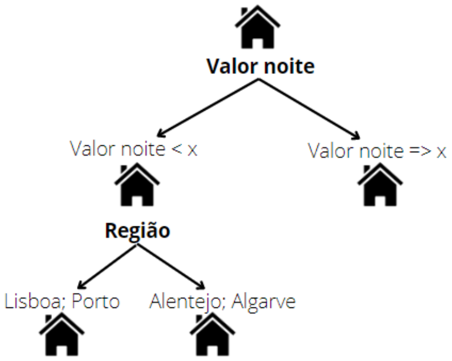


Figura 20. Exemplo de árvore de decisão binária.
Fonte: Elaboração própria, com recurso à aplicação *canva*.

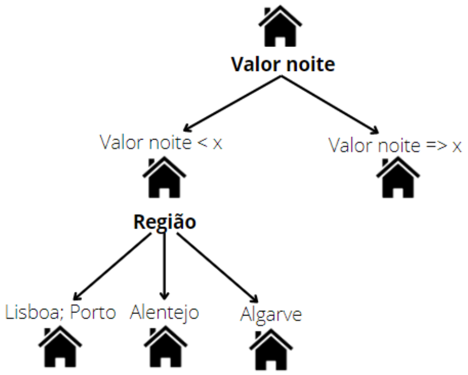


Figura 21. Exemplo de árvore de decisão não necessariamente binária.
Fonte: Elaboração própria, com recurso à aplicação *canva*.

A principal razão da escolha dos modelos a utilizar recair sobre as árvores de decisão, apresentadas anteriormente, prende-se com a necessidade de garantir a explicabilidade dos resultados para com a Feels Like Home. Dessa forma não foram escolhidos algoritmos mais complexos como redes neuronais ou com aprendizagens mais profundas, dado que estes muito dificilmente permitiram explicar quais as razões e regras que levam o algoritmo a dizer que certo alojamento é ou não atrativo.

3.5. Avaliação

A fase de avaliação está subdividida segundo várias etapas, em que algumas delas começam ainda durante a fase da modelação, mais precisamente a avaliação dos modelos através de métricas específicas.

No caso dos modelos de segmentação, estes são avaliados não só através da análise do conteúdo dos clusters e se teoricamente fazem sentido, mas também através do coeficiente de silhueta, que permite avaliar o grau de semelhança dentro dos grupos e o grau de diferenças entre si, com o intuito de medir quão boa é a atribuição da classe aos registos da mesma.

Um valor positivo mais elevado indica que a atribuição é boa, enquanto que um valor próximo de zero é considerado uma atribuição fraca, já um valor negativo é considerado mal classificado (Larose & Larose, 2015).

No que toca aos modelos de classificação, estes exigem um processo ainda mais rigoroso. O procedimento de avaliação das arvores de decisão passa pela divisão da tabela em dois conjuntos, o de treino e o de teste, com uma proporção de 70% / 30%, respetivamente (Quinn, 2020).

O conjunto de treino é balanceado através do aumento de registos da classe minoritária (Berthold *et al*., 2020), até que se apresente a mesma percentagem de casos positivos e negativos. Posteriormente, o modelo treinado é usado para efetuar previsão no conjunto de teste, permitindo a avaliação dos resultados da classificação versus valores reais.

Para os modelos que permitem a funcionalidade, é aplicada validação cruzada, os dados originais são particionados em k subconjuntos independentes, de dimensão e distribuição de classes semelhante, e o modelo é então construído usando os dados de k−1 subconjuntos para treino, usando-se o k-ésimo subconjunto como conjunto de teste. O resultado dos testes será uma média dos resultados dos k modelos executados (Larose & Larose, 2015), sabendo-se assim, para cada modelo, qual a média de acerto total e a média do erro dos vários conjuntos.

Para os modelos que em que são utilizados os conjuntos treino e teste, os erros de classificação são avaliados através de métricas contruídas a partir da matriz de confusão. A Tabela 9 mostra a matriz de confusão, onde as linhas representam as classificações reais e as colunas representam os valores previstos (Berthold *et al*., 2020).

Tabela 9. Matriz de Confusão.

		Previsão	
		0	1
Real	0	Verdadeiros Negativos (VN)	Falsos Positivos (FP)
	1	Falsos Negativos (FN)	Verdadeiros Positivos (VP)

As métricas utilizadas na avaliação dos modelos são descritas de seguida.

A Exatidão ou *accuracy* (3.5.1) que mede a taxa de acerto total. Quanto mais próximo de 1 mais semelhantes são as previsões dos valores reais. Esta utiliza todas as classes da matriz de confusão no seu cálculo (Berthold *et al*., 2020) e é dada por:

$$Exatidão = \frac{VN + VP}{VN + VP + FN + FP}\,. \tag{3.5.1}$$

A Sensibilidade ou *recall* (3.5.2) que mede a capacidade do modelo reconhecer corretamente a classe positiva, e a Especificidade (3.5.3) que mede a capacidade do modelo reconhecer o que não pertence à classe positiva (Berthold *et al*., 2020). Estas métricas são dadas por:

$$Sensibilidade \ = \frac{VP}{VP + FN}\,; \tag{3.5.2}$$

$$Especificidade \ = \frac{VN}{VN + FP}\,. \tag{3.5.3}$$

A Precisão (3.5.4) de uma classe é o rácio entre o número de registos corretamente classificados como pertencentes a essa classe e o número total de registos classificados como sendo dessa classe. Esta também pode ser considerada uma medida de exatidão (Berthold *et al*., 2020). No caso de classificação binária, e em relação à classe positiva, a precisão é dada por:

$$Precisão \ = \frac{VP}{VP + FP}\,. \tag{3.5.4}$$

Finalmente, a Medida-F (3.5.5) pode ser interpretada como uma média ponderada entre a precisão e sensibilidade, que atinge o seu melhor valor quando é 1 e o pior quando é 0 (Berthold *et al*., 2020).

$$F = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \tag{3.5.5}$$

São ainda tidas em conta as métricas *area under curve* AUC, ou seja, a área por baixo da curva ROC, um outro indicador de quão bem o classificador resolve o problema. A curva ROC descreve a relação, tendencialmente inversa, entre a sensibilidade e a especificidade, mais precisamente curva é a taxa de verdadeiros positivos (sensibilidade) e função da taxa de falsos positivos (1- Especificidade). Quanto maior a área por baixo da curva, melhor a solução para o problema de classificação (Berthold *et al*., 2020).

É importante salientar que se optou por utilizar um grande conjunto de métricas com o intuito de controlar, da melhor forma possível, os erros de cada modelo, visto que, apesar de todas as métricas serem importantes, a forma de cálculo de algumas pode induzir em erro. Por exemplo, a exatidão é uma das medidas mais utilizadas, mas que pode ser facilmente enviesada caso exista uma grande diferença na quantidade de casos de cada classe do alvo.

Logo, pode-se dizer que apesar de todas as métricas serem tidas em conta, a medida-F e a especificidade serão as duas que terão um maior peso, primeiro a medida-F, porque engloba a precisão e a sensibilidade e depois, como complemento, a especificidade, porque se foca na classe que falta, a classe 0 (não atrativos).

As métricas são analisadas tanto para o conjunto de treino, como para o conjunto de teste de forma a perceber como podem ser melhorados os modelos. No caso de modelos sobreajustados, com resultados bons no treino mas baixos no teste, a solução poderá passar por experimentar valores um pouco mais altos para o número mínimo de casos dos nós pais e filhos ou por reduzir a profundidade, não permitindo que o modelo se torne tão complexo e assim "demasiado" ajustado aos dados de treino. No caso dos modelos estarem subajustados, ou seja, não conseguirem captar padrões e que ainda apresentarem uma clara margem para aprender no treino, pode ser uma opção reduzir o número mínimo dos nós filhos e deixar a árvore crescer, permitindo que o modelo se torne mais complexo.

Esta fase compreende ainda a avaliação e discussão dos resultados obtidos, a sua comparação com os critérios analíticos e de sucesso do negócio, a validação por parte do negócio e a comparação com os resultados presentes na literatura.

Por fim, são determinados os próximos passos, refletindo sobre todo o procedimento realizado no estudo e verificando se ainda existem pormenores a serem melhorados, como variáveis ou análises que obriguem a voltar a alguma fase anterior. Caso isso não aconteça, pode-se dar por encerrado o estudo.

3.6. Implementação

Na última fase, a implementação, é crucial sumarizar as conclusões e contributos tanto científicos como para o negócio, as limitações, as recomendações e os passos futuros. Este conteúdo é apresentado nos dois capítulo seguinte, onde primeiro são descritos e discutidos os resultados a que se chegaram e depois, na conclusão, é sumarizado todo o trabalho realizado, recomendações, limitações e sugestões de pesquisas futuras.

Neste caso a implementação culminou na escrita e apresentação dos resultados da investigação, ou seja, na elaboração do documento da dissertação e sua apresentação. Adicionalmente originou também a escrita e publicação de um artigo científico.

Apesar que já ser fora do âmbito desta dissertação, poderia também fazer parte da implementação a colocação do modelo final em produção, ou seja, o modelo poderia ser colocado no software da FLH, ajudando assim na tomada de decisão. Quando se pretendesse estudar a atratividade dos alojamentos atuais, com base em todas as características que entraram no modelo, bastaria correr o modelo e este seria capaz de identificar os alojamentos considerados mais atrativos e com que probabilidade.

3.7. Testes alternativos

No âmbito deste trabalho foram realizadas variações às análises já mencionadas, para as quais não foram apresentadas metodologias detalhadas ou resultados, por terem sido obtidos modelos com piores avaliações. Ainda assim, julga-se conveniente referir brevemente a metodologia aplicada nessas análises.

Uma alternativa foi a criação de uma tabela com chave primária casa + mês, pretendendo-se ter em conta a sazonalidade e eliminar as variáveis construídas pelas médias mensais, visto que uma casa com piscina ou perto do mar poderia ser mais atrativa nos meses de época alta do que nos de época baixa. Esta alternativa não teve sucesso muito possivelmente pelo ruído que pode ter sido gerado, dado que quando se cria um registo para cada casa por cada mês, há repetição de uma grande parte da informação que é comum em todos os meses.

Testou-se também a utilização da mediana como valor de limiar para o alvo, obtendo-se um conjunto de dados totalmente balanceado entre as classes atrativa e não atrativa, mas que conduziu a resultados piores. Em particular, observou-se uma grande diminuição no valor da especificidade passando esta a ficar inferior à sensibilidade. Ao considerar que 50% dos alojamentos são atrativos, os modelos não conseguiram encontrar padrões específicos, o que não é estranho dado que os alojamentos mais atrativos são bastante diferentes e também têm valores de atratividade muito díspares entre si. Muito possivelmente o fator diferenciador varia de alojamento para alojamento, e os únicos fatores que podem ser comuns a um maior número de alojamentos são o valor por noite e a localização, dado que são os mais valorizados pelos modelos.

Houve também a tentativa de modelar atribuindo custos de classificação errada, ou seja, atribui-se um peso maior aos Falsos Negativos ou aos Falsos Positivos, de forma a tentar melhorar os resultados dos modelos obtidos até então. Porém ambas as alternativas não foram viáveis uma vez que pequenas alterações nos pesos produziam resultados muito distintos, muito possivelmente devido à dimensão relativamente pequena do conjunto de dados.

A última alternativa prendeu-se com a utilização de modelos de previsão, como redes neuronais e algoritmos como *random trees* ou *random forest* que foram avaliados através das mesmas métricas. Porém estes também não superaram os resultados obtidos e os padrões que encontraram resumem-se às variáveis valor por noite e localização. Uma possível razão para tal poderá ser um elevado sobreajuste aos dados que resulta da maior complexidade destes modelos, ou seja, os modelos aprendem padrões demasiado específicos no treino que não se generalizam bem para o teste.

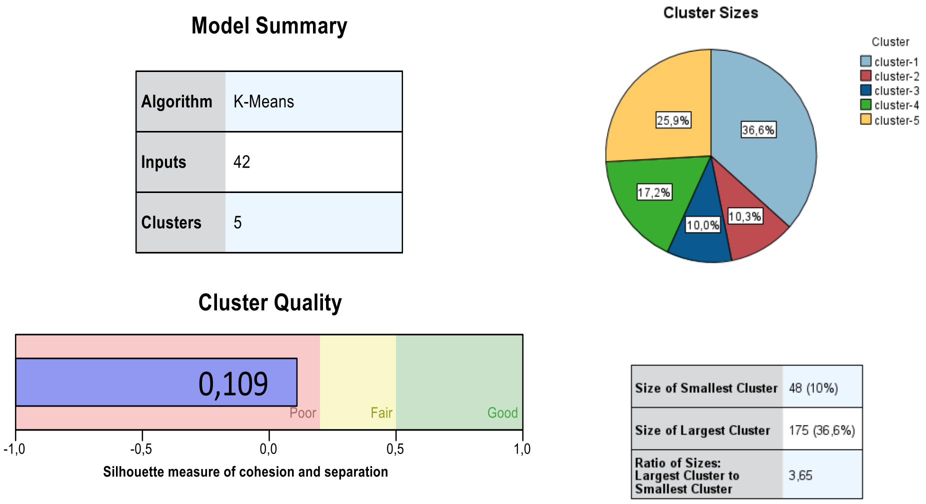
CAPíTULO 4: Resultados e Discussão

4.1. Caracterização dos alojamentos da FLH

Antes de responder pragmaticamente à principal questão de investigação, a qual se prende com a importância das características das imagens colocadas no anúncio de cada alojamento, faz-se uma análise geral das características dos alojamentos.

Esta análise visa encontrar as principais características dos alojamentos da FLH e perceber se existem padrões entre os diversos grupos de alojamentos e, no limite, se esses grupos podem ser diferenciados pela atratividade dos alojamentos. Procura-se assim, em traços gerais, as principais características e padrões dos alojamentos e se a atratividade é uma delas.

A Figura 22 mostra a segmentação do conjunto de dados em *clusters* (grupos ou segmentos). A modelação apresenta um baixo coeficiente de silhueta (0,109), pelo que não existem garantias que todos os grupos são completamente diferentes entre si. Pode também acontecer que um grupo contenha alojamentos com algumas características distintas. Porém, analisando os resultados do modelo produzidos pelo programa é possível verificar que os alojamentos são divididos em 5 *clusters*.



Model Summary

Algorithm	K-Means
Inputs	42
Clusters	5

Size of Smallest Cluster	48 (10%)
Size of Largest Cluster	175 (36,6%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	3,65

Figura 22. Resultado do modelo de segmentação.
Adaptado do *output* do IBM Modeler.

As variáveis com maior importância para diferenciar os alojamentos são apresentadas na Figura 23. Neste caso são: a presença de elevador, a divisão que se apresenta na primeira fotografia do anúncio, o número de imagens de sala de estar e a divisão que se apresenta nas segunda e terceira fotografias do anúncio.

Tendo em conta as cinco primeiras variáveis que mais diferenciam os alojamentos, quatro são relacionadas com as características das fotografias do site. Isto confirma que não existe um padrão na forma de decoração das casas da FLH, ao contrário do que poderia acontecer por exemplo em diferentes quartos de um mesmo hotel. Tal falta de padrão é facilmente explicável pelo facto de os alojamentos serem de particulares e geralmente manterem a sua decoração quando passam a ser geridos pela empresa.

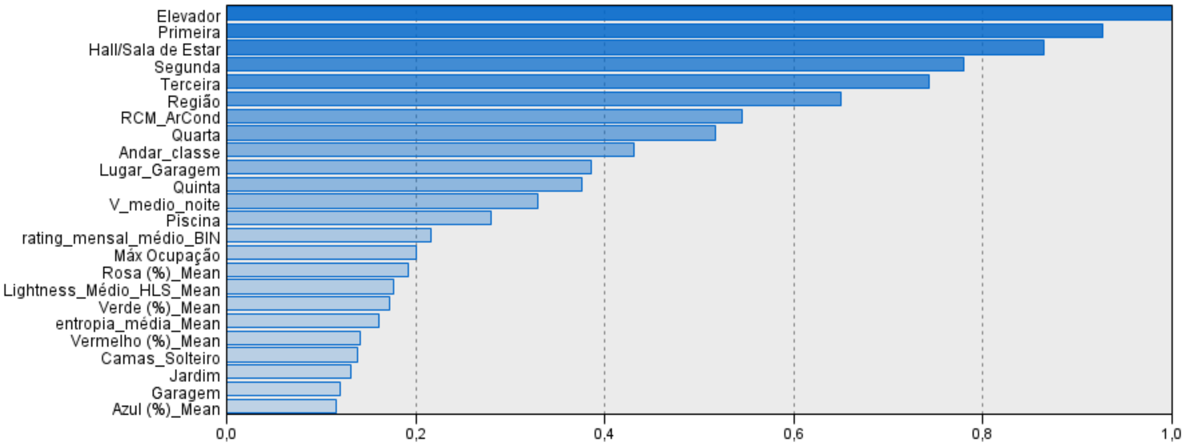


Figura 23. Importância das variáveis para a segmentação.
Adaptado do *output* do IBM Modeler.

A Figura 24 permite analisar a relação da atratividade e os vários grupos criados. Para a melhor interpretação, deve-se saber que na legenda está apresentado o alvo, que a vermelho (1) representa os alojamentos atrativos e a azul (0) os não atrativos. O gráfico à esquerda mostra a contagem de alojamentos de cada cluster e quantos são atrativos e não atrativos, enquanto o da direita mostra a mesma informação, mas em percentagem, ou seja, para cada cluster a percentagem dos seus alojamentos que são atrativos e não atrativos.

Observa-se que a atratividade não se consegue diferenciar de forma clara nos cinco grupos. Porém, pode-se observar que o *cluster* 4 se distingue por ter menos alojamentos atrativos (4%). Contrariamente, os *clusters* 1 e 5 são os que têm mais alojamentos atrativos, respetivamente 27% e 25% dos alojamentos pertencentes a estes grupos são atrativos.

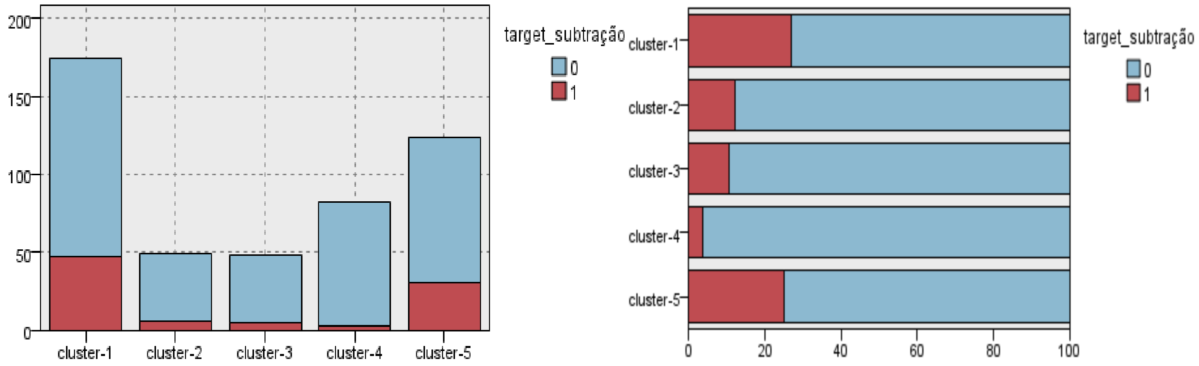


Figura 24. Relação entre atratividade e os clusters.
Adaptado do *output* do IBM Modeler.

A figura adaptada do programa, presente no Anexo E possibilita a leitura das características mais importantes e que melhor caracterizam cada *cluster*, dessa forma e com base nas 5 variáveis mais importantes é possível criar uma caracterização para cada um.

Caracterização dos clusters [2]:

1) Alojamentos sem elevador, cuja moda da primeira imagem é a divisão categorizada com 11 (outros), que têm poucas fotos do hall/sala de estar (apenas 1 a 2 fotos) e a moda da segunda e terceira imagem é a divisão 3 (quarto).
2) Alojamentos sem elevador, cuja moda da primeira imagem é a divisão 6 (piscina), que têm poucas fotos do hall/sala de estar (apenas 1 a 2 fotos), a moda da segunda imagem é a divisão categorizada com 11 (outros) e da terceira a divisão 3 (quarto).
3) Alojamentos sem elevador, cuja moda da primeira imagem é a divisão 12 (Hall/Sala de Estar), que têm algumas fotos do hall/sala de estar (mais do que 3 fotos) e a moda da segunda e terceira imagem é a divisão 12 (Hall/Sala de Estar).
4) Alojamentos com elevador, cuja moda da primeira imagem é a divisão 12 (Hall/Sala de Estar), que têm algumas fotos do hall/sala de estar (mais do que 3 fotos) e a moda da segunda e terceira imagem é a divisão 12 (Hall/Sala de Estar).
5) Alojamentos com elevador, cuja moda da primeira imagem é a divisão categorizada com 11 (outros), que têm poucas fotos do hall/sala de estar (1 a 2 fotos) e a moda da segunda e terceira imagem é a divisão 3 (quarto).

Analisando a caracterização dos clusters com mais alojamentos atrativos, através dos gráficos de comparação de clusters presente no Anexo F, é possível perceber que os clusters 1 e 5 são muito semelhantes, dado que apresentam padrões semelhantes, exceto em relação ao elevador, ao andar da casa e ao ar condicionado. O cluster 1 não tem elevador e a moda é ter alojamentos em andares mais baixos, já o 5 tem elevador e alojamentos em andares mais altos. Estas duas características são complementares, no sentido que, os alojamentos em andares mais baixos não necessitam de elevador, mas nos que se localizam em andares mais altos o elevador é bastante importante. Esta complementaridade justifica o baixo impacto na atratividade. Já a presença ou não de ar condicionado pode influenciar a atratividade, neste caso o *cluster* 5 apresenta maioritariamente alojamentos com ar condicionado. Esta é a razão pela qual não se sugere a junção de ambos os grupos.

Analisando os outros *clusters* em termos de atratividade, o cluster 2 é o mais semelhante com o 1 e 5 e de facto é o terceiro em termos de percentagem de alojamentos atrativos incluídos no cluster. Depois fica o 3 e por último o 4, mais semelhantes entre si e diferentes dos primeiros, mas ainda assim existem certas características que talvez não fossem as esperadas, como a frequente presença de ar condicionado nos alojamentos do cluster 4.

Na próxima secção, procura-se clarificar as relações entre as características dos alojamentos e a sua atratividade, se as características das imagens também são importantes para explicar a atratividade dos alojamentos e quais são as características dos alojamentos atrativos e dos não atrativos.

---

[2] A categoria 11 - outros pode ser ambígua, pois referem-se a fotográficas que podem apanhar mais do que uma parte da propriedade, como explicado mais à frente

4.2. A atratividade dos alojamentos da FLH

Dado que se pretende analisar características das imagens e dos alojamentos, antes de iniciar a modelação foram criadas algumas visualizações que permitem fazer uma análise bi-variada entre algumas dessas variáveis e a atratividade. Para esta análise foram escolhidas as variáveis que se consideram mais importantes, tanto do ponto de vista do negócio como estatisticamente, tais como o preço por noite e a região da casa. Foram também consideradas algumas variáveis extraídas das imagens, que segundo as métricas de correlação têm uma maior relação com o alvo.

De referir que as análises seguintes são apenas entre duas variáveis pelo que é de esperar que não expliquem de forma clara e inequívoca o porquê de um alojamento ser ou não atrativo. Na sua maioria um alojamento é ou não considerado atrativo devido a um vasto conjunto de variáveis e perceções do cliente (Lien et al., 2015).

4.2.1. Relação entre atratividade e as outras características

Tendo em conta a relação entre a região e a atratividade (Figura 25) é possível perceber que as regiões que têm os alojamentos mais atrativos são, por ordem decrescente, Lisboa, Porto, Algarve e Madeira. Esta relação pode indiciar que os alojamentos de Lisboa ou Porto são mais atrativos. No entanto, não tem em conta outras variáveis importantes como a relação entre a oferta e a procura da região, a qual pode também ter um grande impacto na atratividade. Por outro lado, não existem alojamentos com alvo positivo no Alentejo, Cascais/Estoril, Centro e Ericeira, o que pode ser explicado pelo facto de serem zonas com poucos alojamentos. A amostra nestas zonas é muito mais pequena do que quando comparado com as outras, em conjunto, as quatro regiões têm 38 alojamentos, de uma amostra total de 487. Uma vez que em todo o conjunto de dados existem muito mais alojamentos considerados não atrativos, é expectável que numa amostra maior existam mais alojamentos atrativos do que numa amostra pequena.

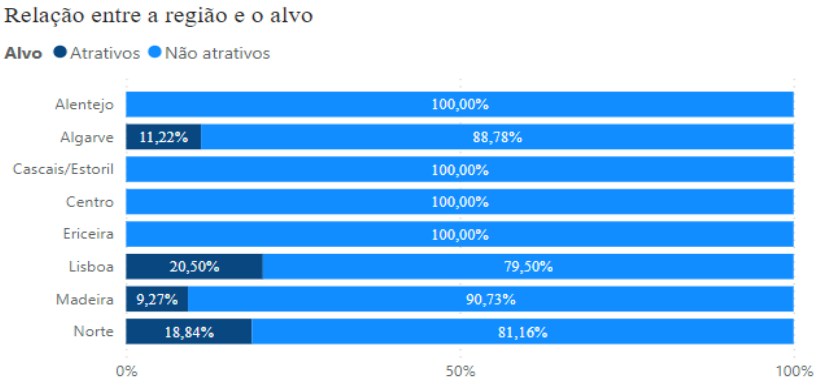


Figura 25. Relação entre a região e o alvo.
Adaptado do output do Power BI.

A Figura 26 mostra que existem mais alojamentos atrativos sem elevador e ar condicionado (respetivamente, 18% e 24%) do que com (respetivamente, 12% e 12%). É algo que à partida não seria de esperar, pois são duas comodidades que trazem conforto e que por isso tornariam a casa mais atrativa.

É aceitável os alojamentos mais atrativos serem, em maior percentagem, os que não têm elevador ou ar condicionado, pois é muito provável que quando os possíveis hospedes procuram um alojamento para alugar não sejam essas as principais características que influenciam a sua decisão. Isto pode ser explicado pelo facto de nesta amostra existirem alojamentos dos bairros antigos de lisboa, por exemplo, que como são mais pequenos não têm espaço para elevador, ou então são moradias ou alojamentos no rés do chão que não têm necessidade de elevador. Para além disso, tratam-se por norma de estadias de curta duração e com finalidade turística. Nestes casos, a proximidade a certos sítios pode ser mais importante do que ter ar condicionado, visto que o tempo passado dentro do alojamento é reduzido.

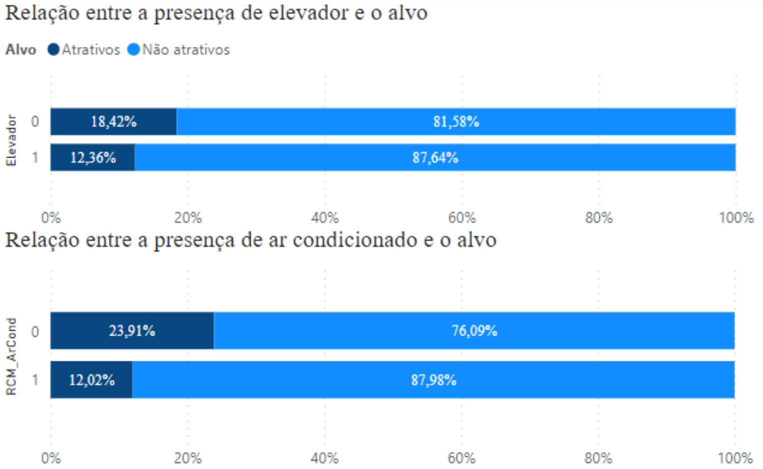


Figura 26. Relação entre presença de elevador e ar condicionado e o alvo.
Adaptado do output do Power BI.

Ainda sobre características do anúncio, e em conformidade com o esperado, a Figura 27 mostra que alojamentos mais atrativos têm um valor médio por noite mais baixo (49,72€) do que os alojamentos menos atrativos (72,09€), o que faz todo o sentido, para situações similares qualquer consumidor prefere pagar o menos possível.

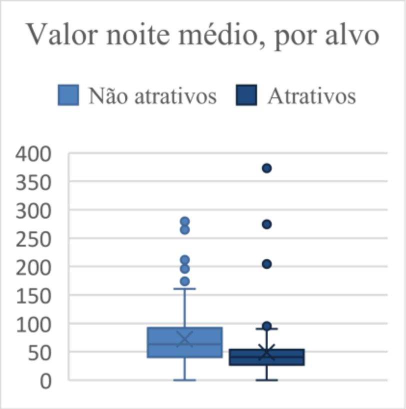


Figura 27. Valor por noite médio, por alvo.
Adaptado do output do Excel.

Tendo em conta as fotografias dos alojamentos, a Figura 28 mostra que, existe uma pequena diferença entre o número de fotografias dos anúncios, os alojamentos mais atrativos têm, em média, menos fotografias no anúncio do que os menos atrativos (respetivamente, 18,8 e 19,7). Este facto está de acordo com o descrito no artigo de Larceneux *et al*. (2018), onde se afirma que, quanto maior for o número de fotografias, mais específica é a qualidade do produto, por isso, se o alojamento não for de excelência, um maior número de fotografias pode mostrar todos os detalhes que comprovem essa falta de excelência logo, diminuir a atratividade.

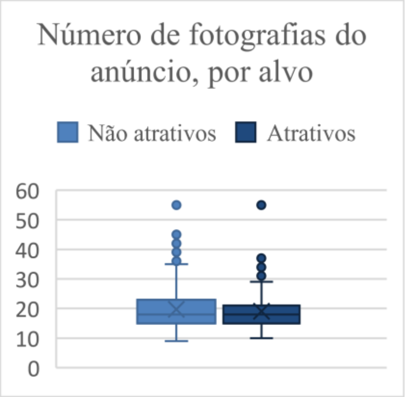


Figura 28. Número de fotografias do anúncio, por alvo.
Adaptado do output do Excel.

Analisando as características das imagens tanto para os alojamentos atrativos como para os não atrativos (Figuras 29 e 30) verificam-se diferenças muito baixas entre as duas classes de atratividade. As casas mais atrativas têm em média uma menor percentagem de amarelos, laranjas, rosas, verdes e vermelhos, sendo que apenas a cor azul existe em maior percentagem para os alojamentos atrativos. Também as casas mais atrativas têm uma menor entropia média e uma maior luminosidade nas suas fotografias. Isto significa que não têm tantas mudanças de profundidade, objetos e cenários diferentes e também que são, em média, imagens mais claras e com menos ruído.

Porém, as diferenças presentes nos gráficos da Figuras 29 e 30, para além de pequenas, muito possivelmente não são importantes individualmente. Quando analisadas todas em conjunto podem ter outra relação com o alvo, porque uma imagem é composta por uma combinação de todas as suas caraterísticas.

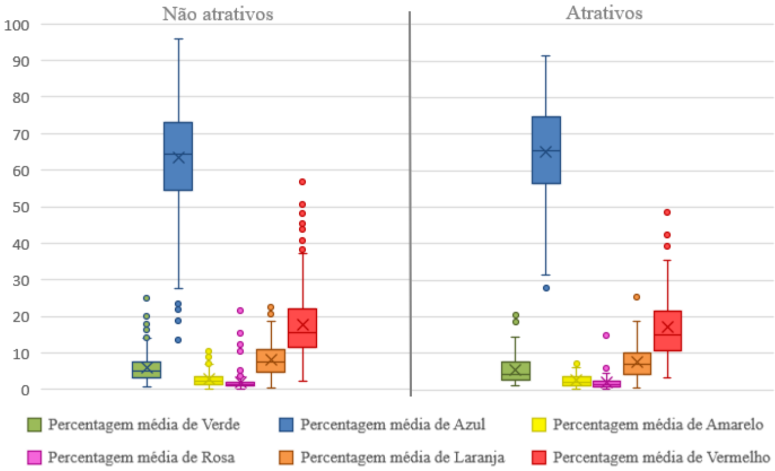


Figura 29. Percentagem média das cores presentes nas fotografias, por alvo.
Fonte: Elaboração própria, no Excel.

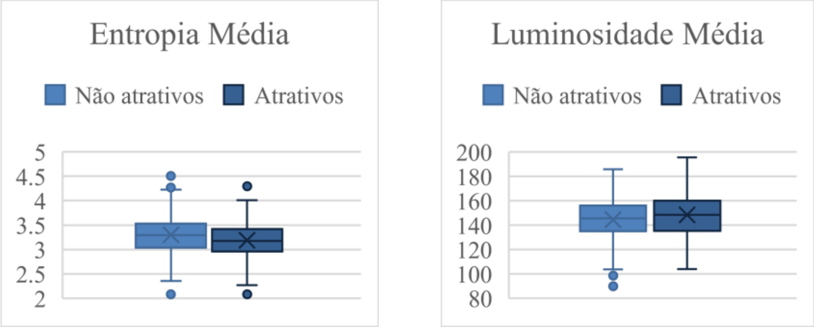


Figura 30. Entropia e luminosidade médias.
Fonte: Elaboração própria, no Excel.

Ainda relativamente às fotografias, nomeadamente às divisões fotografadas e à sua ordem no anúncio, a Figura 31 é possível ver para as cinco primeiras fotografias quais são as divisões que pertencem a uma maior percentagem de alojamentos atrativos.

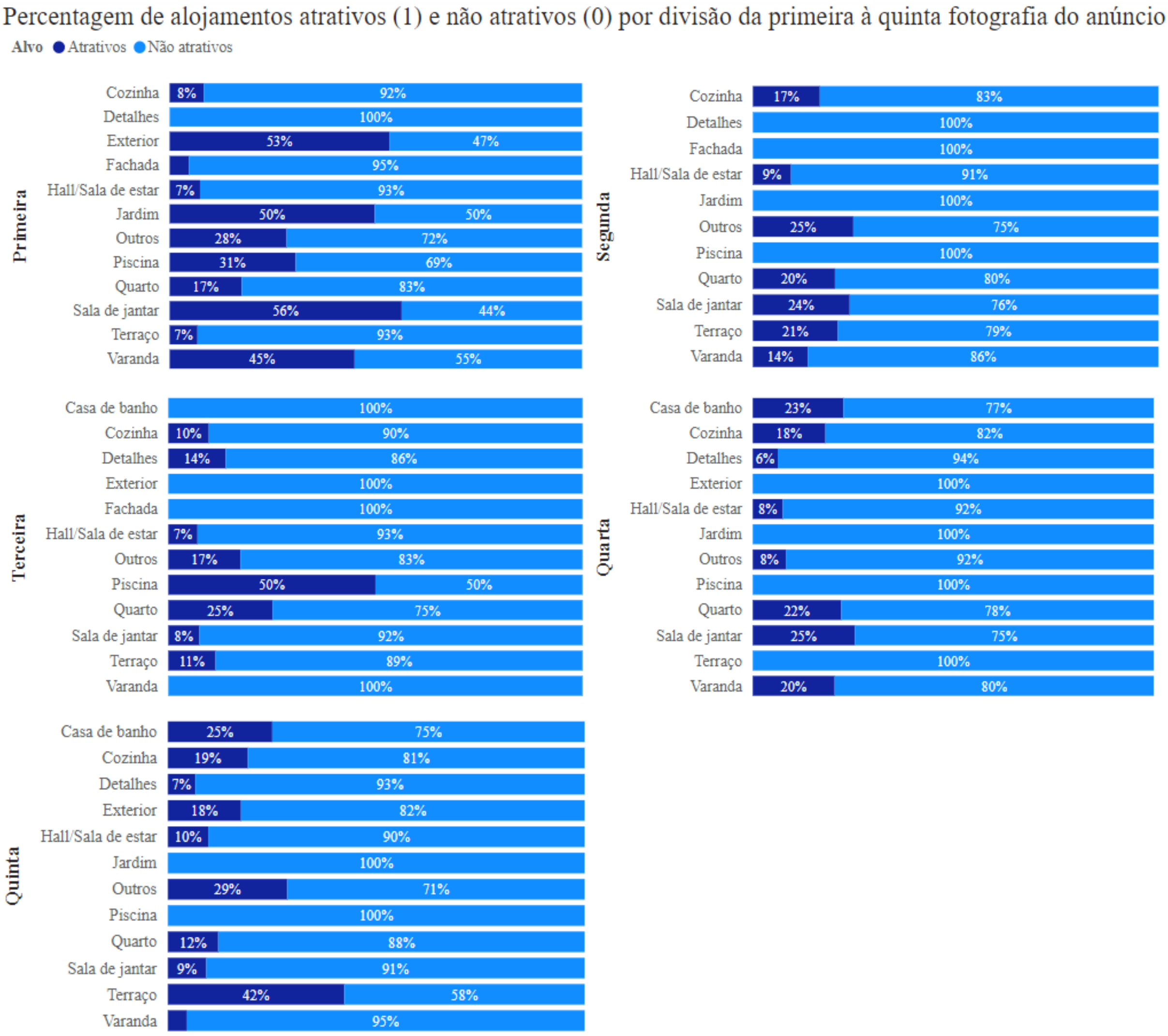


Figura 31. Relação entre a divisão, da primeira à quinta fotografia do anúncio, e o alvo.
Adaptado do output do Power BI.

Para a primeira fotografia existe uma maior percentagem de alojamentos atrativos com imagens da sala de jantar, do exterior, do jardim e da varanda, dando a ideia de que imagens mais gerais e do ambiente do alojamento poderão funcionar melhor como primeiras fotografias. Para a segunda fotografia já não se encontra nenhum padrão específico, observa-se uma grande semelhança nas divisões com mais alojamentos atrativos: a sala de jantar, outros, terraço e quarto, logo devido a esta distribuição não existe uma divisão que salte à vista como a mais importante para o segundo lugar. Para a fotografia que aparece em terceiro lugar, as divisões que apresentam uma maior percentagem de alojamentos atrativos são a piscina e o quarto. Isto leva a crer que depois de obter uma ideia do ambiente geral da casa, o consumidor pretende saber mais detalhes da mesma. Para a quarta fotografia não há uma divisão que se destaque por ter uma elevada percentagem de alojamentos atrativos, não se conseguindo retirar um padrão fixo. Já para a quinta fotografia, voltam a estar em destaque as imagens do terraço.

Algo que é transversal a quase todas as posições das fotografias é a presença considerável de alojamentos atrativos quando a divisão apresentada é outros. Como podemos ver nas imagens apresentadas na Figura 32, isto pode confundir o modelo, dado que as imagens desta classe apresentam várias divisões em simultâneo.

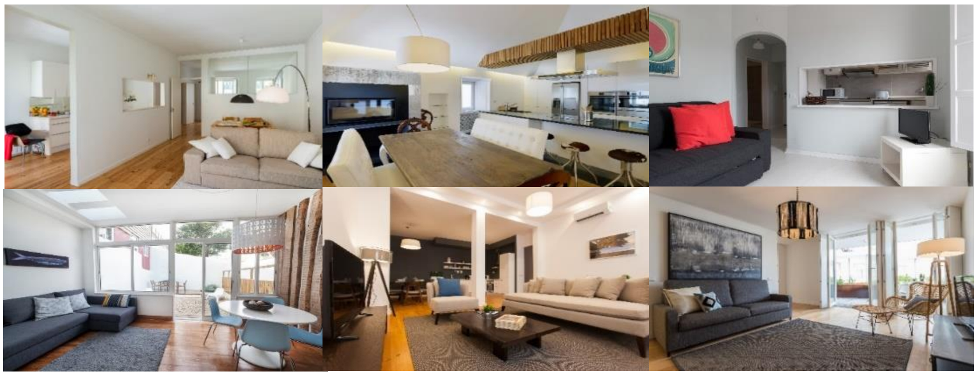


Figura 32. Exemplos de divisões categorizadas como Outros.
Fonte: Base de dados FLH.

4.2.2. Perfis da atratividade dos alojamentos

4.2.2.1. Avaliação dos modelos

Dado que a atratividade de um alojamento pode assentar num variado conjunto de variáveis, a próxima análise procura encontrar padrões que permitam explicar o porquê de uma casa ser atrativa ou não. Para perceber se as informações retiradas das imagens são importantes para explicar a atratividade, utilizaram-se diferentes conjuntos de variáveis e vários modelos com diferentes parâmetros, como foi explicado anteriormente na metodologia na secção 0. Neste capítulo é apenas apresentado o melhor modelo para cada conjunto de dados, apresentados na Tabela 8 da secção 3.3.4.

É importante ter a ideia de que, para se chegar a estes modelos finais, foi realizado um enorme conjunto de experiências para cada um dos quatro conjuntos de variáveis. Existiram testes com as variáveis contínuas e também com a categórica criada, com o objetivo de ficar com aquelas que apresentassem melhor desempenho (visto que ambas poderiam ser modeladas nas árvores de decisão). Foram testados, modelos com parametrização *default* para cada conjunto de variáveis, sendo então cada um desses modelos analisado e os seus parâmetros alterados, com o intuito de o tornar mais ou menos complexo, no caso de se apresentar, respetivamente, sub ou sobre ajustado. Todos os modelos iam sendo comparados com o melhor de cada conjunto até então, e eliminados no caso dos resultados serem inferior ou substituindo o melhor, no caso de serem realmente melhores. Para além destas tentativas existiram também as que estão apresentadas na secção 3.7. que acabaram por não ter resultados favoráveis.

A Tabela 10 apresenta os melhores resultados para cada conjunto de dados, sendo que o melhor modelo foi em todos os casos o C5.0. Ao observar os resultados percebe-se que os modelos conseguem aprender bastante bem os padrões existentes no treino, mas depois não conseguem aplicar da melhor forma esses padrões no teste, ou seja os modelos estão bastante sobreajustados.

Tabela 10. Resultados dos modelos para cada conjunto de dados.

Parâmetros \ Modelos	**Caract. das casas e de todas as imagens**	**Caract. das casas e da primeira imagem**	**Caract. das casas e das 5 primeiras imagens**	**Só caract. das casas**
Profundidade Máxima	11	9	6	12
Mínimo de casos: nós Filho	2	2	10	10
Modo	*Simple*	*Simple*	*Expert*	*Expert*
Nós	81	78	33	21
Nós terminais	48	46	21	13
Treino				
Exatidão	97%	97%	85%	80%
Sensibilidade	100%	100%	89%	81%
Especificidade	94%	94%	82%	80%
Precisão	95%	94%	82%	80%
Medida-F	97%	97%	86%	80%
AUC	0,98	0,98	0,91	0,85
Teste				
Exatidão	72%	78%	71%	68%
Sensibilidade	43%	50%	57%	50%
Especificidade	80%	85%	75%	72%
Precisão	35%	45%	36%	31%
Medida-F	39%	48%	44%	38%
AUC	0,63	0,66	0,69	0,67
Valid. Cruzada: Média	-	-	75,2	75,3
Valid. Cruzada: Erro padrão	-	-	1,9	1,8

Nota: Os valores da validação cruzada não existem para todos os modelos, uma vez que apenas são apresentados pelo programa quando se escolhe o modo *expert*

Para combater o sobreajustamento foram tomados vários caminhos. Tentou-se reduzir a complexidade do modelo, impossibilitando a aprendizagem de tantos detalhes no treino, com o propósito de conseguir acertar mais no teste, através de diferentes parametrizações, diferentes custos de má classificação, o aumento do número de filhos dos nós e/ou redução da profundidade máxima para não permitir que o modelo se torne tão complexo. Contudo não se conseguiu combater esse sobreajustamento, quando se impossibilitava a aprendizagem o modelo tinha piores resultados tanto no treino como no teste. Este sobreajustamento poder-se-á dever principalmente ao facto da atratividade de um alojamento depender de um elevado número de características, fazendo com que o reduzido número de amostras não seja suficiente para determinar padrões.

Através da métrica especificidade é possível perceber que, em todos os conjuntos de dados, os alojamentos menos atrativos são mais facilmente reconhecidos. Isto pode dever-se a dois factos: no conjunto de teste existe maior número de alojamentos não atrativos do que atrativos; ou à facilidade em identificar algumas características que claramente não agradam aos hospedes, como por exemplo um valor por noite elevado, uma má localização ou maus acessos. Já a dificuldade em definir o porquê de um alojamento ser considerado atrativo (transparecida pela sensibilidade e pela Medida-F), no seguimento do que se já referiu, pode estar ligada ao facto de haver poucos alojamentos atrativos no teste. Comparando os resultados nos quatro conjuntos de dados e focando mais precisamente na medida F, todos os modelos com informações de imagens apresentam melhores resultados no teste face ao conjunto que contém apenas informação da casa.

4.2.2.2. Importância das variáveis

Sendo que o modelo com informações da casa e da primeira imagem do anúncio é considerado o melhor, este será utilizado para comparação com o modelo que contém apenas informação da casa. Esta análise ajuda a entender melhor o que acontece quando se introduz informação das imagens na modelação.

Na Figura 33 e na Figura 34 apresentam-se os gráficos da importância das variáveis do modelo apenas com as características das casas e do modelo que também inclui a informação da primeira imagem do anúncio, respetivamente.

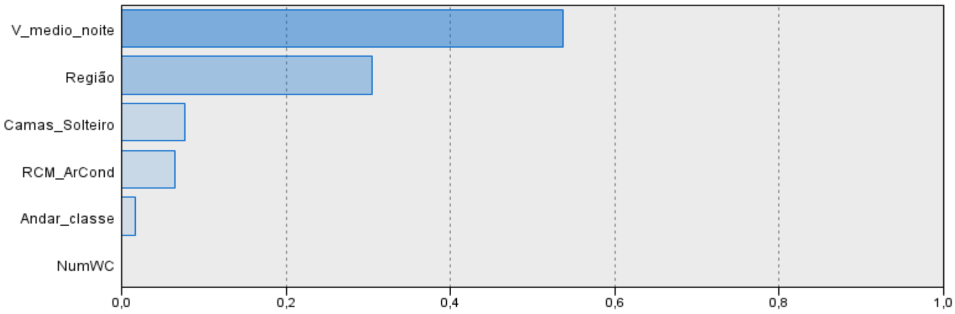


Figura 33. Gráfico de importância – modelo apenas com caraterísticas das casas. Adaptado do output do IBM Modeler.

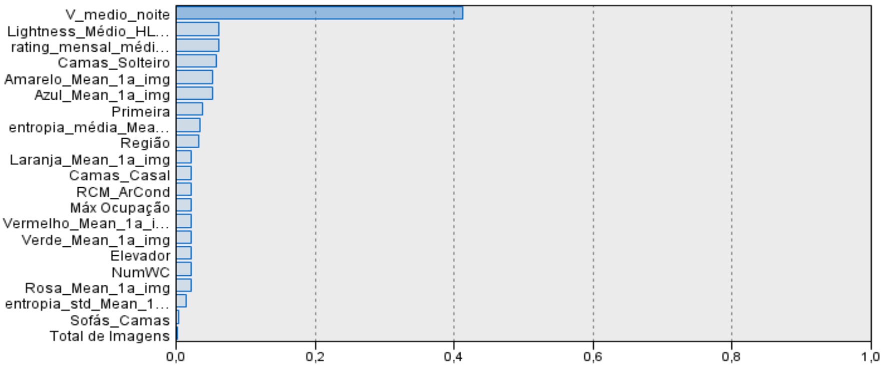


Figura 34. Gráfico de importância – modelo com caraterísticas das casas e da 1ª imagem. Adaptado do output do IBM Modeler.

É possível de observar que, em ambos os modelos, a variável mais importante para classificar a casa como atrativa é o seu valor por noite, sendo inegável a importância desta variável. Pode-se também constatar que as características das imagens ocupam várias posições de importância no modelo, o que vem corroborar as conclusões retiradas por Kostic & Jevremovic em 2020. Neste caso, na Figura 33 a região é a segunda variável mais importante, mas quando se juntam as informações das imagens à equação (Figura 34) esta já passa para a nona posição, isto mostra que os padrões explicativos da atratividade são bastante diferentes nos dois modelos. Os valores de importância de valor por noite e região diminuem, respetivamente, de aproximadamente 0,55 para 0,4 e de 0,3 para 0,05 quando as características das imagens são incluídas, dando importância a estas novas variáveis.

Nas Figuras 35 e 36 estão representados os gráficos de ganhos para os dois modelos referidos anteriormente.

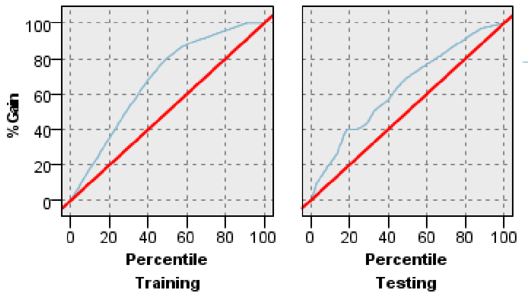


Figura 35. Gráfico de ganhos do modelo apenas com caraterísticas das casas. Adaptado do output do IBM Modeler.

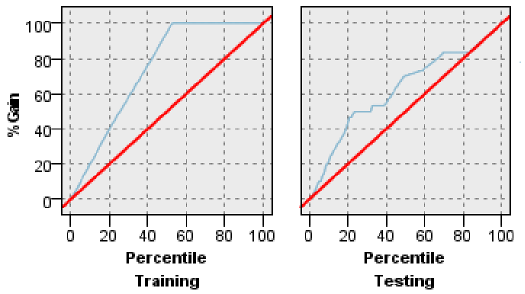


Figura 36. Gráfico de ganhos do modelo com caraterísticas das casas e da 1ª imagem. Adaptado do output do IBM Modeler.

O gráfico de ganhos representa a percentagem de casos positivos, isto é, a percentagem de alojamentos que realmente são atrativos, em função do percentil de casos mais prováveis de serem atrativos, ou seja, no eixo das abcissas os casos são ordenados por ordem decrescente da probabilidade de atratividade. Na Figura 36, considerando a partição do teste, para os 20% dos alojamentos com maior probabilidade de serem atrativos, acerta-se em cerca de 50%, ou seja, consegue-se chegar a 50% dos alojamentos que são realmente atrativos. Já na Figura 35 constata-se que para os mesmos 20% dos alojamentos com maior probabilidade de serem atrativos só se atinge cerca de 40% dos atrativos.

Através da interpretação dos gráficos nota-se que, recorrendo a um menor número de alojamentos, o modelo com imagem tem uma maior taxa de acerto do que o modelo com apenas informação da casa. Isto realça novamente a importância de incorporar as imagens e as suas características neste tipo de estudos.

4.2.2.3. Regras do modelo

Com base nas regras das árvores de decisão construídas pelos dois modelos em análise, visíveis no Anexo G para o modelo apenas com informação da casa e no Anexo H para o modelo com informação da casa e da primeira imagem do anúncio, foram criados os perfis de alojamentos atrativos. Foram selecionadas as regras com maior confiança e maior número de registos de suporte. Estes perfis permitem mais claramente perceber quais as variáveis importantes e de que forma contribuem ou não para a atratividade do alojamento.

Modelo apenas com informação da casa

Perfis de alojamentos atrativos:

- Com 90,9 % de confiança e suporte de 22 alojamentos, os alojamentos com um valor por noite inferior ou igual a 55,37€, na zona de lisboa, sem camas de solteiro e com no mínimo uma casa de banho, tendem a ser atrativos.
- Com 90,9 % de confiança e suporte de 11 alojamentos, os alojamentos com um valor por noite superior a 269,50€ tendem a ser atrativos.

Perfis de alojamentos não atrativos:

- Com 81% de confiança e suporte de 21 alojamentos, os alojamentos com um valor por noite inferior ou igual a 55,37€, na Ericeira, Cascais-Estoril ou Algarve, tendem a ser não atrativos.
- Com 100% de confiança e suporte de 46 alojamentos, os alojamentos com um valor por noite entre 108,50€ e 269,50€ tendem a ser não atrativos.

Através do modelo com base apenas nas características das casas pode-se reter que os alojamentos atrativos estão na sua maioria em Lisboa e, ou têm valores por noite baixos (<55,37€), ou então muito altos (>269,50€). Isto pode significar que existem dois tipos de hóspedes, os que preferem valores baixos e por isso não é necessário ter muitas comodidades e apenas as necessidades básicas, e os que preferem opções mais luxuosas e com preços mais elevados, tornando-se o alojamento atrativo por existir uma qualidade de excelência garantida. Por outro lado, os alojamentos menos atrativos encontram-se na Ericeira, em Cascais-Estoril ou no Algarve, ou então têm um preço mediano (>55,37€ e <269,50€), que muito possivelmente não serve a nenhum dos tipos de hóspedes que foram referidos anteriormente.

Modelo com informação da casa e da primeira imagem do anúncio

Perfis de alojamentos atrativos:

- Com 100 % de confiança e suporte de 30 alojamentos, os alojamentos com um valor por noite inferior ou igual a 55,37€ e uma luminosidade média da primeira imagem do anúncio superior a 178,93, tendem a ser atrativos.
- Com 96,7 % de confiança e suporte de 30 alojamentos, os alojamentos com um valor por noite inferior ou igual a 55,37€, uma luminosidade média da primeira imagem do anúncio entre 103,86 e 178,93, na região Norte, com uma percentagem média de azul e rosa na primeira imagem do anúncio, inferior ou igual a 74,65% e 9,38%, respetivamente e nenhum sofá cama, tendem a ser atrativos.
- Com 94,1 % de confiança e suporte de 17 alojamentos, os alojamentos com um valor por noite inferior ou igual a 55,37€, uma luminosidade média da primeira imagem do anúncio inferior ou igual a 178,93, na região de Lisboa, sem camas de solteiro, com uma entropia média na primeira imagem do anúncio superior a 2,88, com 1 casa de banho, com o quarto na primeira imagem,

sem elevador e com uma percentagem média de verde na primeira imagem do anúncio superior a 1,52%, tendem a ser atrativos.

Perfis de alojamentos não atrativos:

- Com 100 % de confiança e suporte de 58 alojamentos, os alojamentos com um valor por noite superior a 55,37€, no máximo 5 camas de solteiro e uma entropia média na primeira imagem do anúncio superior a 1,13, tendem a ser não atrativos.
- Com 100 % de confiança e suporte de 28 alojamentos, os alojamentos com um valor por noite superior a 55,37€, no máximo 5 camas de solteiro, uma entropia média na primeira imagem do anúncio inferior ou igual a 1,13, a primeira imagem categorizada como "outros", com uma luminosidade média da primeira imagem do anúncio inferior ou igual a 167,88 e uma percentagem média de rosa na primeira imagem do anúncio superior a 0,08%, tendem a ser não atrativos.
- Com 100 % de confiança e suporte de 45 alojamentos, os alojamentos com um valor por noite superior a 55,37€, no máximo 5 camas de solteiro, uma entropia média na primeira imagem do anúncio superior a 2,66, a primeira imagem classificada como "hall/sala de estar", com uma percentagem média de azul na primeira imagem do anúncio superior a 21,82, tendem a ser não atrativos.

Um baixo valor por noite (<55,37€) e pertença à região lisboa continuam a ser dois fatores que diferenciam alojamentos atrativos. Adicionalmente retém-se que na região do Norte, alojamentos sem camas de solteiro, sofás-cama e elevador tendem a ser atrativos. Isto pode ser explicado pelo facto de em Lisboa e no Norte existirem alojamentos bastantes antigos, sendo aceitável pelos hóspedes que não haja elevador à disposição.

Os alojamentos não atrativos apresentam um maior número de camas de solteiro e já não existe a menção das regiões Ericeira, Cascais-Estoril ou Algarve, as razões pelo qual isto acontece não é claro, o que se percebe é que quando comparado com a análise bi-variada e o modelo anterior, que mostravam que estas três regiões apresentavam uma elevada percentagem de alojamentos não atrativos, conclui-se agora que a região pode não ser uma causa da não atratividade. Muito possivelmente existe um denominador comum nos alojamentos destas regiões que torne as casas menos atrativas, sugerindo uma necessidade de realizar uma análise semelhante, mas apenas para os alojamentos desta zona.

Mais ainda, em relação à primeira imagem, são atrativos alojamentos que tenham a fotografia do quarto. Em contrapartida, alojamentos com a primeira imagem da categoria "sala" ou "outros" tendem a ser menos atrativos. Já a luminosidade e a entropia da primeira fotografia não apresentam regras muito específicas, apenas se pode apreender que luminosidade e entropia muito baixas são característica dos alojamentos menos atrativos. Por último e em relação à cor da primeira fotografia, salienta-se que o verde, muito provavelmente ligado à natureza, tal como Kostic & Jevremovic (2020) afirmaram, é uma característica dos alojamentos mais atrativos. Por outro lado, o rosa é uma característica dos menos atrativos.

CAPÍTULO 5: Conclusões e Recomendações

5.1. Conclusão

Esta dissertação procurou responder às duas questões de investigação elaboradas para otimizar a estratégia de escolha e disposição das fotografias nos anúncios da Feels Like Home (FLH), uma empresa de gestão de propriedades para aluguer de alojamento local. A análise foi realizada seguindo a metodologia CRISP-DM e utilizando os dados fornecidos pela FLH.

A revisão sistemática da literatura (RSL) permitiu perceber quais os tópicos/conteúdos menos abordados. Com base nos critérios de seleção utilizados, 23 artigos foram selecionados a partir de um universo inicial de 241, indiciando que literatura sobre esta temática apresenta lacunas que esta dissertação visou preencher.

A maioria dos estudos analisados concentram-se em anúncios de aluguer de alojamentos turísticos, imóveis e imagens turísticas, sejam de atrações ou publicadas em redes sociais, sendo que do conjunto de artigos selecionados, 15 analisam imagens, porém apenas um deles é sobre alojamento turístico.

Adicionalmente, a partir da revisão realizada, pode-se concluir que estudos de outras áreas podem ser utilizados como referência para este trabalho, visto que há pontos em comum entre o alojamento turístico e as imagens de imóveis e atrações turísticas como forma de análise de imóveis e/ou preferências turísticas.

Os objetivos analíticos definidos inicialmente prenderam-se numa primeira fase com a extração das características visuais das imagens dos alojamentos. Numa segunda fase, passaram por perceber se existia algum padrão ou forma de agrupar os alojamentos em função das suas características. Numa última fase, procurou-se separar os alojamentos em atrativos e não atrativos e perceber quais são as características mais importantes para modelar a atratividade.

Na metodologia foram preparados os dados, cumprindo-se dessa forma o primeiro objetivo analítico, relativo à extração de características visuais das imagens dos alojamentos da FLH e sua apresentação numa tabela. Verificou-se também se as características relacionadas com as imagens podem ser consideradas importantes, primeiro de uma formais mais simples, através de análises bi-variadas, e depois de uma forma mais completa, através da modelação com árvores de decisão.

Com o objetivo de identificar características dos alojamentos da FLH que permitam definir padrões de segmentação dos alojamentos e perceber se esses segmentos podem ser diferenciados pela atratividade, foi realizada uma segmentação. A análise do coeficiente de silhueta mostrou que a criação dos segmentos dos alojamentos não tem resultados excelentes, visto que tem um valor de 0,109, o que significa que a classificação é fraca. Ainda assim, analisando os grupos foi possível retirar algumas conclusões.

As variáveis com mais importância para segmentar os alojamentos são a presença de elevador e características das imagens no site, nomeadamente a divisão da primeira fotografia, o número de imagens de sala de estar e a divisão da segunda e terceira fotografias. Isto mostra que, para além de presença de elevador, são as características visuais que mais diferenciam os alojamentos da FLH, o que é esperado visto que estes são de particulares e o seu visual não é definido pela FLH.

Adicionalmente, analisando os 5 segmentos de alojamentos criados tendo em conta o alvo criado com base nos cliques, na tentativa de representar a atratividade dos mesmos, percebe-se que o modelo não os consegue distinguir de uma forma inequívoca. Ainda assim, sobressaem algumas características que os distinguem.

Nos dois segmentos com maior percentagem de alojamentos atrativos, a primeira imagem é maioritariamente da categoria "outros" (categoria da imagem que mostra mais do que uma divisão da casa ou outra coisa que não seja uma divisão da casa), têm em média poucas fotos do hall/sala de estar e a segunda e terceira divisão são maioritariamente do quarto. Já para o segmento com menor percentagem de alojamentos atrativos a primeira, segunda e terceira fotografia são maioritariamente do hall/sala de estar.

Com o objetivo de entender o comportamento entre as características das casas e a sua atratividade, foram criadas e analisadas relações bi-variadas. Observou-se que as regiões que têm uma maior presença de alojamentos atrativos são Lisboa e Porto, que a maioria não têm elevador e ar condicionado e que há um maior número de alojamentos atrativos com um valor médio por noite mais baixo e menor número de fotografias no anúncio.

Focando ainda na análise bi-variada, tendo em conta as divisões do alojamento que aparecem e a sua ordem no anúncio, é percetível que para a primeira fotografia existe uma maior percentagem de alojamentos atrativos com imagens da sala de jantar, do exterior, do jardim e da varanda, dando a ideia de que imagens mais gerais e do ambiente do alojamento poderão funcionar melhor como primeiras fotografias. Algo que é transversal a quase todas as posições das fotografias no anúncio, é o facto de os alojamentos com a foto do tipo "outros" terem tendência a ser mais atrativos.

A última análise realizada procurou encontrar os padrões que permitam explicar o porquê de uma casa ser atrativa ou não. Para tal, foram utilizadas quatro técnicas de modelação com arvores de decisão C5.0, CART, QUEST e CHAID, sendo que os melhores modelos resultaram da utilização do algoritmo C5.0. O modelo com melhores métricas apresentou na amostra de teste 78% de exatidão, 50% de sensibilidade, 85% de especificidade, 48% na medida F e 0,66 de AUC. Algo que deve ser tido em conta, é que todos os modelos, mesmo depois de todos os esforços para o combater, apresentaram sobreajustamento.

O modelo escolhido responde afirmativamente à primeira questão de investigação "Qual será a importância relativa do conteúdo visual na atratividade dos alojamentos da FLH, quando comparadas com as restantes características dos mesmos?". Apesar da variável mais importante para classificar a casa como atrativa ser claramente o seu valor por noite, quando consideradas as características das imagens na análise, estas ocupam posições de elevada importância no modelo, até mesmo antes de outras características da casa.

Por último, foi possível construir alguns perfis que explicam as características dos alojamentos mais e menos atrativos permitindo responder à segunda questão de investigação “Quais as características visuais das imagens dos alojamentos da FLH que mais atraem os potenciais clientes?”. Concluiu-se que alojamentos atrativos tendem a ter um valor por noite muito alto ou então um valor por noite baixo, tendem a estar localizados em Lisboa e a ter como primeira imagem fotografias do quarto e fotografias com uma maior percentagem de verde, ou então tendem a ser da região do Norte e não ter camas de solteiro, sofás-cama nem elevador.

Por outro lado, os alojamentos não atrativos apresentam um maior número de camas de solteiro e como primeira imagem a sala ou outros e também valores de luminosidade e entropia muito baixas. Por último, em relação à cor da primeira fotografia, salienta-se que nos alojamentos menos atrativos existe uma maior presença da cor rosa.

5.2. Limitações

Pode referir-se que as limitações sentidas durante a investigação estão principalmente associadas a cinco fatores.

1. Ao facto de existirem variáveis que podem influenciar a atratividade de um alojamento, mas que não estão disponíveis, como os dados demográficos do hóspede ou as características da zona onde se situa o alojamento,
2. Uma vez que são dados apenas relativos a 2020 e 2021, as conclusões podem não ser generalizáveis por se tratar de anos atípicos em termos de turismo devido à influência da pandemia,
3. Os anúncios são controlados pelas plataformas, pelo que estas podem aplicar técnicas para destacar os alojamentos consoante as suas próprias regras e interesses. Consequentemente, fatores desconhecidos podem influenciar o clique e a atratividade de um alojamento,
4. A forma de categorização da divisão da imagem, mais precisamente das imagens categorizadas como 11 - outros, poderia ser mais específica. Numa situação ideal, estas imagens estariam separadas pela categoria detalhes e pela categoria da divisão que está mais presente na imagem,
5. O tamanho muito reduzido do conjunto de dados, assim como ao facto de o alvo não estar balanceado torna a modelação muito mais complexa.

5.3. Recomendações

Tendo em conta que se comprovou que as características visuais podem ter influência na atratividade, é importante a FLH investir na forma como capta e dispõe as suas fotos nos anúncios. Para tal, do presente trabalho conseguem-se retirar algumas recomendações que permitem à Feels Like Home refinar a sua estratégia de escolha e disposição de imagens. Recomenda-se que a Feels Like Home:

1. Tenha um cuidado especial em relação à qualidade e divisão da primeira imagem do anúncio,

2. Defina, para os alojamentos da região de Lisboa, como primeira fotografia do anúncio, uma fotografia do quarto,
3. Procure colocar fotografias com uma maior percentagem de verde, ou seja, com elementos de natureza, como fotografias onde se vejam plantas ou jardim,
4. Não utilize imagens com valores de luminosidade extremamente baixos, ou seja, fotografias muito escuras, seja em termos de luz ou decoração,
5. Não utilize fotografias demasiado homogéneas, por exemplo com grande predominância de paredes ou superfícies grandes e lisas (valores de entropia extremamente baixos).

Deve por fim fazer-se notar à FLH que tais regras são obtidas de modelos com uma relativa margem de erro, não devendo ser vistas como regras rígidas. A perceção e experiência do profissional que coloca fotografias no anúncio continuam a ser consideradas essenciais, devendo as recomendações apresentadas ser vistas como tendências a seguir.

5.4. Trabalho futuro

Como trabalho futuro sugere-se que se tome um de dois caminhos. Um primeiro caminho, com o objetivo de aprofundar esta análise, procurando enriquecer o conjunto de dados utilizado com mais informação sobre outros alojamentos do Airbnb e do Booking, por exemplo. De tal conjunto de dados seria possível retirar conclusões aplicáveis a outros países e com um maior histórico, aproveitando para incluir variáveis sobre atrações perto de cada alojamento. Ainda neste caminho também há a hipótese de aprofundar a análise da atratividade do alojamento por tipo de hospede, por região e/ou por valor/noite, visto que, a atratividade pode ser tida de formas diferentes, consoante o contexto.

O segundo caminho possível passaria pela obtenção informação sobre a atratividade de cada imagem, nomeadamente os cliques nas imagens. Tal permitiria aprofundar a análise por imagem e perceber quais são as características intrínsecas às imagens atrativas. Este segundo caminho poderia até resultar num modelo de análise e recomendação automática, definindo quais as melhores imagens a colocar no anúncio.

Referências Bibliográficas

Agustí, D. P. (2018). Characterizing the location of tourist images in cities. Differences in user-generated images (Instagram), official tourist brochures and travel guides. *Annals of Tourism Research*, *73*(August), 103–115. https://doi.org/10.1016/j.annals.2018.09.001

Balomenou, N., & Garrod, B. (2019). Photographs in tourism research: Prejudice, power, performance and participant-generated images. *Tourism Management*, *70,* 201–217. https://doi.org/10.1016/j.tourman.2018.08.014

Balomenou, N., Garrod, B., & Georgiadou, A. (2017). Making sense of tourists' photographs using canonical variate analysis. *Tourism Management*, *61*, 173–179. https://doi.org/10.1016/j.tourman.2017.02.010

Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). *Guide to Intelligent Data Science - How to Intelligently Make Use of Real Data* (D. Gries & O. Hazzan (eds.); Second Edi). Springer. https://doi.org/10.1007/978-3-030-45574-3

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Casado-Díaz, A. B., Pérez-Naranjo, L. M., & Sellers-Rubio, R. (2017). Aggregate consumer ratings and booking intention: the role of brand image. *Service Business*, *11*(3), 543–562. https://doi.org/10.1007/s11628-016-0319-0

FeelsLikeHome. (2020). *About us.* https://rentals.feelslikehome.pt/about-us

Forge, O. (n.d.). *Funtion hsv*. Retrieved May 18, 2022, from https://octave.sourceforge.io/octave/function/hsv.html

Gan, Y. S., Wang, S. Y., Huang, C. E., Hsieh, Y. C., Wang, H. Y., Lin, W. H., Chong, S. N., & Liong, S. T. (2021). How Many Bedrooms Do You Need? A Real-Estate Recommender System from Architectural Floor Plan Images. *Scientific Programming*, *2021*, 1–15. https://doi.org/10.1155/2021/9914557

Kang, Y., Cho, N., Yoon, J., Park, S., & Kim, J. (2021). Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos. *ISPRS International Journal of Geo-Information*, *10*(3), 1–20. https://doi.org/10.3390/ijgi10030137

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society*, *29*(2), 119–127. https://doi.org/https://doi.org/10.2307/2986296

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. In *Keele University Technical Report TR/SE-0401*. https://doi.org/10.1145/3328905.3332505

Kitchenham, B., & Brereton, P. (2013). A systematic review of systematic review process research in software engineering. *Information and Software Technology*, *55*(12), 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010

Kleinlein, R., García-Faura, A., Jiménez, C. L., Montero, J. M., Díaz-De-maría, F., & Fernández-Martínez, F. (2019). Predicting image aesthetics for intelligent tourism information systems. *Electronics*, *8*(6), 1–15. https://doi.org/10.3390/electronics8060671

Kostic, Z., & Jevremovic, A. (2020). What Image Features Boost Housing Market Predictions? *IEEE Transactions on Multimedia*, *22*(7), 1904–1916. https://doi.org/10.1109/TMM.2020.2966890

Larceneux, F., Bezançon, M., & Lefebvre, T. (2018). Asymmetric revelation effect: The influence of an increased number of photos on mental imagery and behavioural responses depending on target market. *Recherche et Applications En Marketing*, *33*(3), 31–60. https://doi.org/10.1177/2051570718785976

Larose, D., & Larose, C. (2015). *Data Mining and Predictive Analytics* (Second Edi). John Wiley &Sons, Inc.

Laureano, R. (2020). *Testes de Hipóteses e Regressão - O meu manual de consulta rápida* (M. Robalo (ed.); 1ª Edição). Edições Sílabo.

Li, M., Bao, Z., Sellis, T., Yan, S., & Zhang, R. (2018). HomeSeeker: A visual analytics system of real estate data. *Journal of Visual Languages and Computing*, *45*, 1–16. https://doi.org/10.1016/j.jvlc.2018.02.001

Li, M., & Fan, N. (2021). Research on Night Tourism Recommendation Based on Intelligent Image Processing Technology. *Scientific Programming*, *2021*. https://doi.org/10.1155/2021/2624621

Li, Y., & Xie, Y. (2020). Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement. *Journal of Marketing Research*, *57*(1), 1–19. https://doi.org/10.1177/0022243719881113

Lien, C. H., Wen, M. J., Huang, L. C., & Wu, K. L. (2015). Online hotel booking: The effects of brand image, price, trust and value on purchase intentions. *Asia Pacific Management Review*, *20*(4), 210–218. https://doi.org/10.1016/j.apmrv.2015.03.005

Lin, C., & Fan, C. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering*, *18*(6), 539–553. https://doi.org/10.1080/13467581.2019.1696203

Loh, W. Y., & Shin, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

Marder, B., Erz, A., Angell, R., & Plangger, K. (2021). The Role of Photograph Aesthetics on Online Review Sites: Effects of Management- versus Traveler-Generated Photos on Tourists' Decision Making. *Journal of Travel Research*, *60*(1), 31–46. https://doi.org/10.1177/0047287519895125

Mariani, M. M., & Borghi, M. (2018). Effects of the Booking.com rating system: Bringing hotel class into the picture. *Tourism Management*, *66*, 47–52. https://doi.org/10.1016/j.tourman.2017.11.006

Meho, L. I., & Yang, K. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125. https://doi.org/https://doi.org/10.1002/asi.20677

Mendes, D., Cruz, F., & Brandão, T. (2022). *The Importance of Accommodation Images in Online Booking Sites - A Systematic Literature Review*.

Mongeon, P., & Paul-Hus, A. (2015). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M.-M., Li, T., Loder, E. W., Mayo-Wilson, E., Mcdonald, S., … Mckenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n160

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). CRISP-DM 1.0. In *CRISP-DM Consortium*. https://the-modeling-agency.com/crisp-dm.pdf

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Quinn, J. (2020). *THE INSIDERS' PREDICTIVE GUIDE TO ANALYTICS* (First Edit). Smart Vision Europe.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students* (Fifth edit). Pearson Education.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Tian, C., & Peng, J. J. (2020). An integrated picture fuzzy anp-todim multi-criteria decision-making approach for tourism attraction recommendation. *Technological and Economic Development of Economy*, *26*(2), 331–354. https://doi.org/10.3846/tede.2019.11412

Wang, Y., & Sparks, B. A. (2016). An Eye-Tracking Study of Tourism Photo Stimuli: Image Characteristics and Ethnicity. *Journal of Travel Research*, *55*(5), 588–602. https://doi.org/10.1177/0047287514564598

Xiao, X., Fang, C., & Lin, H. (2020). Characterizing tourism destination image using photos' visual content. *ISPRS International Journal of Geo-Information*, *9*(12), 1–18. https://doi.org/10.3390/ijgi9120730

Yang, L., Wu, L., Liu, Y., & Kang, C. (2017). Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr. *ISPRS International Journal of Geo-Information*, *6*(11), 1–18. https://doi.org/10.3390/ijgi6110345

You, Q., Pang, R., Cao, L., & Luo, J. (2017). Image-Based Appraisal of Real Estate Properties. *IEEE Transactions on Multimedia*, *19*(12), 2751–2759. https://doi.org/10.1109/TMM.2017.2710804

Yu, J., & Egger, R. (2021). Color and engagement in touristic Instagram pictures: A machine learning approach. *Annals of Tourism Research*, *89*, 1–15. https://doi.org/10.1016/j.annals.2021.103204

Zhang, Q., Liu, Y., Liu, L., Lu, S., Feng, Y., & Yu, X. (2021). Location identification and personalized recommendation of tourist attractions based on image processing. *Traitement Du Signal*, *38*(1), 197–205. https://doi.org/10.18280/TS.380121

Zhu, L., Davis, L. S., & Carr, A. (2021). Visualising natural attractions within national parks: Preferences of tourists for photographs with different visual characteristics. *PLoS ONE*, *16*(6), 1–15. https://doi.org/10.1371/journal.pone.0252661

Anexos

Anexo A: Tabelas e variáveis modelo de dados inicial

Tabela: Reservas

Campos	Tipo	Descrição	Ação tomada
ID	QtD	ID da reserva	Preparação de dados
ID_Cliente	QtD	ID automático da casa (IDAuto_cliente)	Tabela de modelação
Cancelada	QlN	A reversa foi cancelada?	Preparação de dados
Data da Reserva	Data	Data em que foi feita a Reserva	Excluir
Check_in	Data	Data de Check_in da reserva	Preparação de dados
Check_Out	Data	Data de Check_Out da reserva	Preparação de dados
Hora Check In	Hora	Hora prevista do Check_in da reserva	Excluir
Via_site	QlN	Site em que foi feita a reserva	Preparação de dados
Ocupantes	QtD	Número de ocupantes	Preparação de dados
País origem	QlN	País origem do cliente	Preparação de dados
Comissão Agenciamento %	QtC	Comissão Agenciamento (%)	Excluir
Outros serviços - Custos	QtC	Custos com outros serviços	Excluir
Valor_Reserva	QtC	Valor da reserva (incluindo comissão paga ao site)	Excluir
Inserido_Por	QlN	Colaborador que inseriu a reserva	Excluir
Inserido_Em	Hora	Hora de inserção da reserva	Excluir
Modificado_Por	QlN	Último colaborador que modificou a reserva	Excluir
Modificado_Em	Hora	Hora da última modificação da reserva	Excluir
Valor_Total	QtC	Valor da reserva (excluindo comissão do site)	Preparação de dados
AverageSiteComission	QtC	Comissão média do site	Excluir

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Tabela: Airbnb_Avaliações

Campos	Tipo	Descrição	Ação tomada
ID_cliques_airbnb	QtD	ID da casa no Airbnb	Preparação de dados
Listing_title	QlN	Nome da casa	Excluir
jul/20	QlN	Avaliação da propriedade em julho de 2020	Preparação de dados
ago/20	QtC	Avaliação da propriedade em agosto de 2020	Preparação de dados
set/20	QtC	Avaliação da propriedade em setembro de 2020	Preparação de dados
out/20	QtC	Avaliação da propriedade em outubro de 2020	Preparação de dados
nov/20	QtC	Avaliação da propriedade em novembro de 2020	Preparação de dados
dez/20	QtC	Avaliação da propriedade em dezembro de 2020	Preparação de dados
jan/21	QtC	Avaliação da propriedade em janeiro de 2021	Preparação de dados
fev/21	QtC	Avaliação da propriedade em fevereiro de 2021	Preparação de dados
mar/21	QtC	Avaliação da propriedade em março de 2021	Preparação de dados
abr/21	QtC	Avaliação da propriedade em abril de 2021	Preparação de dados
mai/21	QtC	Avaliação da propriedade em maio de 2021	Preparação de dados
jun/21	QtC	Avaliação da propriedade em junho de 2021	Preparação de dados
jul/21	QtC	Avaliação da propriedade em julho de 2021	Preparação de dados
ago/21	QtC	Avaliação da propriedade em agosto de 2021	Preparação de dados
set/21	QtC	Avaliação da propriedade em setembro de 2021	Preparação de dados
out/21	QtC	Avaliação da propriedade em outubro de 2021	Preparação de dados
nov/21	QtC	Avaliação da propriedade em novembro de 2021	Preparação de dados
dez/21	QtC	Avaliação da propriedade em dezembro de 2021	Preparação de dados
jan/22	QtC	Avaliação da propriedade em janeiro de 2022	Preparação de dados

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Tabela: Casas

Campos	Tipo	Descrição	Ação tomada
IDAuto_Cliente	QtD	ID automático da casa	Preparação de dados
Identificação	QtD	ID da casa (visível aos clientes e FLH)	Preparação de dados
Nome da casa	QtD	Nome da casa	Excluir
CoordsGPS	QlN	Coordenadas de GPS	Excluir
Cidade	QlN	Cidade	Preparação de dados
Bairro	QlN	Bairro	Excluir
Inativo	QlN	A casa está inativa?	Excluir
Equipa	QlN	Equipa a que pertence a casa	Excluir
Localização Apartamento	QlN	Localização Apartamento	Preparação de dados
Data_Inicio	Data	Data de entrada da casa na FLH	Preparação de dados
Data_Inativo	Data	Data em que a casa fica inativa	Preparação de dados
Aquecedor	QlN	Tem aquecedor?	Para modelação
WC_CodPostal	QlN	4 digitos do código postal	Excluir
WC_ZonaPostal	QlN	3 digitos do código de zona	Excluir
AirbnbId	QtD	ID da casa no Airbnb	Preparação de dados
Andar	QtD	Andar da casa	Preparação de dados
Elevador	QlN	Tem elevador?	Para modelação
Máx Ocupação	QtD	Nº de ocupantes máximos	Preparação de dados
Tipologia	QlN	Tipologia da casa	Excluir
Tipologia_Class	QlN	Tipologia da casa - Classe	Excluir
Camas_Casal	QtD	Nº de camas de casal	Para modelação
Camas_Solteiro	QtD	Nº de camas de solteiro	Para modelação
Sofás_Camas	QlN	Tem sofas cama?	Para modelação
Camas Extra	QlN	Tem camas extra?	Para modelação
NumWCBasico	QtD	Nº de WC básicos	Preparação de dados
NumWCDuche	QtD	Nº de WC com duche	Preparação de dados
Lugar_Garagem	QlN	Lugar de estacionamento	Preparação de dados
RCM_MaqRoupa	QlN	Tem máquina da roupa	Preparação de dados
RCM_MaqLoica	QlN	Tem máquina da loiça	Preparação de dados
RCM_ArCond	QlN	Tem ar condicionado	Preparação de dados
Welcome_Center	QlN	Welcome Center da casa	Excluir
AutoCheckIn	QlN	CheckIn automático	Excluir

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Tabela: NUTS

Campos	Tipo	Descrição	Ação tomada
Cidade / Conselho	QlN	Cidade ou concelho da casa	Preparação de dados
Distrito / Ilha	QlN	Distrito ou Ilha da casa	Preparação de dados
NUTS I	QlN	NUTS I da casa	Preparação de dados
NUTS II	QlN	NUTS II da casa	Preparação de dados
NUTS III	QlN	NUTS III da casa	Preparação de dados

Nota: QLN - Qualitativa Nominal.

Tabela: Divisões

Campos	Tipo	Descrição	Ação tomada
ID_Divisão	QtD	ID da divisão persente na imagem	Preparação de dados
Descrição divisão	QlN	Descrição do ID_Divisão	Preparação de dados

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Tabela: Airbnb_Search_to_listing

Campos	Tipo	Descrição	Ação tomada
Listing ID	QtD	ID da casa no Airbnb	Preparação de dados
Nome da casa	QlN	Nome da casa	Excluir
Cidade	QlN	Cidade	Excluir
fev/20	QtC	Cliques na página da propriedade em fevereiro de 2020	Preparação de dados
mar/20	QtC	Cliques na página da propriedade em março de 2020	Preparação de dados
abr/20	QtC	Cliques na página da propriedade em abril de 2020	Preparação de dados
mai/20	QtC	Cliques na página da propriedade em maio de 2020	Preparação de dados
jun/20	QtC	Cliques na página da propriedade em junho de 2020	Preparação de dados
jul/20	QtC	Cliques na página da propriedade em julho de 2020	Preparação de dados
ago/20	QtC	Cliques na página da propriedade em agosto de 2020	Preparação de dados
set/20	QtC	Cliques na página da propriedade em setembro de 2020	Preparação de dados
out/20	QtC	Cliques na página da propriedade em outubro de 2020	Preparação de dados
nov/20	QtC	Cliques na página da propriedade em novembro de 2020	Preparação de dados
dez/20	QtC	Cliques na página da propriedade em dezembro de 2020	Preparação de dados
jan/21	QtC	Cliques na página da propriedade em janeiro de 2021	Preparação de dados
fev/21	QtC	Cliques na página da propriedade em fevereiro de 2021	Preparação de dados
mar/21	QtC	Cliques na página da propriedade em março de 2021	Preparação de dados
abr/21	QtC	Cliques na página da propriedade em abril de 2021	Preparação de dados
mai/21	QtC	Cliques na página da propriedade em maio de 2021	Preparação de dados
jun/21	QtC	Cliques na página da propriedade em junho de 2021	Preparação de dados
jul/21	QtC	Cliques na página da propriedade em julho de 2021	Preparação de dados
ago/21	QtC	Cliques na página da propriedade em agosto de 2021	Preparação de dados
set/21	QtC	Cliques na página da propriedade em setembro de 2021	Preparação de dados
out/21	QtC	Cliques na página da propriedade em outubro de 2021	Preparação de dados
nov/21	QtC	Cliques na página da propriedade em novembro de 2021	Preparação de dados
dez/21	QtC	Cliques na página da propriedade em dezembro de 2021	Preparação de dados
jan/22	QtC	Cliques na página da propriedade em janeiro de 2022	Preparação de dados

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua

Tabela: Airbnb_Listing_to_booking

Campos	**Tipo**	**Descrição**	**Ação tomada**
Listing ID	QtD	ID da casa no Airbnb	Preparação de dados
Nome da casa	QlN	Nome da casa	Excluir
Cidade	QlN	Cidade	Excluir
fev/20	QtC	Conversão da propriedade em fevereiro de 2020	Preparação de dados
mar/20	QtC	Conversão da propriedade em março de 2020	Preparação de dados
abr/20	QtC	Conversão da propriedade em abril de 2020	Preparação de dados
mai/20	QtC	Conversão da propriedade em maio de 2020	Preparação de dados
jun/20	QtC	Conversão da propriedade em junho de 2020	Preparação de dados
jul/20	QtC	Conversão da propriedade em julho de 2020	Preparação de dados
ago/20	QtC	Conversão da propriedade em agosto de 2020	Preparação de dados
set/20	QtC	Conversão da propriedade em setembro de 2020	Preparação de dados
out/20	QtC	Conversão da propriedade em outubro de 2020	Preparação de dados
nov/20	QtC	Conversão da propriedade em novembro de 2020	Preparação de dados
dez/20	QtC	Conversão da propriedade em dezembro de 2020	Preparação de dados
jan/21	QtC	Conversão da propriedade em janeiro de 2021	Preparação de dados
fev/21	QtC	Conversão da propriedade em fevereiro de 2021	Preparação de dados
mar/21	QtC	Conversão da propriedade em março de 2021	Preparação de dados
abr/21	QtC	Conversão da propriedade em abril de 2021	Preparação de dados
mai/21	QtC	Conversão da propriedade em maio de 2021	Preparação de dados
jun/21	QtC	Conversão da propriedade em junho de 2021	Preparação de dados
jul/21	QtC	Conversão da propriedade em julho de 2021	Preparação de dados
ago/21	QtC	Conversão da propriedade em agosto de 2021	Preparação de dados
set/21	QtC	Conversão da propriedade em setembro de 2021	Preparação de dados
out/21	QtC	Conversão da propriedade em outubro de 2021	Preparação de dados
nov/21	QtC	Conversão da propriedade em novembro de 2021	Preparação de dados
dez/21	QtC	Conversão da propriedade em dezembro de 2021	Preparação de dados
jan/22	QtC	Conversão da propriedade em janeiro de 2022	Preparação de dados

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Tabela: Imagens

Campos	Tipo	Descrição	Ação tomada
Nome_Imagem	QlN	Nome da Imagem	Excluir
ID_Casa	QtD	ID da casa	Preparação de dados
ID_Ordem_imagem	QtD	Ordem pela qual a casa aparece no anúncio	Preparação de dados
ID_Divisão_imagem	QtD	ID da divisão persente na imagem	Preparação de dados
Verde (%)	QtC	Percentagem de verde na imagem	Preparação de dados
Azul (%)	QtC	Percentagem de azul na imagem	Preparação de dados
Amarelo (%)	QtC	Percentagem de amarelo na imagem	Preparação de dados
Rosa (%)	QtC	Percentagem de rosa na imagem	Preparação de dados
Laranja (%)	QtC	Percentagem de laranja na imagem	Preparação de dados
Vermelho (%)	QtC	Percentagem de vermelho na imagem	Preparação de dados
Nº cores encontradas	QtD	Número de cores encontradas entre: verde, azul, amarelo, rosa, laranja e vermelho	Preparação de dados
entropia_média	QtC	Entropia média da imagem	Preparação de dados
entropia_std	QtC	Desvio padrão da entropia da imagem	Preparação de dados
Lightness_Médio_HLS	QtC	Lightness médio retirado da escala de cores HLS	Preparação de dados
cluster de cor + freq	QtD	Cluster de cor mais frequente retirado da escala de cores HSV	Preparação de dados
Amarelos_Cluster _HSV	QtD	Presença de cluster da cor amarela retirado da escala de cores HSV	Preparação de dados
Azuis_Cluster _HSV	QtD	Presença de cluster da cor azul retirado da escala de cores HSV	Preparação de dados
Laranjas_Cluster _HSV	QtD	Presença de cluster da cor laranja retirado da escala de cores HSV	Preparação de dados
Rosas_Cluster _HSV	QtD	Presença de cluster da cor rosa retirado da escala de cores HSV	Preparação de dados
Verdes_Cluster _HSV	QtD	Presença de cluster da cor verde retirado da escala de cores HSV	Preparação de dados
Vermelhos_Cluster _HSV	QtD	Presença de cluster da cor vermelha retirado da escala de cores HSV	Preparação de dados
Nº de clusters encontrados	QtD	Número de clusters de cores encontrados entre os clusters: verde, azul, amarelo, rosa, laranja e vermelho	Preparação de dados

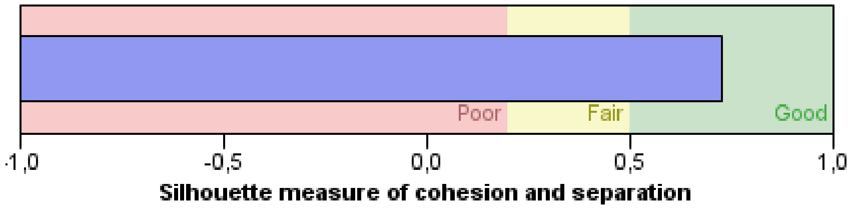
Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua.

Anexo B - Kmeans com os cliques e k=2, para descobrir o limiar do alvo

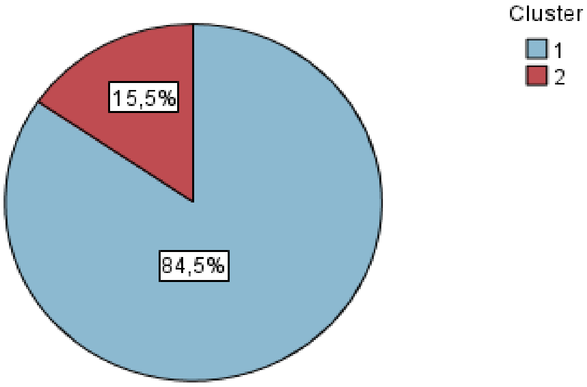
Model Summary

Algorithm	K-Means
Inputs	1
Clusters	2

Cluster Quality



Cluster Sizes



Size of Smallest Cluster	74 (15,5%)
Size of Largest Cluster	404 (84,5%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	5,46

Anexo C – Tabela de modelação: Medida de Associação

Conjunto de variáveis exclusivas da casa

Campo	Tipo	Descrição	Medida de Associação
IDAuto_Cliente	QtD	ID automático da casa	-
Região	QlN	Localização da casa	0.196
Andar_classe	QlN	Classe do andar da casa	0.07
Andar_Recode	QlO	Andar da casa codificado	0.087
Elevador	QlN	Tem elevador (1) ou não (0)	0.062
Máx Ocupação	QtD	Nº de ocupantes máximos	0.174
Máx_Ocupação_Classe	QlN	Classe do número de ocupantes máximo	0.072
Camas_Casal	QtD	Nº de camas de casal	0.119
Camas_Solteiro	QtD	Nº de camas de solteiro	0.114
Sofás_Camas	QlN	Tem sofas cama?	0.069
Camas Extra	QlN	Tem camas extra?	0.058
NumWC	QtD	Número de casas de banho	0.127
Lugar_Garagem	QlN	Lugar de estacionamento (1) ou não (0)	0.036
RCM_ArCond	QlN	Tem ar condicionado (1) ou não (0)	0.109
V_medio_noite	QtC	Valor médio por noite	0.984
target_subtração	QlN	Alvo: é atrativo ou não	-
rating_mensal_médio_BIN	QlO	Classe da avaliação mensal média do Airbnb	0.064
V_medio_noite_BIN	QlO	Classe do valor noite médio	0.192

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua; QlO - Qualitativa Ordinal

Conjunto de variáveis relativas a todas as fotografias

Campo	**Tipo**	**Descrição**	**Medida de Associação**
Cozinha	QtD	Número total de fotografias de cozinha no anúncio	0.108
Fachada	QtD	Número total de fotografias da fachada no anúncio	0.167
Quarto	QtD	Número total de fotografias de quartos no anúncio	0.136
Casa de banho	QtD	Número total de fotografias de casa de banho no anúncio	0.134
Jardim	QtD	Número total de fotografiasdo jardim no anúncio	0.13
Piscina	QtD	Número total de fotografias da piscina no anúncio	0.17
Terraço	QtD	Número total de fotografias do terraço no anúncio	0.071
Garagem	QtD	Número total de fotografias da garagem no anúncio	0.059
Exterior	QtD	Número total de fotografias do exterior no anúncio	0.209
Detalhes	QtD	Número total de fotografias de detalhes no anúncio	0.147
Hall/Sala de Estar	QtD	Número total de fotografias do hall ou sala no anúncio	0.169
Varanda	QtD	Número total de fotografias da varanda no anúncio	0.104
Sala de Jantar	QtD	Número total de fotografias da sala de jantar no anúncio	0.085
Total de Imagens	QtD	Número total de fotografias no anúncio	0.228
Primeira	QlN	Divisão que aparece em primeio nas fotografias do anúncio	0.286
Segunda	QlN	Divisão que aparece em segundo nas fotografias do anúncio	0.157
Terceira	QlN	Divisão que aparece em terceiro nas fotografias do anúncio	0.236
Quarta	QlN	Divisão que aparece em quarto nas fotografias do anúncio	0.179
Quinta	QlN	Divisão que aparece em quinto nas fotografias do anúncio	0.208
Verde (%)_Mean	QtC	Percentagem média de verde nas fotografias do anúncio	1
Azul (%)_Mean	QtC	Percentagem média de azul nas fotografias do anúncio	1
Amarelo (%)_Mean	QtC	Percentagem média de amarelo nas fotografias do anúncio	1
Rosa (%)_Mean	QtC	Percentagem média de rosa nas fotografias do anúncio	1
Laranja (%)_Mean	QtC	Percentagem média de laranja nas fotografias do anúncio	1
Vermelho (%)_Mean	QtC	Percentagem média de vermelho nas fotografias do anúncio	1
Nº cores encontradas_Mean	QtD	Número de cores encontradas nas fotografias do anúncio, em média	0.505
entropia_média_Mean	QtC	Entropia média das fotografias do anúncio	1
entropia_std_Mean	QtC	Desvio padrão da entropia das fotografias do anúncio	1
Lightness_Médio_HLS_Mean	QtC	Luminosidade média das fotografias do anúncio	1

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua; QlO - Qualitativa Ordinal

Conjunto de variáveis relativas à primeira fotografia

Campo	Tipo	Descrição	Medida de Associação
Verde_Mean_1a_img	QtC	Percentagem de verde na primeira fotografia do anúncio	0.889
Azul_Mean_1a_img	QtC	Percentagem de azul na primeira fotografia do anúncio	0.981
Amarelo_Mean_1a_img	QtC	Percentagem de amarelo na primeira fotografia do anúncio	0.822
Rosa_Mean_1a_img	QtC	Percentagem de rosa na primeira fotografia do anúncio	0.773
Laranja_Mean_1a_img	QtC	Percentagem de laranja na primeira fotografia do anúncio	0.929
Vermelho_Mean_1a_img	QtC	Percentagem de vermelho na primeira fotografia do anúncio	0.97
Nº cores encontradas_Mean_1a_img	QtD	Número de cores encontradas na primeira fotografia do anúncio	0.039
entropia_média_Mean_1a_img	QtC	Entropia média da primeira fotografia do anúncio	0.988
entropia_std_Mean_1a_img	QtC	Desvio padrão da entropia da primeira fotografia do anúncio	0.988
Lightness_Médio_HLS_Mean_1a_img	QtC	Luminosidade média da primeira fotografia do anúncio	1

Nota: QtD - Quantitativa Discreta; QLN - Qualitativa Nominal; QtC – Quantitativa Contínua; QlO - Qualitativa Ordinal

Conjunto de variáveis relativas às cinco primeiras fotografias

Campo	Tipo	Descrição	Medida de Associação
Verde_Mean_5_imgs	QtC	Percentagem média de verde nas primeiras 5 fotografias do anúncio	0.962
Azul_Mean_5_imgs	QtC	Percentagem média de azul nas primeiras 5 fotografias do anúncio	0.986
Amarelo_Mean_5_imgs	QtC	Percentagem média de amarelo nas primeiras 5 fotografias do anúncio	0.98
Rosa_Mean_5_imgs	QtC	Percentagem média de rosa nas primeiras 5 fotografias do anúncio	0.966
Laranja_Mean_5_imgs	QtC	Percentagem média de laranja nas primeiras 5 fotografias do anúncio	0.986
Vermelho_Mean_5_imgs	QtC	Percentagem média de vermelho nas primeiras 5 fotografias do anúncio	0.99
Nº cores encontradas_Mean_5_imgs	QtD	Número de cores encontradas nas primeiras 5 fotografias do anúncio, em média	0.076
entropia_média_Mean_5_imgs	QtC	Entropia média das primeiras 5 fotografias do anúncio	0.997
entropia_std_Mean_5_imgs	QtC	Desvio padrão da entropia das primeiras 5 fotografias do anúncio	0.997
Lightness_Médio_HLS_Mean_5_imgs	QtC	Luminosidade média das primeiras 5 fotografias do anúncio	1

Anexo D – Tabela de modelação: Medida de Associação

Variáveis exclusivas da casa

Campo	Média	Desvio-padrão	Mín.	Mediana	Máx.	Não nulos	Nulos	Valores distintos
IDAuto_Cliente	898	504	1	790	1800	478	0	478
Máx Ocupação	4	2	1	4	12	478	0	12
Camas_Casal	1	1	0	1	6	478	0	6
Camas_Solteiro	1	2	0	2	9	478	0	9
Sofás_Camas	0	1	0	0	2	478	0	3
Camas Extra	0	0	0	0	2	478	0	3
NumWC	2	1	0	1	8	478	0	8
V_medio_noite	67.79	45.65	0.00	57.57	372.88	478	0	465

Variáveis relativas a todas as fotografias

Campo	Média	Desvio-padrão	Mín.	Mediana	Máx.	Não nulos	Nulos	Valores distintos
Cozinha	2	1	0	2	8	478	0	9
Fachada	1	1	0	1	6	478	0	7
Quarto	5	3	0	4	24	478	0	16
Casa de banho	2	1	0	2	10	478	0	8
Jardim	0	1	0	0	9	478	0	9
Piscina	0	1	0	0	7	478	0	8
Terraço	0	1	0	0	8	478	0	8
Garagem	0	0	0	0	2	478	0	3
Exterior	2	3	0	0	19	478	0	15
Detalhes	2	2	0	1	10	478	0	11
Hall/Sala de Estar	2	2	0	1	10	478	0	11
Varanda	1	1	0	0	14	478	0	7
Sala de Jantar	0	1	0	0	4	478	0	5
Total de Imagens	20	7	9	18	55	478	0	35
Verde (%)_Mean	5.77	3.67	0.55	4.82	24.96	478	0	478
Azul (%)_Mean	63.84	13.85	13.63	64.58	95.87	478	0	478
Amarelo (%)_Mean	2.73	1.97	0.05	2.17	10.49	478	0	477
Rosa (%)_Mean	2.01	2.49	0.18	1.37	21.48	478	0	478
Laranja (%)_Mean	8.02	4.60	0.30	7.44	25.35	478	0	478
Vermelho (%)_Mean	17.62	9.43	2.13	15.47	57.30	478	0	478
Nº cores encontradas_Mean	5.91	0.18	4.52	6.00	6.00	478	0	108
entropia_média_Mean	3.28	0.40	2.08	3.26	4.51	478	0	478
entropia_std_Mean	1.17	0.15	0.75	1.16	1.75	478	0	478
Lightness_Médio_HLS_Mean	148.33	12.75	93.07	148.00	176.87	478	0	478

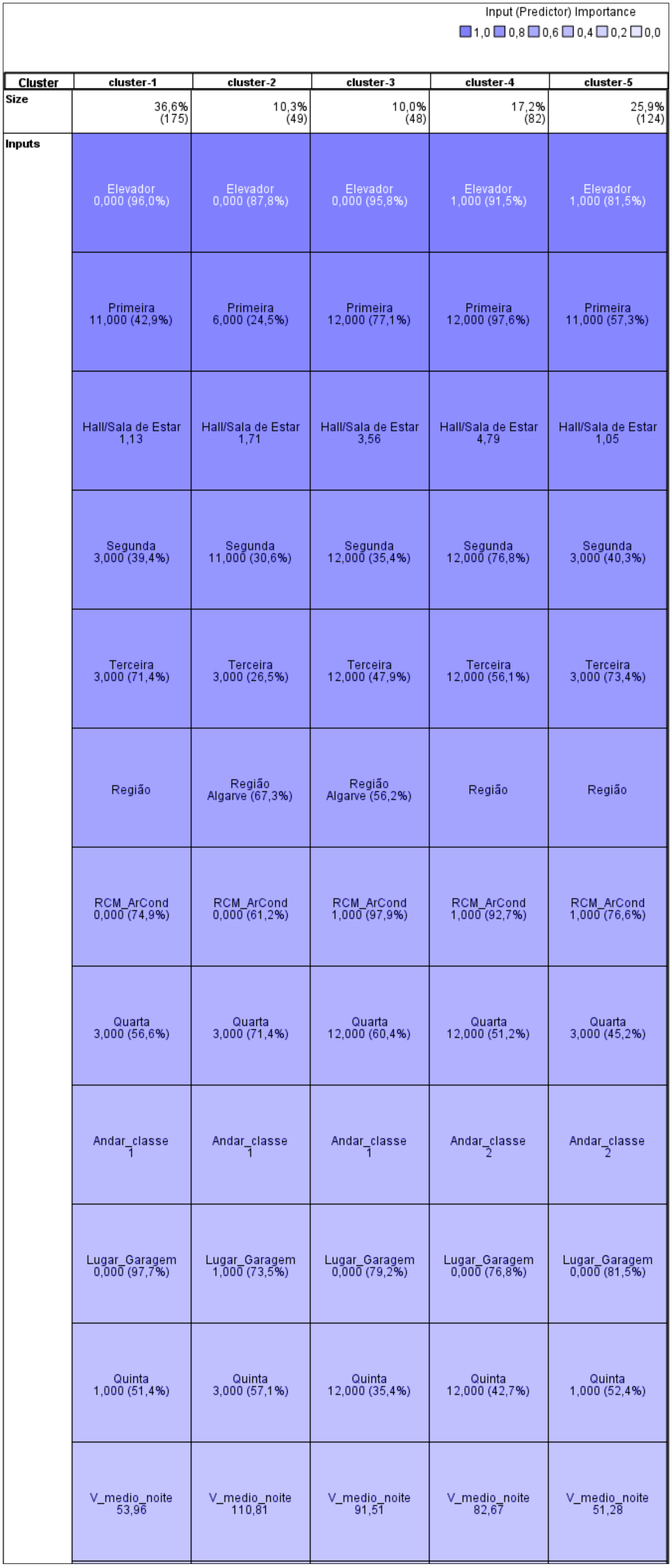
Variáveis relativas à primeira fotografia

Campo	Média	Desvio-padrão	Mín.	Mediana	Máx.	Não nulos	Nulos	Valores distintos
Verde_Mean_1a_img	5.55	7.60	0.01	3.16	63.99	478	0	375
Azul_Mean_1a_img	64.59	19.78	5.94	66.25	99.02	478	0	463
Amarelo_Mean_1a_img	2.82	3.61	0.00	1.62	30.54	478	0	327
Rosa_Mean_1a_img	2.20	3.91	0.02	1.07	40.14	478	0	278
Laranja_Mean_1a_img	8.22	8.48	0.00	5.44	44.57	478	0	417
Vermelho_Mean_1a_img	16.62	13.86	0.16	12.69	75.51	478	0	452
Nº cores encontradas_Mean_1a_img	6	0	4	6	6	478	0	3
entropia_média_Mean_1a_img	3.36	0.52	2.27	3.28	5.10	478	0	471
entropia_std_Mean_1a_img	1.24	0.23	0.53	1.22	1.91	478	0	471
Lightness_Médio_HLS_Mean_1a_img	145.33	16.97	89.91	145.90	195.57	478	0	478

Variáveis relativas às cinco primeiras fotografias

Campo	Média	Desvio-padrão	Mín.	Mediana	Máx.	Não nulos	Nulos	Valores distintos
Verde_Mean_5_imgs	4.72	3.92	0.14	3.62	33.63	478	0	454
Azul_Mean_5_imgs	65.80	15.65	8.23	67.12	97.62	478	0	470
Amarelo_Mean_5_imgs	2.57	2.48	0.01	1.84	19.94	478	0	453
Rosa_Mean_5_imgs	2.16	2.80	0.04	1.37	22.39	478	0	440
Laranja_Mean_5_imgs	7.31	5.15	0.05	6.58	27.10	478	0	471
Vermelho_Mean_5_imgs	17.43	10.81	1.60	15.26	66.59	478	0	473
Nº cores encontradas_Mean_5_imgs	5.97	0.16	4.60	6.00	6.00	478	0	8
entropia_média_Mean_5_imgs	3.21	0.41	2.11	3.18	4.49	478	0	477
entropia_std_Mean_5_imgs	1.18	0.18	0.70	1.17	1.89	478	0	477
Lightness_Médio_HLS_Mean_5_imgs	148.13	13.30	95.97	149.39	180.87	478	0	478

Anexo E - Características dos clusters, adaptado do output do IBM Modeler.

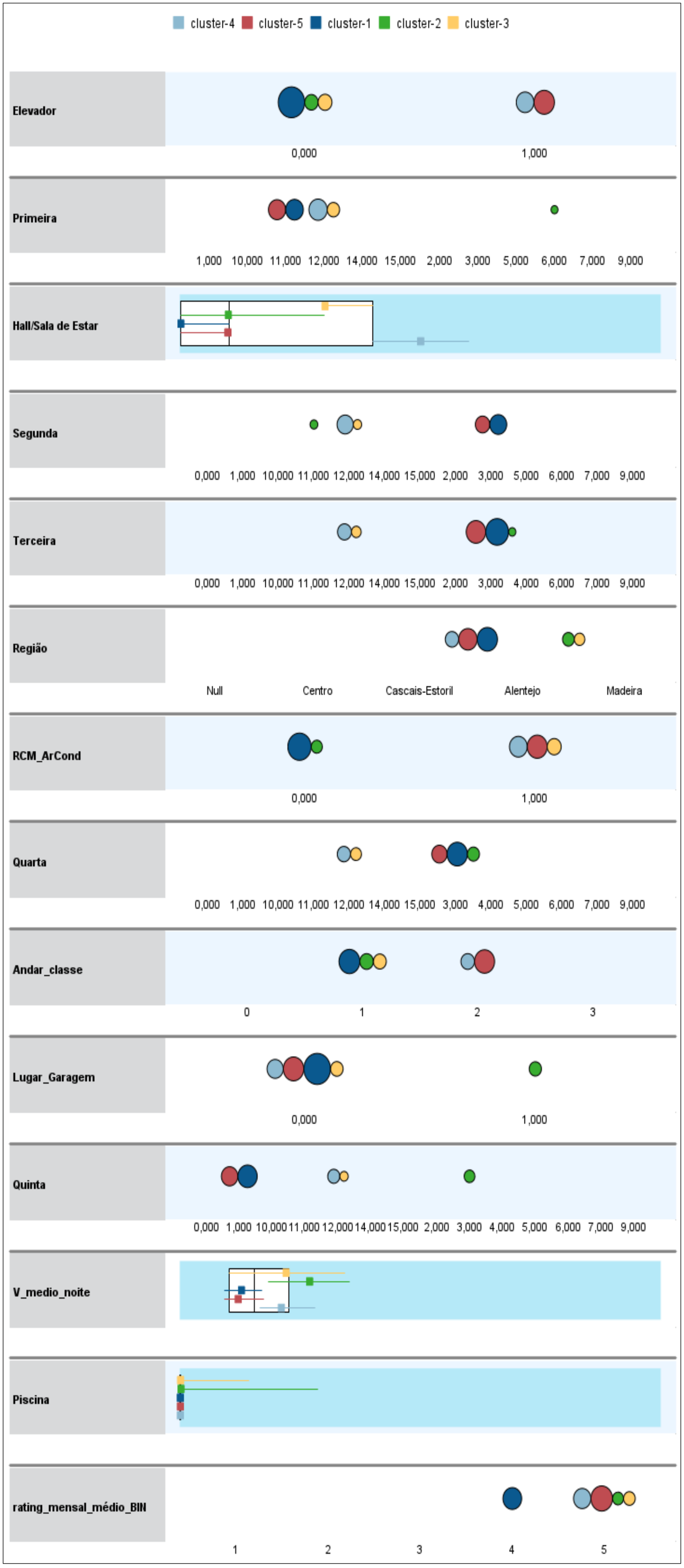


Input (Predictor) Importance

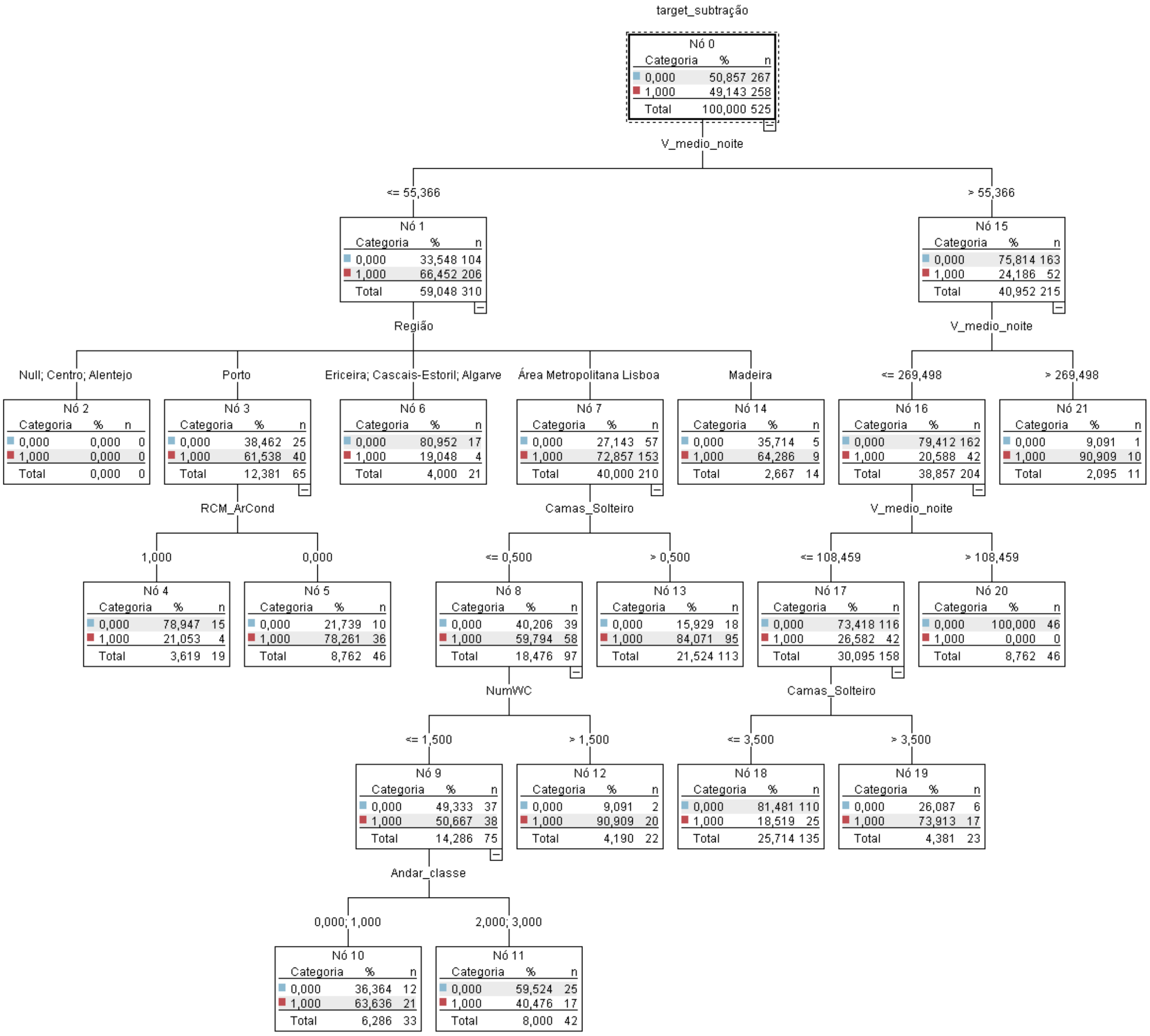
■1,0 ■0,8 ■0,6 ■0,4 ■0,2 □0,0

Cluster	cluster-1	cluster-2	cluster-3	cluster-4	cluster-5
Size	36,6% (175)	10,3% (49)	10,0% (48)	17,2% (82)	25,9% (124)
Inputs	Elevador 0,000 (96,0%)	Elevador 0,000 (87,8%)	Elevador 0,000 (95,8%)	Elevador 1,000 (91,5%)	Elevador 1,000 (81,5%)
	Primeira 11,000 (42,9%)	Primeira 6,000 (24,5%)	Primeira 12,000 (77,1%)	Primeira 12,000 (97,6%)	Primeira 11,000 (57,3%)
	Hall/Sala de Estar 1,13	Hall/Sala de Estar 1,71	Hall/Sala de Estar 3,56	Hall/Sala de Estar 4,79	Hall/Sala de Estar 1,05
	Segunda 3,000 (39,4%)	Segunda 11,000 (30,6%)	Segunda 12,000 (35,4%)	Segunda 12,000 (76,8%)	Segunda 3,000 (40,3%)
	Terceira 3,000 (71,4%)	Terceira 3,000 (26,5%)	Terceira 12,000 (47,9%)	Terceira 12,000 (56,1%)	Terceira 3,000 (73,4%)
	Região	Região Algarve (67,3%)	Região Algarve (56,2%)	Região	Região
	RCM_ArCond 0,000 (74,9%)	RCM_ArCond 0,000 (61,2%)	RCM_ArCond 1,000 (97,9%)	RCM_ArCond 1,000 (92,7%)	RCM_ArCond 1,000 (76,6%)
	Quarta 3,000 (56,6%)	Quarta 3,000 (71,4%)	Quarta 12,000 (60,4%)	Quarta 12,000 (51,2%)	Quarta 3,000 (45,2%)
	Andar_classe 1	Andar_classe 1	Andar_classe 1	Andar_classe 2	Andar_classe 2
	Lugar_Garagem 0,000 (97,7%)	Lugar_Garagem 1,000 (73,5%)	Lugar_Garagem 0,000 (79,2%)	Lugar_Garagem 0,000 (76,8%)	Lugar_Garagem 0,000 (81,5%)
	Quinta 1,000 (51,4%)	Quinta 3,000 (57,1%)	Quinta 12,000 (35,4%)	Quinta 12,000 (42,7%)	Quinta 1,000 (52,4%)
	V_medio_noite 53,96	V_medio_noite 110,81	V_medio_noite 91,51	V_medio_noite 82,67	V_medio_noite 51,28

Anexo F - Comparação dos clusters, adaptado do output do IBM Modeler.



Anexo G – Arvore de decisão do modelo sem informação de imagens


target_subtração
Nó 0
Categoria % n
0,000 50,857 267
1,000 49,143 258
Total 100,000 525
V_medio_noite
<= 55,366
> 55,366
Nó 1
Categoria % n
0,000 33,548 104
1,000 66,452 206
Total 59,048 310
Nó 15
Categoria % n
0,000 75,814 163
1,000 24,186 52
Total 40,952 215
Região
V_medio_noite
Null; Centro; Alentejo
Porto
Ericeira; Cascais-Estoril; Algarve
Área Metropolitana Lisboa
Madeira
<= 269,498
> 269,498
Nó 2
Categoria % n
0,000 0,000 0
1,000 0,000 0
Total 0,000 0
Nó 3
Categoria % n
0,000 38,462 25
1,000 61,538 40
Total 12,381 65
Nó 6
Categoria % n
0,000 80,952 17
1,000 19,048 4
Total 4,000 21
Nó 7
Categoria % n
0,000 27,143 57
1,000 72,857 153
Total 40,000 210
Nó 14
Categoria % n
0,000 35,714 5
1,000 64,286 9
Total 2,667 14
Nó 16
Categoria % n
0,000 79,412 162
1,000 20,588 42
Total 38,857 204
Nó 21
Categoria % n
0,000 9,091 1
1,000 90,909 10
Total 2,095 11
RCM_ArCond
Camas_Solteiro
V_medio_noite
1,000
0,000
<= 0,500
> 0,500
<= 108,459
> 108,459
Nó 4
Categoria % n
0,000 78,947 15
1,000 21,053 4
Total 3,619 19
Nó 5
Categoria % n
0,000 21,739 10
1,000 78,261 36
Total 8,762 46
Nó 8
Categoria % n
0,000 40,206 39
1,000 59,794 58
Total 18,476 97
Nó 13
Categoria % n
0,000 15,929 18
1,000 84,071 95
Total 21,524 113
Nó 17
Categoria % n
0,000 73,418 116
1,000 26,582 42
Total 30,095 158
Nó 20
Categoria % n
0,000 100,000 46
1,000 0,000 0
Total 8,762 46
NumWC
Camas_Solteiro
<= 1,500
> 1,500
<= 3,500
> 3,500
Nó 9
Categoria % n
0,000 49,333 37
1,000 50,667 38
Total 14,286 75
Nó 12
Categoria % n
0,000 9,091 2
1,000 90,909 20
Total 4,190 22
Nó 18
Categoria % n
0,000 81,481 110
1,000 18,519 25
Total 25,714 135
Nó 19
Categoria % n
0,000 26,087 6
1,000 73,913 17
Total 4,381 23
Andar_classe
0,000; 1,000
2,000; 3,000
Nó 10
Categoria % n
0,000 36,364 12
1,000 63,636 21
Total 6,286 33
Nó 11
Categoria % n
0,000 59,524 25
1,000 40,476 17
Total 8,000 42

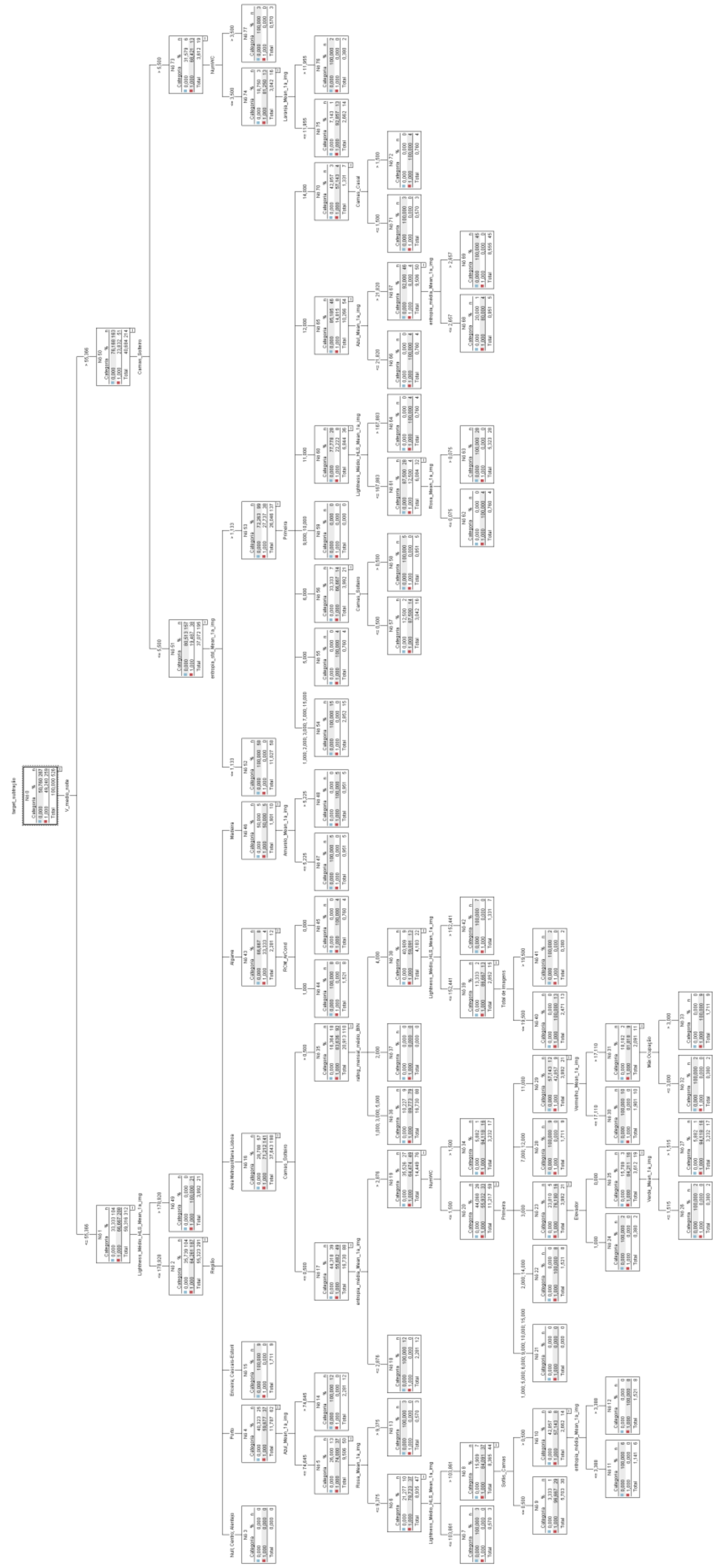
Regras de alojamentos atrativos:

- Nó 12: V_medio_noite <= 55,366; Região in [5]; Camas_Solteiro <= 0,500; NumWC > 1,500 (22; 0,909).
- Nó 21: V_medio_noite > 55,366; V_medio_noite > 269,498 (11; 0,909).

Regras de alojamentos não atrativos:

- Nó 6: V_medio_noite <= 55,366; Região in [3, 4, 7] (21; 0,81)
- Nó 20: V_medio_noite > 55,366; V_medio_noite <= 269,498; V_medio_noite > 108,459 (46; 1,0)

Anexo H - Arvore de decisão do modelo com informação da primeira imagem



Regras de alojamentos atrativos:

- Nó 49: V_medio_noite <= 55,366; Lightness_Médio_HLS_Mean_1a_img > 178,928 (21; 1,0)
- Nó 9: V_medio_noite <= 55,366; Lightness_Médio_HLS_Mean_1a_img <= 178,928; Região in [1]; Azul_Mean_1a_img <= 74,645; Rosa_Mean_1a_img <= 9,375; Lightness_Médio_HLS_Mean_1a_img > 103,861; Sofás_Camas <= 0,500 (30; 0,967)
- Nó 27: V_medio_noite <= 55,366; Lightness_Médio_HLS_Mean_1a_img <= 178,928; Região in [5]; Camas_Solteiro <= 0,500; entropia_média_Mean_1a_img > 2,876; NumWC <= 1,500; Primeira in [3.000]; Elevador = 0,000; Verde_Mean_1a_img > 1,515 (17; 0,941)

Regras de alojamentos não atrativos

- Nó 52: V_medio_noite > 55,366; Camas_Solteiro <= 5,500; entropia_std_Mean_1a_img <= 1,133 (58; 1,0)
- Nó 63: V_medio_noite > 55,366; Camas_Solteiro <= 5,500; entropia_std_Mean_1a_img > 1,133; Primeira in [11.000]; Lightness_Médio_HLS_Mean_1a_img <= 167,883; Rosa_Mean_1a_img > 0,075 (28; 1,0)
- Nó 69: V_medio_noite > 55,366; Camas_Solteiro <= 5,500; entropia_std_Mean_1a_img > 1,133; Primeira in [12.000]; Azul_Mean_1a_img > 21,820; entropia_média_Mean_1a_img > 2,657 (45; 1,0)