INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Aplicação De Modelos De *Machine Learning* Para Previsão Do *Housing Price* Em Singapura

António Muinga Francisco

Mestrado em Economia Monetária e Financeira

Orientadora:
Dra. Sofia de Sousa Vale, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Co-Orientadora:
Dra. Diana Elisabeta Aldea Mendes, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Setembro, 2024

CIÊNCIAS SOCIAIS
E HUMANAS

Departamento de Economia Política

Aplicação De Modelos De *Machine Learning* Para Previsão Do *Housing Price* Em Singapura

António Muinga Francisco

Mestrado em Economia Monetária e Financeira

Orientadora:
Dra. Sofia de Sousa Vale, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Co-Orientadora:
Dra. Diana Elisabeta Aldea Mendes, Professora Associada,
ISCTE - Instituto Universitário de Lisboa

Setembro, 2024

Dedicatória

Dedico este trabalho à minha família!

Agradecimento

Antes de tudo, agradeço primeiramente a Deus pela vida e saúde, força e por permitir que este trabalho fosse feito, sem Ele nada é possível: “posso todas as coisas Naquele que me fortalece”.

Agradeço à minha família, a minha orientadora *Dra. Sofia Vale* e co-orientadora, *Dra. Diana Mendes*, pelos esforços incansáveis que fizeram durante a realização deste trabalho.

Um agradecimento especial aos demais professores do *Mestrado em Economia Monetária e Financeira*, principalmente: o Dr. Sérgio Lagoa, Dr. Emanuel Leão, Dra. Sofia Vale e o Dr. José Dias Curto, pelas ricas lições e conhecimentos transmitidos durante este percurso pois sinto-me verdadeiramente como o produto do vosso trabalho.

Os meus agradecimentos ao *ISCTE (Instituto Universitário de Lisboa)* pelos incontáveis recursos académicos que tem disponibilizado, de formas a garantir aos estudantes um verdadeiro ambiente de pesquisas científicas, que contribuam para a geração de novos conhecimentos e inovações.

A todos, a minha eterna gratidão.

Resumo

A habitação constitui uma das necessidades mais essenciais da vida humana assim como outras necessidades fundamentais como a comida, água, segurança. Mas simultaneamente é um bem de investimento e os imóveis são ativos reais cujo preço se relaciona com o ciclo económico e, portanto, com a inflação, a taxa de juro, etc., razão pela qual, flutuações no preço da habitação sempre constituem preocupações para a sociedade, tornando o setor imobiliário num campo interessante de pesquisa.

Este trabalho representa um contributo inovador pelos métodos e técnicas utilizados ao aplicar *Machine Learning* (ML), uma abordagem recente, para fazer previsões com elevado nível de acurácia e precisão. O trabalho utiliza algoritmos de regressão *LASSO* e *Random Forest*, para criar modelos de ML e, como principais métricas para avaliar a sua *performance*, o RMSE (*Root Mean Squared Error*), R^2, (coeficiente de determinação) e MAE (*Mean Absolute Error*).

O mercado de habitação analisado é Singapura, para o qual existe informações sobre as transações de venda de imoveis disponibilizados pelo *Housing & Development Board* (HDB), e o período considerado foi Janeiro de 1990 a Dezembro de 2023. O modelo *Random Forest* apresentou a melhor performance, tendo sido selecionado como o modelo final para a previsão do preço da habitação, cujos resultados ajudaram identificar os principais fatores determinantes do preço da habitação, e fornecem informações valiosas para formulação de políticas de habitação, especialmente para o controle e regulação de preços da habitação em Singapura.

Palavras-chave: Habitação, Preço, *Machine Learning*, LASSO, *Random Forest*, Previsão.

Abstract

Housing constitutes one of the most essential needs of human life, as well as other fundamental needs such as food, water and security. But at the same time, it is an investment good and real estate is a real asset whose price is related to the economic cycle and, therefore, to inflation, interest rates, etc., which is why fluctuations in the price of housing are always concerns for society, making the real estate sector an interesting field of research.

This work represents an innovative contribution due to the methods and techniques used when applying Machine Learning (ML), a recent approach, to make predictions with a high level of accuracy and precision. The work uses LASSO and Random Forest regression algorithms to create ML models and, as main metrics to evaluate their performance, RMSE (Root Mean Squared Error), R^2, (coefficient of determination) and MAE (Mean Absolute Error).

The housing market analysed is Singapore, for which there is information on property sales transactions made available by the Housing & Development Board (HDB), and the period considered was January 1990 to December 2023. The Random Forest model presented the best performance, having been selected as the final model for forecasting housing prices, the results of which helped identify the main factors determining housing prices, and provide valuable information for the formulation of housing policies, especially for the control and regulation of housing prices in Singapore.

Keywords: Housing, Price, Machine Learning, LASSO, Random Forest, Forecast.

Índice

Índice de quadros e figuras

Glossário

ML – *Machine Learning*

R – RStudio

IA – Inteligência Artificial

SVM – *Suporte Vector Machine*

RF – *Randon Forest*

HDB - *Housing & Development Board*

ONS - *Office for National Statistics - ONS & Registry*

RNA - Redes Neuronais Artificiais

PIB – Produto Interno Bruto

ONU – Organização das Nações Unidas

BC – Banco Central

IDH – Índice de Desenvolvimento Humano

INE – Instituto Nacional de Estatística

LASSO - *Least Absolute Shrinkage and Selection Operator*

AD – Árvores de Decisão

CV – *Cross Validation*

RMSE – *Root Mean Squared Error*

MAE – *Mean Absolute Error*

RMSPE - *Root Mean Squared Percentage Error*

MAPE - *Mean Absolute Percentage Error*

CAPÍTULO I

Introdução

1.1. Introdução

Diariamente, são gerados dados na sociedade fruto da interação entre os principais agentes económicos (famílias, empresas e governos), na realização de transações relacionadas com o consumo, investimento, e poupança alocando recursos na economia (Abreu et al., 2018). Estes dados de natureza e características diversas, impactam os vários setores da economia de forma direta ou indireta quando recolhidos, armazenados, processados e analisados por diversas entidades (Kaffash et al., 2021; Yaseen & Obaid, 2020; Li et al., 2015). Agregam maior valor e geram informações úteis que servem de suporte para tomadas de decisões dos agentes económicos.

Como nos demais setores económicos, no setor da habitação podemos encontrar diversos dados dentre os quais, os relacionados com o preço de casas tanto para venda, aluguer para habitação ou de curta duração, são gerados em grande volume, variedade e velocidade (*Housing & Development Board* - HDB, 2022; *Office for National Statistics - ONS & Registry*, 2024). Estas e outras entidades demonstram que o preço da habitação é uma variável cuja importância suscita interesse por parte do Estado, empresas e famílias ao redor do mundo e que, os dados sobre a habitação encontram-se disponíveis em diversas fontes e formatos.

Atualmente, centenas de plataformas *online* atuam como agregadores de anúncios de diversas fontes, incluindo imobiliárias, proprietários particulares, de modo a oferecer serviços e informações sobre o preço da habitação em cada país. A nível de países podemos encontrar sites como: *Rightmoveo* e *OnTheMarket* no Reino Unido, *Zillow* e *Realtor* nos Estados Unidos da América, *Huttons e Savills Singapore*, em Singapura, *Imovirtual e BPI Expresso Imobiliário* em Portugal, etc. Outros como o *Century 21 Global, Remax Global, Idealista, Airbnb*, etc, também oferecem uma ampla seleção de imóveis para venda ou aluguer para moradia e de curta duração ao redor do mundo gerando volumes de dados importantes sobre o preço da habitação.

As questões relacionadas com a habitação sempre despertaram a atenção da sociedade em geral e dos governos em particular, principalmente após a crise financeira de 2008 com origem nos Estados Unidos da América (EUA), cujas consequências afetaram as principais economias, quando o preço da habitação caiu 30% após ter registado várias subidas entre 2000 e 2007, causando uma grande crise financeira que despertou maior interesse em estudos sobre os impactos macroeconómicos da habitação e do mercado imobiliário (Blanchard et al., 2021; Camões & Vale, 2018).

O Estado, através dos mecanismos que tem à sua disposição, recolhe dados sobre o preço da habitação, processa e analisa formas de desenhar políticas públicas para este setor e medir os seus impactos, empresas privadas e setores da economia, ao observarem o comportamento desta variável, podem extrair informações úteis que servem de base para a formulação de estratégias e tomada de decisões, assim como para explorar novas oportunidades de investimentos (Kang et al., 2020;

Zaman et al., 2021; Zulkifley et al., 2020). E por fim também as famílias, na aquisição de casa para habitação ou como forma de investimento, procuram informações sobre o preço da habitação que tem sido parte fundamental no portfólio de ativo das famílias (Camões & Vale, 2018).

Com o desenvolvimento tecnológico verificado nos últimos anos, acompanhado de economias de escala, os dados, dentre os quais sobre o preço da habitação, são gerados em grande volume, velocidade e variedade, transcendendo os métodos e técnicas tradicionais de processamento e análise de dados, bem como a capacidade humana de extrair informações a partir de grandes volumes de dados com elevada rapidez e precisão para previsão e tomada de decisões (Prüfer & Schottmüller, 2021). Estas limitações, deram lugar a aplicação de algoritmos de *Machine Learning* (ML) em grande escala nos vários setores de atividades económica (Obschonka & Audretsch, 2020).

Assim, tendo em conta os aspetos acima expostos, a importância que a habitação tem para o bem-estar de uma sociedade e o impacto que o preço da habitação tem sobre a economia (Baldominos et al., 2018; Teixeira et al., 2010), este trabalho propõe-se responder a um conjunto de questões, a saber: De que forma podemos aplicar algoritmos de *Machine Learning* (ML) para fazer previsão do preço da habitação? Qual é a importância do preço da habitação para economia? Quais são os atributos específicos que mais influenciam o preço da habitação? Qual é o modelo de ML mais adequado para fazer previsão do preço da habitação com base nas caraterísticas específicas dos dados? Desta forma, o principal objetivo deste trabalho é aplicar modelos de *Machine Learning* para prever o preço da habitação em Singapura durante o período compreendido entre Janeiro de 1990 e Dezembro de 2023.

Prever o preço da habitação não é uma tarefa nova, mas este trabalho representa um contributo inovador pelos métodos e técnicas utilizados ao aplicar *Machine Learning*, uma abordagem recente para fazer previsões com elevado nível de acurácia e precisão não alcançados muitas vezes com os métodos tradicionais de previsão, principalmente com o surgimento do *Big Data* em vários setores da economia (Mahesh, 2018).

As metodologias de ML destacam-se cada vez mais por oferecer uma ampla quantidade de ferramentas e técnicas que permitem lidar com dados mais complexos, gerados de forma continua e em grandes volumes, não dependem de suposições estatísticas rígidas e são capazes de fornecer previsões mais precisas e generalizáveis (Akansu et al., 2016; Zulkifley et al., 2020). As metodologias de ML são assim uma escolha atraente para aplicações numa época em que a complexidade e a quantidade dos dados que são gerados excedem os métodos tradicionais de previsão, exigindo novos paradigmas que alteram a nossa forma de olhar e tratar os dados com objetivo de criar soluções mais eficazes (Mei & Shi, 2024; Shafiee et al., 2021).

Este trabalho diferencia-se dos estudos realizados sobre a previsão do preço da habitação na medida em que emprega, não somente modelos de ML, mas também as técnicas automáticas baseadas em ML para a seleção dos atributos mais relevantes, de forma a construir modelos mais simples que possam ser generalizados para novos dados e calcular a previsão com elevada precisão, o que representa um

diferencial em relação aos muitos estudos e trabalhos feitos neste domínio (Emmert-Streib & Dehmer, 2019; Ranstam & Cook, 2018; Zulkifley et al., 2020; Zhang et al., 2021).

1.2. Contexto

Singapura, é um país localizado no sudeste da Ásia e faz parte dos tigres asiáticos, isto é, Hong Kong, Singapura, Coreia do Sul e Taiwan, um grupo de quatro países do Sudeste Asiático que experimentaram um rápido crescimento económico e industrialização entre as décadas de 1960 e 1990. São vistos como termómetros económicos importantes para o mundo, ao destacarem-se no fornecimento de suprimentos tecnológicos como *chips* que são importantes para redes de telecomunicações 5G e processamento de *Big Data* ((Basu & Ferreira, 2020; The Economist, 2019).

De acordo com os relatórios do fórum económico mundial, Singapura foi classificada no top 10 entre as economias mais competitivas do mundo no período de 2017 a 2018 (Schwab, 2017), título que continuou a obter nos anos subsequentes, isto é, entre 2019 e 2023 sempre entre as maiores economias do mundo e outras mais competitivas como os: Estados Unidos da América, a Alemanha, o Reino Unido, a Suíça, etc, (Bris et al., 2020; Schwab, 2019; World Bank, 2019).

Singapura oferece um ambiente regulatório para negócios dos mais favoráveis do mundo, e o crescimento do seu PIB tem sido um dos mais altos do mundo crescendo em média 7,7% desde a independência. Hoje é considerada como uma das cidades mais habitáveis e com um índice de capital humano mais elevado do mundo (World Bank, 2019).

Quando se fala de habitação em Singapura, a principal referência que surge é o *Housing & Development Board* (HDB), a autoridade de habitação pública de Singapura fundada em 1960 durante a crise habitacional que afetou o país. O seu objetivo é planear e desenvolver vários conjuntos habitacionais, isto é, construir casas e transformar cidades para criar um ambiente de vida de qualidade para todos os cidadãos de Singapura (HDB, 2022). Atualmente, de acordo com este organismo, a habitação pública de Singapura abriga uma nação inteira e os apartamentos do HDB representam 80% da população residente de Singapura, dos quais cerca de 90% possuem casa própria (HDB, 2022; The Economist, 2017).

O setor da habitação em Singapura é fortemente influenciado pelas políticas governamentais através da autoridade da habitação publica de Singapura, o HDB, que aplica rigorosos regulamentos ao setor imobiliário (público e privado) de formas a garantir habitação para todos os residentes de Singapura (Bian et al., 2019). Através do HDB, o governo vende apartamentos a preços acessíveis aos cidadãos de Singapura e quando estes apartamentos são revendidos, os preços são influenciados por políticas governamentais ajustadas aos fatores macroeconómicos do país (Lin-Heng, 2020).

Pelos feitos realizados, o HDB já foi agraciado em 2010 com o prémio *Scroll of Honour* da ONU-HABITAT, um programa das Nações Unidas lançado em 1989 que trabalha para um futuro urbano melhor, de formas a promover um desenvolvimento social e ambientalmente sustentável, pelas melhores iniciativas do mundo que fizeram contribuições notáveis em fornecer moradias e qualidade de vida

urbana contribuindo de forma significativa para a criação de “cidades e comunidades sustentáveis” (The United Nations Human Settlements Programme [ONU-HABITAT], 2024).

1.3. Estrutura do trabalho

O trabalho está estruturado da seguinte forma: o primeiro capítulo é a introdução, no segundo capítulo fez-se uma revisão da literatura relacionada com o *preço da habitação*, sua importância para a sociedade e na economia, bem como sua relação com certas variáveis macroeconómicas.

O terceiro capítulo faz referência aos dados e a metodologia utilizada para tratamento dos dados, onde se faz uma breve abordagem sobre Machine Learning (ML), principais tipos de algoritmos de ML, e a definição dos algoritmos de ML usado neste trabalho, seguido de uma explicação sobre a recolha, análise e exploração dos dados, e as fases para a construção de um modelo de ML e sua aplicação prática.

O quarto capítulo estará relacionado com a análise e interpretação de resultados, indicando dentre vários modelos de ML aquele que apresentar melhor *performance* ou desempenho na previsão do *preço da habitação*, isto é, a escolha do modelo final. Vindo a seguir as principais conclusões sobre o trabalho no quinto capítulo.

CAPÍTULO II

Revisão da Literatura

2.1. Preço da habitação

Quando olhamos para as principais agendas de governos, é difícil não encontrarmos políticas públicas relacionadas com a habitação. Pela importância que a habitação tem para a sociedade é muitas vezes mencionada em diversos protocolos de âmbito internacional tais como: na Declaração Universal dos Direitos Humanos de 10 de Dezembro de 1948, artigo 25º, e no Pacto Internacional sobre Direitos Económicos, Sociais e Culturais de 16 de Dezembro de 1966, no artigo 11º como um direito ao declararem de certa forma que, "todos têm direito a um padrão de vida capaz de assegurar o bem estar, inclusive alimentação, vestuário, habitação, cuidados médicos e os serviços sociais indispensáveis, etc.," (ONU, 1948, pp. 12–13, 1966). Estas declarações pela sua abrangência têm reflexo na lei constitucional de vários países ao redor do mundo.

A casa constitui uma das necessidades mais essenciais da vida humana juntamente com outras necessidades fundamentais como a comida, água, segurança, etc., cuja procura (por casas) cresceu rapidamente ao longo dos anos à medida que os padrões de vida das pessoas melhoraram (Zulkifley et al., 2020). Enquanto algumas pessoas fazem de sua casa um investimento e propriedade, a maioria das pessoas ao redor do mundo compra uma casa como abrigo ou meio de subsistência (Zaman et al., 2021).

Um estudo realizado por Vale & Camões, (2018), sobre a avaliação da habitação, perceção de riqueza e composição do portfólio dos proprietários, consideram a habitação como, um ativo não líquido de risco médio com custos significativos de transação económica e que, as famílias fazem esforços financeiros com a finalidade de adquirir uma habitação e esta por sua vez, tende a se tornar o ativo mais importante como parte da riqueza destas famílias. Desta maneira, qualquer flutuação no preço da habitação afeta a riqueza destas famílias e de certa forma a economia.

A crise financeira (*subprime*) de 2008 com origem nos EUA (Estados Unidos da América) resultante da queda no preço da habitação, foi uma das evidências mais clara e notória de como o preço da habitação está fortemente relacionado com a economia, ao se propagar rapidamente transformando-se numa grande crise económica que afetou negativamente os mercados de ações não só dos Estados Unidos como também, a área do euro e as economias emergentes desde 2007 até ao final de 2010 (Blanchard et al., 2021).

Pelas pesquisas realizadas, vários estudos e organizações internacionais enfatizam a importância da habitação mostrando que ela está enraizada profundamente na estrutura económica, financeira e política de um país e que, flutuação no preço da habitação sempre constituiu uma preocupação para os agentes económicos (Rawool et al., 2021; Wang et al., 2014). Nos últimos anos tem-se assistido um aumento no preço da habitação em vários países influenciados por vários fatores (Kang et al., 2020).

Estes fatores, muitas vezes estão estritamente relacionados entre si assim como o mercado imobiliário está fortemente relacionado com a economia e, é considerado um dos mercados mais importante das economias desenvolvidas pelos níveis de investimentos e como fonte de garantia para os empréstimos, tornando-se um campo interessante de pesquisa em uma perspetiva macro e micro, ao relacionar-se intimamente com o sector financeiro e empresarial sendo uma parte importante da análise económica de qualquer país (Lagoa et al., 2004; Rącka & Khalil, 2018).

Pela importância da habitação para a economia, o preço da habitação relaciona-se com certas variáveis macroeconómicas cuja variação tende a influenciá-la positiva ou negativamente. Entre estas variáveis encontramos as principais como: a inflação, o ciclo económico, e a taxa de juro.

2.1.1. Preço da habitação e inflação

Um dos fatores macroeconómicos que afeta o preço da habitação é a inflação, comumente definida por muitos autores como um aumento do nível geral de preços em uma economia durante um certo período, e cuja variação difere de um país para outro ao longo do tempo (Dornbusch et al., 2013; Mankiw, 2015). Ela mede a variação do nível geral de preços permitindo que os formuladores de políticas monitorem a inflação e tomem medidas para controlá-la (Mendonça et al., 2021).

Um aumento na inflação tende a afetar o preço da habitação de várias formas, isto é, um aumento da inflação tende a provocar um aumento no preço da habitação reduzindo o poder de compra da moeda, e por outro lado, a inflação quando não controlada pode reduzir o valor do dinheiro a longo prazo tornando o investimento em propriedades (casa) como melhor alternativa em relação aos outros ativos (Lewandowska et al., 2023; Musarat et al., 2021).

Dado o impacto que as altas taxas de inflação têm sobre o preço da habitação e não só, torna-se vital um controlo sobre esta variável (inflação) visto que taxas altas e instáveis de inflação acarretam várias consequências negativas para a economia na medida em que, " maior incerteza na taxa de inflação, leva os mercados financeiros a cobrar prémios de riscos mais elevados, cria surpresas desagradáveis a quem faz depósitos a prazo, bem como torna menos eficaz a afetação de recursos na economia", (Leão et al., 2019, pp. 313–314). Estes eventos de certa forma acabam por afetar o preço da habitação em função da inter-relação existente entre os vários setores da economia.

A inflação e o desemprego geram preocupações macroeconómicas que muitas vezes, levam os formuladores de políticas económicas a tomarem decisões que representam um *trade-off* entre o desemprego e inflação num esforço para diminuir a inflação (Mankiw, 1998). Embora os custos da inflação sejam menos óbvios do que os do desemprego, a inflação cria distúrbios nas relações de preços (dentre os quais o da habitação) como conhecemos, e impacta negativamente a eficiência do sistema de preços da economia (Dornbusch et al., 2013; Mankiw, 1998).

2.1.2. Preço da habitação e ciclo económico

O crescimento do Produto Interno Bruto (PIB) tende a afetar o preço da habitação na medida em que um crescimento do PIB aumenta o nível da atividade económica geral, o nível de renda e do emprego que por sua vez, podem aumentar a procura por habitação para residência ou como uma forma de investimento sugerindo de certa forma uma relação positiva visto que, esse investimento em habitação é muitas vezes considerado como parte significativa do PIB em muitas economias e, portanto, é crucial para a atividade económica, (Kohler et al., 2023; Lewandowska et al., 2023).

Os preços das casas relacionam-se com o ciclo económico e podem segui-lo, considerando aqui o ciclo económico como, flutuações não regulares na atividade económica em geral em torno de uma tendência de longo prazo (Sachs & Larrain, 1995). O ciclo económico também é definido como, "flutuações periódicas da produção nacional em torno de sua tendência de longo prazo" (Sloman & Wride, 2009, p. 400). Durante expansões económicas, os preços das casas podem subir devido a uma procura crescente. Em contraste, durante recessões, os preços podem estagnar ou cair fruto da queda da produção nacional (Blanchard et al., 2021).

Estudos realizados por Case et al., (2005) e Mian et al., (2013) demostram que mudanças nos preços das casas tendem a impactar o consumo, investimento e decisões de empréstimo das famílias, bem como o seu comportamento financeiro em geral.

Desta forma a literatura revela que durante períodos de crescimento económico, os preços das casas tendem a subir, visto que o aumento do emprego e da renda geralmente impulsionam a procura por habitação afetando positivamente os preços da habitação. Por outro lado, os aumentos dos preços da habitação tendem a estimular os gastos dos consumidores e conduzem a um maior crescimento económico com o efeito riqueza, visto que as famílias também fazem da habitação uma forma de investimento (Pinheiro, 2012; Yi et al., 2022).

2.1.3. Preço da habitação e taxa de juro

O preço da habitação também está relacionado com a taxa de juro, considerada uma das mais importantes entre as principais variáveis macroeconómicas. Na condução da política monetária, através da gestão da taxa de juro, um Banco Central (BC) pode influenciar a quantidade de oferta e de procura de moeda a nível da economia, e com isso afetar as demais variáveis macroeconómicas que por sua vez também impactam o preço da habitação, (Leão et al., 2019; Mankiw, 2015).

Embora o investimento em habitação esteja sujeito aos ciclos económicos, estes ciclos por sua vez são influenciados por fatores como taxas de juro, condições do mercado imobiliário e políticas governamentais (Blanchard et al., 2021). Durante *booms* imobiliários, o investimento em habitação tende a aumentar e impulsiona a atividade económica. "A habitação pode ser usada como garantia para

novos empréstimos, e as flutuações nos preços das casas afetam simultaneamente a capacidade de empréstimo das famílias e o retorno da produção de novas casas (Camões & Vale, 2018, p. 1956)".

As políticas de um Banco Central (BC), incluindo mudanças nas taxas de juros, podem ter um impacto significativo no preço das casas. Políticas expansionistas podem impulsionar os preços, enquanto políticas restritivas podem ter o efeito oposto (Snyder & Vale, 2022). Através dos instrumentos à sua disposição tais como: operações de mercado aberto, política de desconto ou de refinanciamento, e reservas mínimas obrigatórias, o Banco Central influência as demais taxas de juro da economia através do sistema bancário, e isto, impacta o crédito bancário, muitas vezes utilizado pelas famílias como a melhor alternativa para a aquisição de habitação (Abreu et al., 2018; Lagoa et al., 2004).

O crédito habitação assim como os demais créditos bancários, são muitas vezes influenciados pela taxa de juro controlada por um BC na economia de um dado país ou grupo de países. Estudos realizados por Favilukis et al., (2017) e Khandani et al., (2013), sobre a dinâmica dos preços da habitação nos Estados Unidos da América de forma a controlar as bolhas especulativas no mercado imobiliário, revelaram que as quedas na taxa de juro, e o relaxamento das restrições de crédito, contribuíram significativamente para o *boom* imobiliário e a valorização dos preços da habitação.

Desde essa época, o setor imobiliário tem ganhado bastante atenção das entidades governamentais e de vários académicos, de modo a evitar-se novamente bolhas especulativas no mercado imobiliário (Eldionara et al., 2014; Pinheiro, 2012; Wu & Lux, 2018).

Podemos entender que o mercado imobiliário tem grande importância para uma economia e que, quando não é controlado, "pode perturbar o desempenho econômico, desde moldar e amplificar os ciclos económicos até alterar o mecanismo de transmissão monetária. Aumentos nos preços das casas induzem efeitos de riqueza nos proprietários e expandem o ciclo económico, acentuando seus picos e recessões" (Snyder & Vale, 2022, p. 244).

Desta forma, para garantir a estabilidade do preço da habitação, tendo em conta as flutuações que ocorrem com os ciclos económicos, alterações da taxa de juro, etc, é necessário que a política económica em geral e a política monetária em particular desempenhem um papel determinante para a estabilização do sistema, por forma a evitar bolhas especulativas, visto que a estabilização e amortecimento dos ciclos económicos têm importância fundamental para um funcionamento regular da economia (Abreu et al., 2018; Blanchard et al., 2021).

2.2. Fatores específicos que influenciam o preço da habitação

Existem inúmeros fatores e atributos específicos que podem influenciar significativamente o preço da habitação, e que podem variar de um país para outro. Tendo em conta a heterogeneidade associada aos custos e preços da habitação, embora seja considerado como um dos bens de primeira necessidade, a sua comparação com os outros bens de primeira necessidade torna-se uma tarefa complexa em função de outros fatores específicos tais como: a localização, o tipo de habitação, a estrutura da casa, (tamanho em metros quadrado, número de quartos, etc.) (Teixeira et al., 2010; Thamarai & Malarvizhi, 2020).

Nos estudos realizados por Zulkifley et al., (2020), e Rawool et al., (2021) sobre a previsão do preço da habitação, os autores identificaram três principais atributos que geralmente influenciam o preço da habitação, a saber: a localização, considerado como o primeiro e principal atributo, isto é, leva em conta a proximidade do imóvel em relação aos pontos de transportes, escolas, estação de metro, vista para o mar, etc.; a estrutura, relacionado com o tamanho da casa medido em metros quadrados, número de quartos, tipo de habitação, garagem, ano de construção, etc.; e, por último, mas não menos importante, o bairro, levando em conta aspetos como a existência de pontos de transportes públicos, a segurança pública, a educação dos moradores, taxa de crimes, etc.

Outros autores como Kang et al., (2020) e Adetunji et al., (2022) defendem que só os atributos acima mencionados não sejam suficientes para explicarem o preço da habitação, deve-se levar também em conta aspetos que contribuam para a saúde física e mental das pessoas, bem como verificar o potencial de crescimento que tende a atrair mais pessoas e investimento nestas localizações. Adicionalmente, identificam a localização como o atributo que mais significativamente influência o preço da habitação.

Através destes atributos, podemos utilizar modelos de ML para fazer previsão do preço da habitação, visto que, prever o preço futuro da habitação é extremamente importante tendo em conta a forte relação que tem com a economia e a importância da habitação para o bem-estar de uma sociedade, estando a habitação fortemente relacionada com cada um de nós (Wang et al., 2014).

Estudos realizados, como o de Bian et al., (2019), sobre o mercado imobiliário em Singapura, também revelam que os preços da habitação são influenciados pelos atributos acima mencionados, isto é, a proximidade dos *Mass Rapid Transport*, (MRP) bem como pontos de autocarros e o centro da cidade.

2.3. Resumo do capítulo

A habitação é de extrema importância para a sociedade e o preço da habitação está fortemente relacionado com as principais variáveis macroeconómicas como a inflação, ciclo económico e taxa de juro, demonstrando ser indispensável o monitoramento do mercado imobiliário a nível da economia, fato que se tornou mais evidente principalmente após a crise *subprime* nos EUA.

Assim como o equilíbrio da economia é comumente perturbado por choques aleatórios, e estes choques por sua vez podem ser temporários ou permanentes, antecipados ou não antecipados (Sachs & Larrain B., 1995; Wickens, 2011), também os mercados, dentre os quais, o mercado imobiliário, em função da inter-relação existente, é constantemente afetado por choques, razão pela qual se torna fundamental analisar e prever o seu comportamento, de modo a garantir o seu equilíbrio e contribuir para o funcionamento regular da economia.

"Um dos objetivos da economia é a capacidade de explicar e prever o comportamento do mercado, (Dornbusch et al., 2013, p. 445)", tendo em conta os registos (dados) passados e outras informações disponíveis sobre um determinado mercado situado em uma região ou país, etc. (Cesa-Bianchi & Lugosi, 2006). Desta forma Singapura pelo seu histórico sobre a habitação e as características do mercado imobiliário, representa a nossa unidade de observação para este trabalho.

O setor imobiliário representa um dos mercados mais importantes da economia cuja previsão a partir de grandes volumes de dados, leva os agentes económicos a obterem informações relevantes sobre o preço da habitação antes de tomarem qualquer decisão. Analisar e entender o mercado imobiliário tem importância fundamental para a economia visto que há uma forte relação entre o preço da habitação e a economia.

CAPÍTULO III

Aspetos Metodológicos

Existem várias formas de fazer previsões sobre o preço da habitação utilizando diferentes metodologias, desde os métodos tradicionais baseados em análise de regressão, precificação hedónica e os vários modelos de ML (Zulkifley, et al., 2020). Para a nossa investigação, usaremos modelos de *Machine Learning*, uma abordagem amplamente utilizada atualmente para a criação de soluções envolvendo grandes volumes de "dados".

3.1. Inteligência Artificial (IA) e Machine Learning (ML)

É difícil falar de *Machine Learning* (ML) sem fazer menção da Inteligência Artificial (IA), cuja ideia remonta do passado através de Alan Turing que em 1950 acreditava que, através da aprendizagem era possível elevar um sistema simples a um nível de inteligência humana, conceito a que se chamou de "máquina-criança" (Bostrom, 2014, p. 49).

Atualmente, com os avanços da tecnologia, há uma forte tendência na descontinuidade dos métodos tradicionais de previsão, dando lugar ao uso crescente da *Inteligência Artificial* onde se emprega *Machine Learning* em vários setores da economia para se criar modelos preditivos com elevado nível de acurácia, levando diversas empresas e indústrias a caminharem rumo para a automação (Obschonka & Audretsch, 2020; Rawool et al., 2021).

Segundo Simons, (1986, p. 54), Inteligência Artificial define-se como, "a ciência de fazer com que máquinas façam coisas que requereriam inteligência se feitas pelos humanos". Outros autores definem IA como, "uma disciplina que tem por objetivo o estudo e construção de entidades artificiais com capacidades cognitivas semelhantes às dos seres humanos (Costa & Simões, 2004, p. 3)".

A IA é um campo de pesquisa onde se estuda e analisa a forma como agentes computacionais com objetivos e limitações, aprendem com as experiências e agem de forma inteligente, ou seja, executam tarefas que geralmente são vistas como exigindo inteligência (Poole & Mackworth, 2010). Com a IA percebe-se um aumento significativo das capacidades não mecânicas do ser humano em resolver problemas (de elevada complexidade), embora haja uma resistência por parte da sociedade em aceitar a possibilidade de que as máquinas tenham esta capacidade (Costa & Simões, 2004).

Já *Machine Learning* é uma área de estudo que, através de um conjunto de instruções, designadas por algoritmos, dá aos computadores a capacidade de aprender sem que sejam explicitamente programados para este efeito (Domingos, 2017; Samuel, 1959). Isto é, a capacidade dos computadores de adquirirem conhecimento e melhorarem o seu desempenho através de experiências e do acesso a dados, em vez de dependerem de instruções programadas simplesmente por um ser humano (Flach, 2012).

Machine Learning, é uma subárea da IA com o foco no desenvolvimento de algoritmos e modelos que permitem aos computadores aprender a partir de dados históricos, baseia-se na ideia de que sistemas podem aprender com dados, identificar padrões e fazer decisões com o mínimo de intervenção humana (Marsland, 2009).

Destas abordagens, podemos entender que a Inteligência Artificial (IA) é um campo amplo que busca criar sistemas que exibam comportamento inteligente, enquanto o *Machine Learning* (ML) é uma abordagem específica dentro da IA que se concentra no desenvolvimento de algoritmos eficientes utilizados para criar modelos, fazer análise e previsão, e constitui uma das técnicas mais proeminentes usadas para alcançar a IA (Theodoridis, 2015).

Consideramos neste contexto um algoritmo como, "uma sequência de instruções que diz a um computador o que fazer (Domingos, 2017, p. 25)". Outra definição não muito diferente da anterior considera que, "um algoritmo é qualquer procedimento computacional bem definido que recebe algum valor, ou conjunto de valores, como entrada e produz algum valor, ou conjunto de valores, como saída. Também podemos ver um algoritmo como uma ferramenta para resolver um problema computacional bem especificado (Cormen et al., 2009, p. 5)".

Com base nestas definições podemos entender que um algoritmo, é uma sequência de passos (com um início e um fim) dado de forma lógica e sequencial com a finalidade de resolver um problema específico. Atualmente existem vários algoritmos de ML utilizados para resolver diversos problemas reais.

3.1.1. Tipos de algoritmos de *Machine Learning*

Vários algoritmos de *Machine Learning* tais como, *Árvores de Decisão* (AD), SVM (*Suporte Vector Machine*), *Random Forest* (RF), Redes Neuronais Artificiais (RNA), entre outros, estão disponíveis para serem usados no setor imobiliário, assim como para fazer a previsão do preço da habitação com elevada *performance*, muitas vezes não alcançados através dos métodos tradicionais de previsão (Wang et al., 2014; Zulkifley et al., 2020). A seguir, resumidamente descrevemos os algoritmos de ML a serem utilizados neste trabalho, tendo em conta o nosso objetivo e as características dos dados existentes.

3.1.1.1. Regressão Linear Multipla (RLM)

O modelo de regressão é um dos modelos mais simples e muito utilizado para determinar a relação entre duas ou mais variáveis, isto é, regressão linear simples e regressão linear múltipla ou multivariada (Thamarai & Malarvizhi, 2020). Considerado como um dos algoritmos de ML mais simples, através do modelo de regressão linear também podemos fazer previsões de preços específicos e verificar quais atributos ou variáveis explicativas são mais relevantes para explicar a variável dependente. Na realidade,

problemas desafiantes em análise preditiva com o modelo de regressão linear são mais de natureza multivariável (Kelleher et al., 2015).

3.1.1.2. *Random Forest* (RF)

Random Forest é um tipo de algoritmo que combina os resultados de várias árvores de decisão para depois apresentar como resultado uma única árvore com maior precisão (Quang et al., 2020). Geralmente cada árvore de decisão é constituída por vários nós que representam atributos a serem classificados ou previstos em caso de regressão e os ramos que representam os valores dos atributos (Mahesh, 2018).

Estudos realizados por Rawool et al., (2021) para prever o preço da habitação revelaram um elevado grau de precisão apresentado pelo modelo ao usarem o algoritmo *Random Forest*, em comparação com outros algoritmos de ML como *Decision Tree*, KNN (*K-Nearest Neighbors*) e Regressão Linear.

Todavia, essa engenhosidade da IA e ML que deixa o mundo maravilhado pelos seus resultados, só é possível através de outro grande requisito designado por "dados" que, gerados em grande volume, velocidade e variedade dão origem ao *Big Data* considerado por muitos como o novo petróleo, o cerne e o motor da aprendizagem de máquina visto que ML baseia-se em modelos treinados a partir de dados históricos para fazer previsões futuras (Mahesh, 2018; Rawool et al., 2021).

3.2. *Big Data*

Quando falamos de aplicação de *Machine Learning* em qualquer área, isso nos remete de certa forma para a utilização de *Big Data*, cuja realidade não podemos ignorar nos dias atuais quando observamos a quantidade de dados gerados em grande volume e alta velocidade, tornando-se em um ambiente importante a ser utilizado pelos académicos, indústrias e governos sempre que é preciso tomar decisões (Shi, 2022).

Atualmente, as diversas áreas desde a ciência, engenharia, economia, negócios e finanças produzem e processam quantidades extraordinárias de dados caracterizando o que hoje chamamos de *Big Data*, (Li et al., 2015; Prüfer & Schottmüller, 2021).

De acordo com um estudo realizado pelo IDC (*International Data Corporation*) em 2011, já se estimava que a quantidade de dados disponíveis no mundo duplicava a cada dois anos (Cui et al., 2016). Li et al., (2015) afirmavam que na era digital em que nos encontramos, são gerados aproximadamente cerca de 2,5 quintilhões de dados todos os dias. Conjugando estes dois indicadores, o tempo em que estes autores realizaram os estudos e o momento em que nos encontramos, podemos ter uma noção da quantidade de dados existente no mundo.

Destes conjuntos de dados quando recolhidos, processados e analisados, resultam as principais decisões informadas que transformam de forma positiva a realidade socioeconómica, visto que as decisões conduzidas por dados ajudam a ter uma visão sobre o futuro. De acordo com estas abordagens

e pelas evidências da literatura, torna-se cada vez mais notório a relevância dos dados para a economia. Silva (2023), durante uma conferência sobre a Antevisão da Economia e Política 2024 sobre Portugal, afirmava que, "os dados representam para as economias no século XXI, aquilo que a terra foi no passado para a agricultura no crescimento das economias".

3.3. Métodos e técnicas

Para a nossa investigação usaremos uma abordagem baseada em *Machine Learning* que, (ML) inclui duas sub-áreas fundamentais, nomeadamente a aprendizagem supervisionada e não-supervisionada, utilizando vários algoritmos de aprendizagem de máquina para treinar modelos, fazer previsões ou tomar decisões baseadas nos dados (Mahesh, 2018; Marsland, 2009).

De acordo com Mahesh (2018), a aprendizagem supervisionada é uma tarefa de aprendizagem que leva uma máquina a aprender uma função que mapeia uma entrada para uma saída com base em pares de entrada e saída designadas como dados de exemplos. E a aprendizagem não supervisionada como o nome indica, refere-se à tarefa através da qual os algoritmos recebem apenas os dados de entrada sem os dados de saída, descobrem por conta própria padrões ocultos e apresentam como resultado estruturas interessante sobre os dados de entrada em forma de classes ou agrupamentos. Assim, quando novos dados são introduzidos, o algoritmo saberá onde agrupá-los de acordo com as suas características.

Geralmente este conjunto de dados de entrada é dividido em dados de treino e dados de teste em que, em dados de treino o algoritmo recebe as variáveis de entrada e saída a serem previstas ou classificadas, e a partir deste conjunto de dados de treino o algoritmo aprende padrões ocultos e aplica estes padrões aprendido ao conjunto de dados de teste, isto é, apenas as entradas ou variáveis explicativas para prever ou classificar a variável de saída também designada como variável alvo (Marsland, 2009).

Atualmente, utilizar algoritmos de ML para tarefas de previsão é uma forte tendência e apresenta grandes vantagens em relação aos métodos tradicionais (Obschonka & Audretsch, 2020), visto que os modelos de ML tendem apresentar maior capacidade de capturar relações não-lineares entre as variáveis e são mais eficazes em grandes conjuntos de dados (Akansu et al., 2016; Zulkifley et al., 2020).

Com ML podemos automatizar a seleção de variáveis mais relevantes para o modelo, maior flexibilidade em diferentes tipos de dados, não dependem de suposições estatísticas rígidas como os modelos tradicionais e têm maior capacidade de generalização ao apresentarem previsões com elevada precisão (Mei & Shi, 2024; Shafiee et al., 2021)

Assim, utilizamos dentre os diversos métodos e técnicas de ML, a aprendizagem supervisionada e os algoritmos de regressão LASSO (Lee et al., 2022) e o *Random Forest* (RF). O primeiro algoritmo (LASSO) será usado para a seleção de atributos e também para criar um dos modelos preditivos, e o segundo algoritmo, *Random Forest* (RF), para construção do outro modelo preditivo, e será feito a seguir a avaliação e seleção do melhor modelo, isto é, o modelo final capaz de prever o preço da habitação com a maior precisão.

Após a recolha dos dados, para melhor analisá-los e obter informações relevantes sobre os mesmos (dados), utilizamos ferramentas como: o software estatístico R, e a linguagem de programação Python, a partir das quais podemos calcular diversas estatísticas e construir gráficos. São ferramentas *open source* e oferecem inúmeros pacotes e funções para se trabalhar em análise, exploração de dados e construir modelos preditivos baseados em ML (Foundation, 2001; Foundation, 2024).

3.4. Fonte dos dados

Os dados utilizados nesta pesquisa foram extraídos da fonte oficial do Governo de Singapura, que disponibiliza diversos volumes de dados sobre o país e dentre os quais, os dados de transações mensais sobre as vendas de habitações realizadas pelo *Housing & Development Board* (HDB), que é a autoridade de habitação pública de Singapura (HDB, 2022).

Para o efeito, foram extraídos um total de 915.374 registos ou observações, considerando o período entre o mês de Janeiro do ano 1990 a Dezembro de 2023, pois acreditamos que, "quanto mais distante olhamos para o passado, maior será a nossa capacidade de entender o presente, e de olhar e prever o futuro (Silva, 2023)".

3.4.1. Descrição do conjunto de dados

Este conjunto de dados, contém como mencionado anteriormente, 915.374 registos ou observações e um total de 10 variáveis ou atributos, onde o preço de revenda ou preço de venda como chamaremos mais adiante para melhor contextualização é a variável dependente (alvo) e os demais atributos constituem as variáveis independentes ou explicativas. Desta forma, antes de prosseguirmos com a nossa análise, apresentamos abaixo um quadro resumo com a explicação das variáveis (Tabela 1).

Tabela 1 - Resumo descritivo das variáveis.

Atributo / Coluna	Tipo de dados	Descrição
Mês	Data	Mês e o ano em que foi realizada a transação de venda
Cidade	Texto	Refere-se à cidade ou região em Singapura onde a habitação está localizada.
Quartos	Texto	Número de quartos do imóvel
Bloco	Texto	Refere-se ao número do bloco de apartamentos dentro de um determinado complexo habitacional ou conjunto residencial.
Nome da rua	Texto	Representa o nome da rua ou avenida onde a habitação está localizada.
Lote	Text	Indica o intervalo de andares onde a unidade habitacional está localizada.
Area em m^2	Numérico	Está relacionado com a área total do imóvel, geralmente em metros quadrados (m^2).
Modelo	Texto	O modelo ou tipo de layout do imóvel de acordo com as tipologias de casa em Singapura
Ano de construção	Numérico	Data de início do contrato de arrendamento do apartamento, geralmente também relacionada com o ano de construção do imóvel.
Preço de venda	Numérico	O preço da transação ou o valor da habitação em termos monetários, expressos em dolar de Singapura. (Considerada como a variável dependente para o nosso estudo).

Fonte: HDB

3.5. Fases para a construção do modelo de *Machine Learning*

Após a identificação da fonte dos dados a serem utilizados, a seguir descrevemos de forma detalhada os passos que foram dados para a construção do modelo de ML, desde a recolha dos dados, análise exploratória dos dados, pré-processamento dos dados, construção do modelo, avaliação, e seleção do modelo (Quang et al., 2020; Rawool et al., 2021).

De acordo com Kelleher et al. (2015), a criação de modelos preditivos envolve muito mais do que escolher um algoritmo de ML e implementá-lo, pois, para que seja bem-sucedido é necessário que se siga um processo padrão baseado nos seguintes passos: Compreensão do problema a ser resolvido, compreensão e recolha dos dados, preparação dos dados, modelação dos dados a partir dos quais diferentes algoritmos são usados para construir modelos de previsão, e o melhor modelo é selecionado, vindo a seguir a avaliação e implementação do modelo.

3.5.1. Análise exploratória dos dados

Uma análise exploratória de dados envolve técnicas de visualização e estatística de forma a melhor descrevê-los e resumi-los, e ajuda-nos a verificar se as características e relações esperadas existem nos dados, bem como ver se existe alguma estrutura inesperada que deve ser levada em conta. Com esta análise podemos detetar falhas no conjunto de dados que devem ser corrigidas caso existam, selecionar uma análise apropriada, e validar as técnicas apropriadas a serem usadas nas etapas seguinte do trabalho, principalmente quando lidamos com dados do tipo multivariado (Siegel, 2016).

De acordo com as características dos dados, na sequência foram dados os seguintes passos com o software R:

Através da função *separate*, dividimos a coluna *Mês* do *dataset* original em duas e obtemos como resultado as colunas *Ano* e *Mês* em que foram realizadas as transações de modo a obter mais informações relevantes a partir dos dados que compõem estas colunas. A seguir, de forma a contextualizar, renomeamos as colunas do conjunto de dados para português obtendo como resultado o conjunto de dados como podemos visualizar os 10 primeiros registos na tabela em Anexo A1.

De seguida fez-se a verificação dos tipos de dados, valores duplicados e existência de valores omissos (NA – *Not Available*) em cada coluna do conjunto de dados. Como resultado verificamos a existência de valores omissos (NA) na coluna "Número de quartos" num total de 69.467 conforme se vê na tabela em Anexo A2. Embora seja um número reduzido comparado com o total de observações (915.374) do conjunto de dados, precisam ser tratados antes de prosseguirmos com a nossa análise.

Para não os remover e correr o risco de perder informações relevantes, usamos a técnica de imputação, isto é, preenchemos os dados em falta (NA) com um dos valores padrão como a mediana da própria variável no valor de 4 (Gama et al., 2012; Moturi, 2020), conforme o resultado do resumo estatístico da mesma variável na Tabela 2 abaixo, visto que este valor (4) se aproxima da média (3,85) de acordo com o cálculo feito e a variável (*numero de quarto*) é do tipo de dado inteiro.

Tabela 2 - Resultado do resumo estatístico da variável número de quarto.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	NA´s
1.00	3.00	4.00	3.85	4.00	5.00	69467.00

Fonte: HDB (cálculo realizado com o software R)

Após esta verificação, de forma a facilitar uma análise detalhada dos dados, separamos as variáveis numéricas das categóricas e resultou assim em dois subconjuntos (*subset*) de dados. Para a análise apresentada a seguir, utilizamos a linguagem de programação Python, com os seus poderosos pacotes como *pandas, numpy, matplotlib, seaborn, sklearn*, entre outros, que pela simplicidade na escrita de códigos oferecem vários recursos para criar visualizações que nos ajudam a compreender melhor os dados e permitem calcular estatísticas relevantes sobre os dados. A Tabela 3 abaixo ilustra essa divisão.

Tabela 3 - Divisão do conjunto de dados em numéricos e categóricos

Descrição	Tipo de dados
Ano	Numérica
Numero de quarto	Numérica
Tamanho em m^2	Numérica
Ano de construção	Numérica
Preço de venda	Numérica
Mês	Categórica
Cidade	Categórica
Bloco	Categórica
Lote/andar	Categórica
Modelo	Categórica
Nome da rua	Categórica

3.5.1.1. Análise de dados numéricos

Para esta categoria de dados o nosso foco foi primeiramente fazer um resumo estatístico das variáveis (Tabela 4) e a construção do mapa de correlação de Pearson mais adiante, de forma a obtermos informações estatisticamente relevantes e visualizações que nos permitam rapidamente compreender as características básicas de cada variável, e estabelecer uma base sólida antes de explorar relações mais complexas entre múltiplas variáveis deste subconjunto de dados e orientar a análise subsequente.

Através da análise univariada procuramos descrever e resumir as variáveis ao utilizarmos medidas de tendência central, dispersão e a forma da distribuição, por meio dos principais aspetos e técnicas de análise univariada. Com o resultado do resumo estatístico, podemos verificar que as propriedades vendidas têm em média cerca de 4 quartos, um tamanho total médio de 95,70 m², e a maior parte foram construídas em torno de 1988, sendo a construção mais antiga de 1966 e a mais recente de 2022. As vendas ocorreram de 1990 a 2023 num total de 915.374 imoveis.

Tabela 4 - Resultados do resumo estatístico das variáveis numéricas

	Ano	Numero_quarto	Tamanho_m2	Ano_construcao	Preco_venda
count	915374.00	915374.00	915374.00	915374.00	915374.00
mean	2005.94	3.86	95.70	1988.15	317237.29
std	9.16	0.76	25.85	10.53	167403.28
min	1990.00	1.00	28.00	1966.00	5000.00
25%	1998.00	3.00	73.00	1981.00	192000.00
50%	2005.00	4.00	93.00	1986.00	295000.00
75%	2013.00	4.00	113.00	1996.00	413000.00
max	2023.00	5.00	307.00	2022.00	1500000.00

Os resultados também mostram que os preços de vendas para habitação em Singapura sofreram variações significativas desde 1990 a 2023, oscilando entre um mínimo de 5.000,00 e um máximo de 1.500.000,00 expressos em dólares de Singapura conforme também é evidenciado pela Figura 1 abaixo, um gráfico que ilustra a evolução do preço de venda da habitação durante este período. Os dados (do resumo estatístico) também indicam que, 75% dos imóveis foram vendidos abaixo dos 413.000,00, 25% abaixo dos 192.000,00 e um desvio padrão de 167.403,28 que demostram variações consideráveis no preço da habitação durante este período.

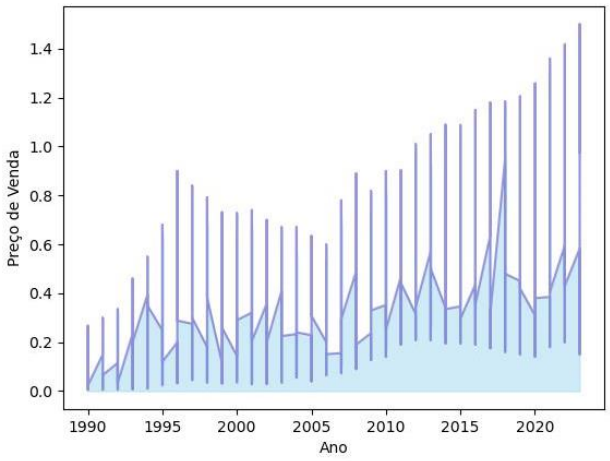


Figura 1 - Evolução do preço de venda da habitação em Singapura desde janeiro de 1990 a Dezembro de 2023.

Na sequência através de visualização gráfica dos *boxplot* (das variáveis) principalmente para a variável dependente (preço de venda) de acordo com a Figura 2 abaixo, também podemos identificar uma possível presença de *outliers*, valores extremos atípicos, que parecem não pertencer aos outros dados porque são muito grandes ou muito pequenos (Siegel, 2016). Para removê-los nas etapas seguintes, usaremos o critério comum que é considerar *outliers* aqueles pontos de dados que estão mais de três desvios padrão de distância (acima ou abaixo) da média.

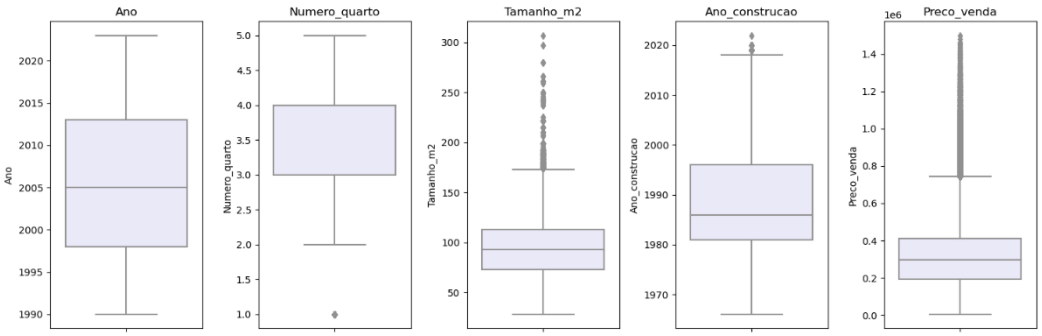


Figura 2 - Resultados da análise univariada das variáveis numéricas, identificação de outliers.

Como boa parte de muitos algoritmos de ML têm como pressuposto de que os dados (a variável dependente, preço de venda) seguem uma distribuição normal, fez-se um teste de normalidade (H_o) de Shapiro-Wilk e o resultado de acordo com o valor do p-value de 2.2e-16 conforme o Anexo A3, revela um valor menor que o nível de significância usual de 0,05 e sugere que temos evidências suficientes para rejeitar a hipótese nula de que os dados vêm de uma distribuição normal, ou seja, os dados não seguem uma distribuição normal (H_1), e para normalizá-lo aplicaremos o logaritmo (Cortinhas & Black, 2012).

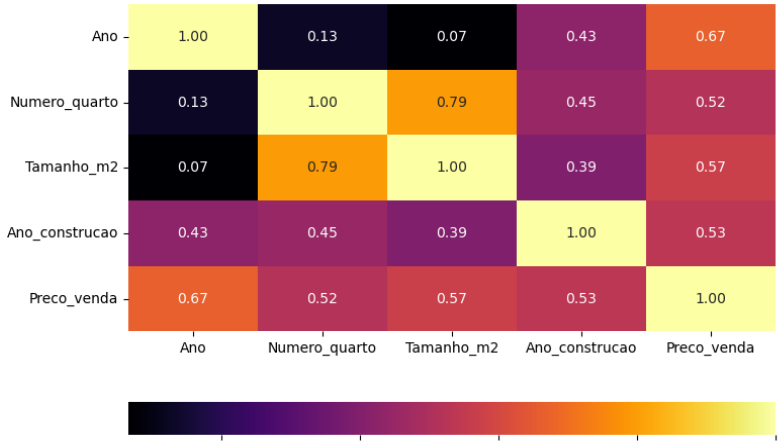


Figura 3 - Resultados do mapa de correlação.

Quanto ao mapa de correlação de Pearson, Figura 3 acima, serviu para medir a intensidade e a direção da relação linear entre as variáveis (Moturi, 2020; Siegel, 2016). O mapa de correlação (Figura 3) indica que existe uma forte correlação positiva entre a variável alvo ou *target* (preço de venda) e as

variáveis explicativas como: o ano de transação e de construção, tamanho em metros quadrado, número de quartos, confirmando o que a literatura diz sobre o impacto destas variáveis no preço de venda da habitação conforme mencionamos no capítulo anterior (Teixeira et al., 2010; Thamarai & Malarvizhi, 2020).

Outro aspeto que chama a nossa atenção é a relação entre algumas variáveis explicativas com um grau de correlação média acima dos 0,60, isto é, as variáveis - tamanho em metros quadrado e o número de quartos - com uma correlação forte positiva no valor de 0,79, indica um sinal de multicolinearidade que pode representar um problema para o modelo ao reforçá-lo com duas variáveis com o mesmo tipo de informação.

As demais variáveis demonstram não haver correlação significativa entre si e outras têm uma relação positiva moderada como podemos observar no mapa acima e no Anexo A4.

3.5.1.2. Análise de variáveis categóricas

Para os dados categóricos o nosso foco de análise foi principalmente obter algumas estatísticas relevantes que nos ajudam a compreender melhor os dados e certos aspetos da realidade habitacional em Singapura. Desta forma ao calcularmos o resumo estatístico destas variáveis e fazer consultas ao subconjunto de dados através do pacote *pandas*, *seaborn e matplotlib* do Python, foi possível obter informações sobre as vendas de habitação em Singapura durante o período em análise (Foundation, 2001).

Tabela 5 - Resultado do resumo estatístico das variáveis categóricas.

	Cidade	Mes	Bloco	Nome_da_rua	Lote_andar	Modelo
contagem	915374	915374	915374	915374	915374	915374
Valores Únicos	27	12	2699	584	25	21
Moda	TAMPINES	7	2	YISHUN RING RD	04 TO 06	MODEL A
Frequência	79133	82051	4563	17454	229876	260408

Fonte: HDB

O resumo estatístico de variáveis categóricas conforme a Tabela 5 mostra que, as vendas ocorreram em 27 cidades, tendo a maior parte ocorrido na cidade de Tampines, foram comercializados 21 modelos de apartamentos e o modelo mais vendido durante este período foi o modelo A. Os gráficos de mapas (coropléticos) na Figura 4 abaixo fornecem mais informação relacionada com o número de habitações vendidas por cidades.

O maior número de apartamentos foi vendido nas cidades de Tampines, Yishun, Bedok, Jurong West, Woodlands, Ang Mo Kio, Hougang, de acordo com os mapas abaixo (Figura 4) a seguir, e a

menor quantidade vendida foi verificada na cidade de Lin Chu Kang no total de 64 apartamentos. Entre o período de 1990 a 2023 foram vendidos em média cerca de 26.922 apartamentos por ano e a maior quantidade vendida foi em 1999, conforme a figura em Anexo A5.

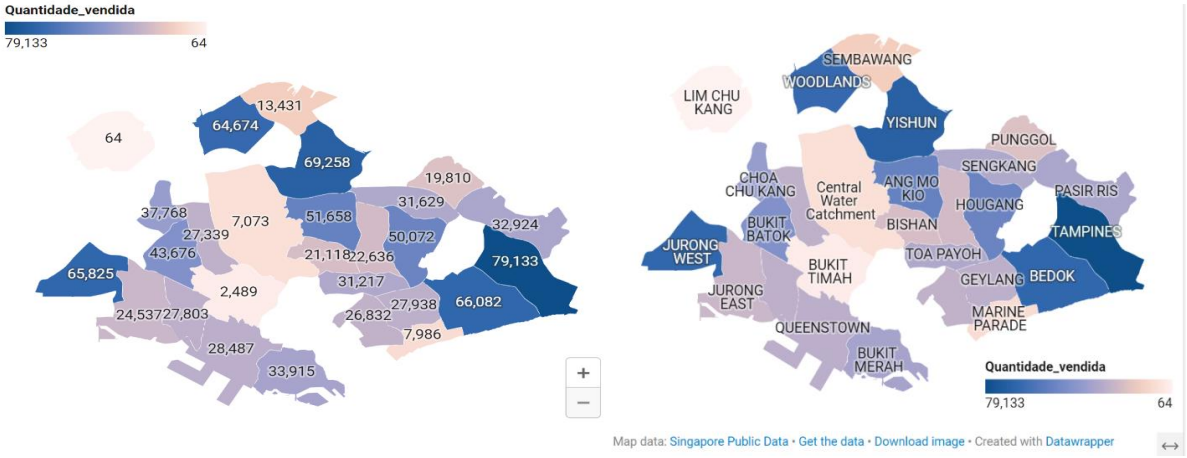


Figura 4 - Mapa de distribuição do número de apartamentos vendidos por cidade no período compreendido entre Janeiro de 1990 a Dezembro de 2023. Do lado esquerdo, o mapa representa as quantidades em cada cidade e no lado direito o mesmo mapa com o nome das cidades.

Pela literatura sabemos que as variáveis categóricas que compõem este subconjunto de dados têm impacto na variação do preço da habitação, isto é, que existe uma relação entre o preço de venda da habitação e as variáveis deste subconjunto de dados de acordo com estudos realizados por Zulkifley et al., (2020), e Rawool et al., (2021). Para perceber melhor esta relação, a nossa análise, na etapa a seguir relacionada com a seleção de atributos, utilizamos o software R para transformar os tipos de dados das variáveis categóricas para o tipo fator, ou seja, no formato numérico, sem alterar a informação original.

A razão para isso e não menos importante senão a principal, é que na aprendizagem de máquina os algoritmos de ML esperam receber os dados no formato numérico para a construção do modelo de aprendizagem (Ozsahin et al., 2022). Esta é na verdade uma das etapas mais importantes e existem várias técnicas como *One-Hot Encoding* e *LabelEncoder* do *sklearn* em Python e outras do pacote *base* R como a função *factor* que nos ajudam para isso, sendo esta última (R) a nossa escolha para esta tarefa de transformação das variáveis categóricas por causa das inúmeras funções e pacotes estatísticos que oferece.

No entanto, antes de realizar esta transformação no pré-processamento dos dados onde faremos a seleção de atributos, vamos remover os *outliers* do nosso conjunto de dados como se viu no *boxplot* principalmente pelas variáveis tamanho em metros quadrado e preço de venda na Figura 2, pois esta foi uma das principais anomalias encontrada até aqui sobre os dados.

Para isso tal como mencionamos anteriormente, usaremos o critério comum que é considerar *outliers* aqueles pontos de dados que estão mais de três desvios padrão de distância (acima ou abaixo) da média (Cortinhas & Black, 2012; Siegel, 2016), critério este conhecido comumente como o "método

do *z-score*" e usamos a linguagem Python para esta operação. Como resultado foram detetados 10.322 registos considerados *outliers* conforme o Anexo A6.

3.5.2. Pré-processamento dos dados

A etapa do pré-processamento é de extrema importância na metodologia adotada para o nosso trabalho, pois envolve a preparação e transformação de dados brutos em um formato adequado para análise e modelação, com o objetivo principal de melhorar a qualidade dos dados para que se crie um modelo de aprendizagem de máquina capaz de fazer previsões precisas e confiáveis (Zaman et al., 2021; Rawool et al., 2021).

Desta forma, para esta etapa, após a remoção dos valores *outliers* (que perfazem 10.322 observações), o conjunto de dados agora tem um total de 905.052 observações e 11 variáveis ou atributos nos quais aplicamos a transformação das variáveis categóricas para o tipo fator como referido anteriormente. Como resultado, temos agora todos os dados no formato numérico como se observa em Anexo A7, o resultado de uma amostra extraída do conjunto de dados.

3.5.2.1. Padronização dos dados

A maior parte dos algoritmos de *Machine Learning* além de receberem os dados no formato numérico, esperam que os dados estejam na mesma escala, ou num intervalo específico geralmente considerado entre 0 e 1 ou -1 e 1, técnicas conhecida como padronização e normalização (Ozsahin et al., 2022), aplicada aos dados com o objetivo de ajustá-los para que diferentes características tenham uma mesma escala numérica e comparável, devido à sensibilidade à escala dos dados demostrada por alguns algoritmos de aprendizagem de máquina (Gama et al., 2012).

A técnica a aplicar geralmente depende do tipo de dados e a natureza do problema que se pretende resolver, e para o nosso caso, tendo em conta as diferentes escalas nos valores do conjunto de dados, usamos a técnica de padronização cuja formula (1) apresentamos abaixo, que consiste em transformar os dados para que os atributos ou recursos caiam num intervalo onde tenham a média como 0 (zero) e o desvio padrão de 1 (Gama et al., 2012; Moturi, 2020).

$$x_p = \frac{x - \mu}{\sigma} \quad (1)$$

Em que:

x_p: é o valor padronizado da variável x;

x: representa cada valor original da variável que está a ser padronizada.

μ: é a média da variável x.

σ: é o desvio padrão da variável x.

Durante a padronização a subtração da média (μ) centraliza os valores em torno de 0, o que faz com que a nova média seja 0, e a divisão pelo desvio padrão (σ) nesta operação, ajusta a escala dos dados para que tenham um desvio padrão de 1, o que significa que os valores padronizados terão a mesma unidade de medida independentemente da escala original dos dados de formas a facilitar, a comparação entre diferentes variáveis e a aplicação em diversos modelos de aprendizagem de máquina (Ahsan et al., 2021; Thara et al., 2019). Para isso usamos a função *StandardScaler* do pacote *sklearn* em Python aplicando-o nas variáveis explicativas a serem usadas para a construção do modelo como se vê em Anexo A8 o resultado desta transformação (Foundation, 2001).

3.5.2.2. Seleção de atributos ou variáveis

Depois de ser feita a normalização dos dados através da padronização das variáveis explicativas e aplicação da transformação logarítmica da variável dependente, de seguida faremos a seleção de atributos, ou seja, a escolha das variáveis mais relevantes para a construção do modelo, ao removermos aquelas (variáveis) que não contribuem significativamente para a previsão do preço da habitação e as que apresentarem multicolinearidade.

Para isso, além do mapa de correlação de Pearson que serviu para obter informação sobre a relação entre as variáveis, dependente (preço de venda) e explicativas na Figura 3 (Cortinhas & Black, 2012; Siegel, 2016), usamos como técnica para a seleção de variáveis, a regressão *LASSO*, bem como o conhecimento obtido sobre o mercado de habitação em Singapura através da literatura e da análise exploratória dos dados nas etapas anterior (Foundation, 2001; Foundation, 2024).

LASSO (*Least Absolute Shrinkage and Selection Operator*), significa em português (Operador de Seleção e Encolhimento Absoluto Mínimo), é uma técnica de aprendizagem de máquina baseada em analise de regressão linear para tarefas de previsão com elevada precisão (Shafiee et al., 2021), ao minimizar ao máximo a diferença entre os valores reais e os valores previstos através da implementação de um parâmetro de regularização designado lambda (λ) (Ranstam & Cook, 2018).

A regressão LASSO também é utilizada como uma poderosa técnica para fazer a seleção de atributos ao aplicar uma regularização do tipo L_1, isto é, o parâmetro penalizador lambda (λ) que controla a força da regularização ao determinar o quanto os coeficientes de regressão são penalizados de formas a gerar um equilíbrio entre a precisão do ajuste do modelo e a simplicidade do modelo (Mei & Shi, 2024). A expressão matemática abaixo representa a fórmula da regressão (LASSO).

$$LASSO = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (2)$$

Em que:

$\frac{1}{2n}$ é um fator de normalização aplicado a soma dos erros quadrados, com a finalidade de normalizar o erro quadrático médio para garantir que a escala da função objetivo não dependa do número de observações.

$\sum_{i=1}^{n}(y_i - \hat{y_i})^2$, representa a soma dos erros quadrados entre os valores observados y_i (valor real da variável dependente para a i-ésima observação) e as previsões feitas pelo modelo $\hat{y}_i$ (valor previsto pelo modelo tendo em conta o conjunto de variáveis independentes ou explicativas).

λ: a expressão lambda nesta fórmula indica o parâmetro de regularização que controla a força de penalização dos coeficientes, isto é, regula o *trade-off* entre a minimização do erro de previsão e a penalização dos coeficientes β_j onde, um λ maior aumenta a penalização e reduz alguns coeficientes a zero de modo a gerar um modelo mais simples e menos propenso a *overfitting*, assim, o valor de λ determina o equilíbrio entre a precisão do ajuste do modelo e a simplicidade do modelo (Mei & Shi, 2024; Ranstam & Cook, 2018).

$\sum_{j=1}^{p}|\beta_j|$, representa a parte da função objetivo responsável pela regularização LASSO onde p é o número de variáveis independentes ou explicativas, e o $|\beta_j|$ é o valor absoluto do coeficiente de regressão para j-ésima variável independente o que permite com que, a regularização LASSO (L_1) tende a reduzir alguns coeficientes a exatamente zero, e promove a seleção de variáveis para gerar modelos mais simples (Emmert-Streib & Dehmer, 2019; Ranstam & Cook, 2018).

Desta forma, além das vantagens já mencionadas anteriormente sobre o ML, é percetível a sua superioridade ao oferecer uma estrutura poderosa e flexível que possibilita tarefas de seleção de variáveis principais que nos permitem construir modelos mais simples e com elevada precisão (Shafiee et al., 2021), ao aplicarmos a regularização do tipo L_1 aos dados, principalmente em amostras onde o número de características ou atributos são maiores que as observações (Lee et al., 2022; Mei & Shi, 2024).

Para a construção do modelo LASSO na seleção das variáveis, foram apresentados ao algoritmo 80% dos dados de entrada e de saída em treino, isto é, com todas as variáveis explicativas e dependente para que o algoritmo realizasse os cálculos necessários. Quando o treino começa, os coeficientes β_j que representam a contribuição de cada variável explicativa no modelo são inicializados (Ranstam & Cook, 2018; Siegel, 2016), o modelo calcula a predição $\hat{y_i}$ para os dados de treino com base nos coeficientes iniciais e nas variáveis explicativas. De seguida, o algoritmo realiza o cálculo da função de custo (LASSO), isto é, calcula o erro de predição e aplica o termo de regularização L_1 onde LASSO ajusta os coeficientes (βj).

Este processo de ajuste dos coeficientes continua de forma iterativa, onde a cada iteração o termo de regularização empurra os coeficientes β_j que não contribuem significativamente para melhorar a previsão a se tornarem exatamente iguais a zero, indicando que estas variáveis correspondentes (a zero) foram excluídas do modelo e as variáveis que mantêm os coeficientes diferentes de zero são aquelas consideradas importantes para a previsão do preço da habitação (Lee et al., 2022; S. Zhang et al., 2021).

Este processo continua até que o algoritmo alcance a convergência, ou seja, quando os ajustes nos coeficientes se tornam tão pequenos que o modelo não melhora significativamente.

Como resultado, o LASSO não só aprendeu os coeficientes ótimos para as variáveis mais relevantes, mas também realizou automaticamente a seleção de variáveis ao descartar algumas (variáveis) que não são úteis para a previsão do preço da habitação como se verifica na Figura 5 abaixo onde foram reduzidas a zero. Essa capacidade de seleção de variáveis durante o treinamento faz do LASSO uma técnica poderosa e interpretável, especialmente útil em situações onde há muitas variáveis explicativas e o objetivo é simplificar o modelo e ao mesmo tempo manter a precisão (do modelo) (Fonti & Belitser, 2017; Ranstam & Cook, 2018; Shafiee et al., 2021)

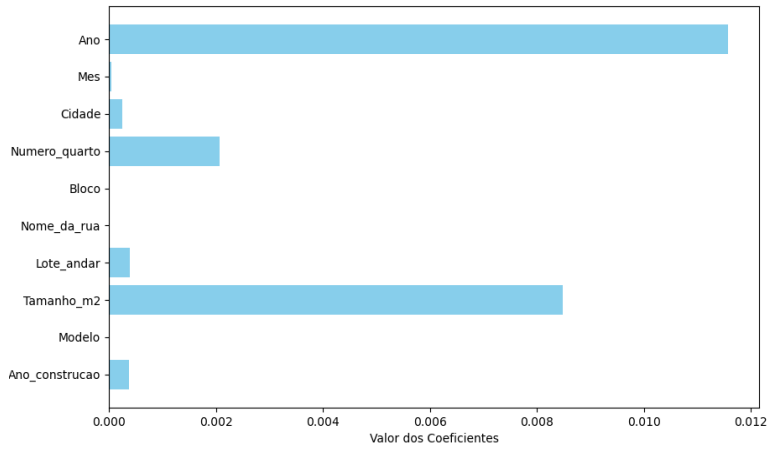


Figura 5 - Resultado do algoritmo LASSO durante a seleção de variáveis.

Com os resultados da construção do modelo de regressão LASSO para a seleção de variáveis, e do conjunto de informações obtidas através da análise exploratória dos dados, selecionamos como principais variáveis para a construção do modelo as seguintes: o *ano (em que foram realizadas as transações), tamanho em metros quadrado, cidade, número de quartos, lote por andar e ano de construção*, por se mostrarem mais relevantes, isto é, melhores preditores para o preço da habitação de acordo com a Figura 5 acima. As demais variáveis cujos coeficientes foram exatamente zero de acordo com o LASSO, foram descartadas do modelo. Como resultado temos um conjunto de dados com 905.052 observações e 7 atributos ou características onde o preço de venda é a variável dependente conforme o Anexo 9.

3.5.2.3. Relevância dos atributos ou variáveis selecionadas

A variável ano em nosso conjunto de dados, reflete em grande parte pelas pesquisas feitas, as condições macroeconómicas de Singapura em cada ano como vimos anteriormente de acordo com os relatórios do Banco Mundial, ao apresentar um crescimento de 7,7% do PIB ao ano desde a independência e um dos

IDH mais alto do mundo (World Bank, 2019). O ano de venda é crucial porque reflete as condições econômicas e de mercado vigentes no momento da transação.

Fatores como taxas de juros, políticas habitacionais e a situação económica geral tendem a influenciar significativamente os preços dos imóveis (Kohler et al., 2023; Lewandowska et al., 2023), sugerindo uma relação positiva entre o ano em que o imóvel foi transacionado e o preço de venda da habitação, influenciado por fatores macroeconómicos no momento (da venda).

O tamanho em metros quadrado e número de quarto representam um dos principais fatores determinantes do preço da habitação. Em geral, quanto maior a área útil em metros quadrado, maior será o valor do imóvel, visto que o preço por metro quadrado é um indicador comum utilizado no mercado imobiliário (INE, 2024; ONS & Registry, 2024) , bem como quanto mais quartos uma casa tem, maior tende a ser o preço.

O atributo cidade relacionado com a localização geográfica é um fator crítico que pode influenciar drasticamente os preços da habitação devido a determinados fatores como oferta e procura, infraestrutura, qualidade de vida, segurança e proximidade de serviços essenciais.

O ano de construção geralmente em termos de venda de habitação, tende a indicar o estado de conservação do imóvel e a modernidade da construção. Os Imóveis construídos recentemente costumam ter melhores padrões de construção e tecnologia, o que pode aumentar seu valor no momento da transação.

Por último a variável lote, entendemos que em terrenos, um lote maior pode permitir expansões futuras enquanto em apartamentos, andares mais altos muitas vezes oferecem melhores vistas e menos ruído, aumentando o valor do imóvel, principalmente em países como Singapura pelas características das construções (HDB, 2022). Essas variáveis são interdependentes e compõem um conjunto complexo de fatores que determinaram o preço final de venda de uma habitação (em Singapura) cuja importância para a sociedade e o impacto para a economia já foi extensamente realçado no capítulo dois deste trabalho.

3.5.3. Construção dos modelos de ML

Com os métodos e técnicas escolhidas até aqui, foi possível realizar o trabalho de análise exploratória dos dados que nos permitiu obter informações relevantes sobre as características das variáveis envolvidas, forneceu-nos uma compreensão mais profunda sobre os dados e as técnicas mais apropriadas a serem utilizadas no pré-processamento, onde fizemos a padronização e a seleção de atributos ou variáveis com o algoritmo de regressão LASSO.

Desta forma, com estas variáveis selecionadas faremos a construção do modelo de ML para a previsão do preço da habitação em Singapura com os algoritmos LASSO e o algoritmo *Random Forest* (RF). Criaremos duas versões de cada modelo, um com os parâmetros base do algoritmo e o outro com os parâmetros mais ajustados. Propomos como objetivo construir os modelos de forma que consigam

explicar pelo menos 70% da variação do preço da habitação, e usaremos como métricas o RMSE, MAE e R^2 amplamente utilizados para avaliação de modelo de ML em tarefas de regressão (Rocha & Figueira, 2017; Siegel, 2016; Thara et al., 2019), e o melhor modelo ML será selecionado em termos de desempenho com base nestas métricas.

3.5.3.1. Construção do modelo Random Forest (RF)

Como mencionado anteriormente, *Random Forest* é um tipo de algoritmo que combina os resultados de várias árvores de decisão para depois apresentar como resultado uma única árvore com maior precisão (Quang et al., 2020). Pode ser usado tanto para tarefa de classificação assim como de regressão e ao combinar várias árvores para fazer previsões com maior precisão, faz do RF atualmente uma escolha adequada para problemas que envolvem previsões competitivas (Borup et al., 2023; Moturi, 2020).

Como o algoritmo RF se baseia em Árvores de Decisão (AD) importa realçar que em ML, considera-se uma árvore de decisão como uma técnica utilizada em aprendizagem supervisionada para prever valores contínuos em caso de regressão ou classificação para prever categorias (Rocha & Figueira, 2017). A árvore de decisão é constituída por um conjunto de "*nós (nó raiz, nó interno e nó folha)"* que representam uma característica dos dados e, cada nó em AD é considerado um ponto de decisão onde o algoritmo de árvore de decisão testa uma condição ou critério de forma a gerar a melhor saída (previsão ou classificação) para a variável dependente representada pelo nó folha (Gama et al., 2012; Moturi, 2020).

Atualmente é um dos algoritmos mais bem-sucedido de propósito geral amplamente utilizado em vários domínios incluindo o económico, pela sua capacidade em fazer previsões com alta precisão cujo modo funcional leva-o a ser melhor caracterizado pela sua abordagem de "*dividir e conquistar*", isto é, ao realizar o *bootstrap* de frações de dados no conjunto de dados, cresce uma árvore de decisão em cada fração e em seguida, agrega essas previsões (Borup et al., 2023). O algoritmo *Random Forest* foi criado primeiramente por Ho em 1995 e desenvolvido por Breiman em 2001, cuja fórmula (3) abaixo representa a sua aplicação para problemas de regressão (Zhang et al., 2021).

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} t_i(x) \qquad (3)$$

Em que:

$\hat{y}$: é o valor previsto pelo modelo de RF para uma dada entrada x e representa a média das previsões feitas por todas as árvores de decisão individuais no modelo.

$\frac{1}{K}$: é o fator de normalização que calcula a média das previsões das árvores e K representa o número total de árvores de decisão no modelo RF.

$$\sum_{k=1}^{K} t_i(x)$$

Este termo é a soma das previsões feitas por todas as K árvores para a entrada de x onde, t_i é um único modelo de regressão de árvore de decisão e $t_i(x)$ representa a previsão feita pela i-ésima árvore de decisão no modelo para as entradas x.

Quanto ao seu funcionamento, para construir cada árvore, t_i, o algoritmo seleciona aleatoriamente com reposição, um subconjunto dos dados de treino através da amostragem *bootstrap* (Gama et al., 2012) e, durante a construção de cada nó de uma árvore, um subconjunto aleatório de características é selecionado e a melhor divisão é escolhida a partir desse subconjunto, através de um critério que por sua vez depende do tipo de problema (classificação ou regressão) a ser resolvido (He et al., 2022; Rocha & Figueira, 2017), essa escolha aleatória de características ou atributos tende a introduzir variabilidade adicional entre as árvores e ajuda a corrigir a correlação entre elas.

De seguida, cada árvore de decisão é construída a partir do subconjunto de dados e subconjunto de características selecionadas, e como as árvores no *Random Forest* podem crescer completamente, sem poda, significa que cada árvore é construída até o ponto onde cada nó folha contém um único exemplo ou não pode mais ser dividida e a previsão final para árvores de decisão em problemas de regressão, será a média das previsões de todas as árvores do modelo RF, isto é, uma vez criada todas as árvores K no modelo, o RF faz uma agregação das previsões de cada árvore para gerar a previsão final (Khajavi & Rastgoo, 2023; Zhang et al., 2021).

Para a construção do modelo RF através da sua implementação em linguagem de programação Python, utilizamos os pacotes *sklearn (pre_processing, model_selection, etc), pandas, numpy, matplotlib* (Foundation, 2001), conforme o Anexo A10 e foram dados os seguintes passos:

Divisão do conjunto de dados em subconjunto de treino e teste (Gama et al., 2012; Marsland, 2009; Zulkifley et al., 2020), isto é, para o nosso estudo decidiu-se dividirmos os dados de formas a utilizarmos 80% dos dados para o treino do modelo e os outros 20% para o teste, através da função *train_test_spli* com as seguintes divisões: *x_treino, x_teste, y_treino, y_teste. x_treino e y_treino*, representam os 80% dos dados para o treino do algoritmo, e os *x_teste e y_teste* representam os dados para o teste e avaliação do modelo (Rocha & Figueira, 2017).

A seguir, foi feita a criação ou construção da primeira versão do modelo, modelo RF 1 (um) como chamaremos para melhor contextualização, de acordo com a fórmula (3) do RF acima, isto é, um modelo base sem muitos ajustes de parâmetros considerando a sua implementação em Python. Desta forma, os principais parâmetros definidos para a construção do modelo RF 1 de formas a controlar o seu comportamento foram:

N_estimator: representa a letra K da fórmula e indica a quantidade de árvores de decisão para a regressão a serem construídas dentro do algoritmo RF, para o nosso modelo foi definido o valor de $K = 100$ árvores de decisão. *Max Depth* (profundidade máxima), define a profundidade máxima da árvore.

Controla o quão profundo a árvore pode crescer, para o nosso modelo foi definido o valor padrão, que faz com que as árvores cresçam até todos os nós serem puros ou conterem menos que o número mínimo de amostras (*Min Samples Leaf*), que definimos com o valor de 10, isto é, o número mínimo de amostras que pode existir em um nó folha de formas a construir árvores mais regularizadas e menos complexas (Gama et al., 2012; Rocha & Figueira, 2017).

Outro parâmetro não menos importante foi *Min Samples Split* (Mínimo de Amostras para Divisão), que controla o número mínimo de amostras que um nó deve ter antes de ser dividido cujo valor padrão é 1 mas, para o nosso estudo definimos um valor de 10 considerado alto de formas a prevenir a criação de muitos nós pequenos e assim, evitar o *overfitting*. E por último definimos os parâmetros *max feature* (número máximo de atributos) a serem considerado, neste caso foram usados todos os atributos considerados relevantes, e o parâmetro *random state,* que também foi definido para garantir que o algoritmo cria uma semente para a geração de números aleatórios de formas a tornar os resultados do trabalho reprodutível mantendo os resultados (Foundation, 2001; Lee et al., 2022; Zhang et al., 2021).

Para melhor compreensão e comparação quanto aos parâmetros para a construção do modelo na primeira versão e segunda (modelo RF 1 e modelo RF 2), apresentamos a tabela em Anexo A12 com um resumo dos principais parâmetros definidos para a construção dos dois modelos RF.

Para a etapa seguinte, após a criação do modelo definido com os parâmetros mencionados anteriormente, foi feito o treino do modelo com a função *fit* do sklearn ao apresentar ao algoritmo RF os dados de *x_treino e y_treino*, para que o algoritmo aprenda a identificar os padrões e relações nos dados que geraram as saídas da variável dependente (y), de formas a ajustar os seus parâmetros internos (critério de divisão) feito de forma iterativa para minimizar o erro entre os valores observados e os valores previstos nos dados de treino.

Os dados de treino se revestem de grande importância para o modelo, pois nesta fase do treino, os dados fornecem ao algoritmo informações para aprender, ajustar os seus hiperparâmetros, construir um modelo eficaz capaz de fazer previsões precisas e generalizar para novas entradas de dados, bem como ajuda a detetar o *overfitting* e *underfitting* (Greener et al., 2022; Guyon & Yao, 1999).

O *overfitting* e *underfitting* em ML representam dois dos principais problemas comuns que os modelos de *Machine Lerning* podem apresentar, o *overfitting* ocorre quando um modelo de ML não aprende os padrões reais subjacente nos dados e apresenta um desempenho muito melhor no subconjunto de dados de treino, mas um desempenho significativamente pior no subconjunto de teste (Agliari et al., 2024; Greener et al., 2022). Já o *underfitting*, ocorre quando um modelo de ML apresenta um desempenho baixo tanto no subconjunto de treino assim como no subconjunto de teste, ou seja, o algoritmo não aprende com os dados de treino e não é capaz de fazer generalizações para novos dados (Ghojogh & Crowley, 2019).

Depois do treino do algoritmo com os dados de treino, a primeira versão do modelo criada, o modelo RF 1, foi testado para avaliação ao ser utilizado para fazer a previsão (*função predict*) com os novos dados (de teste), onde apresentamos apenas ao modelo o subconjunto dos 20% de dados de entrada

(*x_teste*) para o modelo prever as saídas e foram calculadas as métricas para a sua avaliação (Foundation, 2001). Os mesmos passos foram dados para construir a segunda versão do modelo com a função *(RandomForestRegressor)*, isto é, o modelo RF 2, com os seus parâmetros mais ajustados conforme a tabela em Anexo A12.

Com os valores dos parâmetros do modelo RF 2 mais ajustados, nosso propósito é de garantir que o modelo final construído seja mais robusto, utilizável e generalizável para que, além de permitir uma avaliação mais rigorosa da *performance* dos modelos ao compará-los, também possa melhorar a precisão, reduzir o risco de *overfitting* ou *underfitting* e proporcionar a criação de um modelo mais adaptado ao problema (Czajkowski & Kretowski, 2019; Zhang et al., 2018; Zhang et al., 2021).

3.5.3.2. Construção do modelo LASSO

Quanto ao modelo de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*), já abordado na etapa de seleção de variáveis, aqui detalharemos apenas a construção do modelo para efeitos de previsão, visto que também é uma técnica de aprendizagem de máquina baseada em analise de regressão linear para tarefas de previsão com elevada precisão (Shafiee et al., 2021), ao minimizar ao máximo a diferença entre os valores reais e os valores previstos através da implementação de um parâmetro de regularização designado lambda (λ) conforme a fórmula abaixo já explicada anteriormente (Ranstam & Cook, 2018).

$$LASSO = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Para a sua construção de acordo com a sua implementação na linguagem Python seguindo a lógica anterior, ao construirmos duas versões do modelo com o mesmo algoritmo pelas razões já mencionada, após dividir o conjunto de dados em treino e teste, utilizamos a função *LassoCV* do *sklearn* para a construção do modelo ao definirmos os seus principais parâmetros como podemos verificar em Anexo A11, com os parâmetros a esquerda e os valores a direita em cada modelo LASSO (modelo 1 e 2).

O parâmetro *alfa*, representa o lambda da fórmula da regressão LASSO e definimos uma lista de valores para o algoritmo testá-los durante o treino para determinar o *alpha* que resulta no melhor desempenho do modelo ao minimizar o erro durante a validação cruzada (Tchakoucht et al., 2024). O segundo parâmetro é o *cv (cross validation)*, uma técnica usada em ML que consiste em dividir o conjunto de dados de treino em múltiplos subconjuntos (*folds*), treinar o modelo em alguns desses subconjuntos e testar em outro (Prusty et al., 2022; Roberts & Nowak, 2014).

Por exemplo para o modelo 1, cujo parâmetro *cv* foi definido para o valor de 20 (*cv* = 20), o conjunto de dados de treino será dividido em 20 partes. Em cada iteração, o modelo será treinado em 19 partes e testado na parte restante, repetindo o processo 20 vezes, o que pode resultar em uma estimativa mais

estável da *performance* do modelo, mas pode aumentar o tempo de execução especialmente para conjuntos de dados grandes (Prusty et al., 2022; Tchakoucht et al., 2024).

Os demais parâmetros como *max_iter, selection, fit_intercept e tol*, também são importantes para a construção e ajuste do próprio modelo. O *max_iter,* define o número máximo de iterações que o algoritmo de otimização pode realizar para encontrar a solução ótima, assim como o *selection,* determina o método usado para selecionar as características a serem atualizadas durante o ajuste do modelo, enquanto o parâmetro *tol*, representa a tolerância para o critério de paragem do algoritmo de otimização, o seu valor padrão é 0,0001 (Foundation, 2001).

A semelhança do que se fez ao criarmos os modelos com o algoritmo RF, depois de criado o modelo com os parâmetros definidos, foi feito o treino do modelo com a função *fit* do *sklearn* ao apresentar ao algoritmo LASSO os dados de *x_treino e y_treino*, para que o algoritmo aprenda a identificar os padrões e relacionamentos nos dados que geraram as saídas da variável dependente (y).

Assim, depois do treino do algoritmo com os dados de treino, os modelos LASSO (modelo 1 e 2) foram criados e submetidos para o teste a fim de serem avaliados ao serem utilizados para fazer a previsão (*função predict*) com os novos dados (de teste), onde apresentamos apenas ao modelo o subconjunto dos 20% de dados de entrada (*x_teste*) para o modelo prever as saídas e foram calculadas a métricas para as suas avaliações (Foundation, 2001). Desta forma, foram criados os modelos de ML através dos algoritmos RF e LASSO, onde os parâmetros padrões ajudaram a garantir que os modelos fossem ajustados de maneira robusta e eficiente, e extraídas as suas principais métricas para avaliação.

3.5.3.3. Métricas de avaliação dos modelos

Foi feita de seguida a avaliação dos modelos de formas a percebermos a sua *performance* ao extrairmos as métricas mencionadas anteriormente (RMSE, MAE e R^2), comumente usadas para avaliar modelos de ML para regressão de acordo com (Khajavi & Rastgoo, 2023; Zhang et al., 2021).

A Raiz Quadrada do Erro Quadrático Médio (RMSE – *Root Mean Squared Error*), mede a média dos quadrados dos erros, ou seja, a diferença entre os valores previstos pelo modelo $\widehat{y_i}$ e os valores reais ou observados de y_i cujo valor quanto menor, melhor, ao indicar que o modelo é capaz de fazer previsões com muita precisão (Rocha & Figueira, 2017). A fórmula (4) a seguir foi utilizada para o cálculo do RMSE (Zhang et al., 2021).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \qquad (4)$$

A outra métrica utilizada foi o coeficiente de determinação ou R^2, que mede a proporção da variabilidade nos dados que é explicada pelo modelo, cujo valor varia de 0 (zero) a 1 (um), onde 1 indica

que o modelo explica perfeitamente a variação nos dados, e 0 indica que o modelo não explica a variação nos dados (He et al., 2022; Zhang et al., 2021), isto é, valores próximos de 1 indicam bom ajuste do modelo aos dados. A fórmula comumente utilizada para o calculá-lo se verifica a seguir:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (5)$$

Também o Erro Médio Absoluto (MAE – *Mean Absolute Error*) foi calculado como uma métrica adicional dada a sua particularidade em relação ao RMSE ao tratar todos os erros igualmente, para melhor percebermos a *performance* dos modelos em termos de precisão (Tchakoucht et al., 2024). O MAE mede as diferenças médias absolutas entre os valores previstos pelo modelo $\hat{y}_i$ e os valores reais ou observados da variável dependente y_i, calculados com base na fórmula (6) abaixo (Karunasingha, 2022; Khajavi & Rastgoo, 2023)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (6)$$

Atendendo as diferentes escalas dos dados e de formas a termos uma interpretação mais intuitiva dos erros, tanto o RMSE e o MAE, serão transformados em erros percentuais, isto é, RMSPE (Root Mean Squared Percentage Error) e o MAPE (Mean Absolute Percentage Error) que já não dependem da escala (Foundation, 2001; Zhang et al., 2021), cujas fórmulas apresentamos a seguir.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| x\ 100 \quad (7)$$

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}\ \ x\ 100 \quad (8)$$

Ambas as métricas, MAPE e RMSPE, expressam o erro em percentagens, o que significa que os resultados independem da escala dos dados, facilitando a comparação entre diferentes modelos e dados. O MAPE é a média do erro absoluto percentual para cada ponto de dado. Esse valor está expresso como uma percentagem do erro médio absoluto em relação aos valores reais.

Assim, essas métricas foram principalmente utilizadas em Python para avaliar os modelos através das funções *mean_suquared_error, r2_score, e mean_absolute_error*, todas do pacote *sklearn.metrics* calculadas com base nos valores observados da variável dependente (*y_teste)* e os resultados das previsões dos modelos feita pela função *predict* (Foundation, 2001). Com base nos resultados destas métricas o melhor modelo será selecionado ou escolhido como modelo final de acordo com os

algoritmos de ML usados, conforme as Tabelas 6 abaixo com os resultados dos modelos de ML com o algoritmo RF e o algoritmo de regressão LASSO.

Modelo	RMSPE (Treino)	R² Score (Treino)	MAPE (Treino)	RMSPE (Teste)	R² Score (Teste)	MAPE (Teste)
LASSO 1	1.1500	0.7275	0.8200	1.1600	0.7275	0.8200
LASSO 2	1.1515	0.7272	0.8174	1.1552	0.7275	0.8193
RF 1	0.3301	0.9772	0.2204	0.3555	0.9738	0.2370
RF 2	0.3301	0.8591	0.6183	0.8159	0.8600	0.6182

Tabela 6 – Resultados em treino e teste das métricas dos modelos RF 1 e 2 e LASSO 1 e 2.

Com base nos resultados obtidos conforme a Tabela 6 acima mencionada, os algoritmos foram avaliados tantos em treino como em teste de formas a percebermos suas *performances* e, para garantir que os modelos não apresentam sinais de *overfitting*, principalmente para o RF por ser um dos algoritmos com alta capacidade preditiva (Borup et al., 2023; Moturi, 2020).

3.6. Resumo do capítulo

Ao concluirmos este capítulo, podemos verificar que os dados apresentaram qualidades para o estudo que se pretende e, os métodos e técnicas utilizadas foram eficazes para a realização do trabalho, ao possibilitar uma análise exploratória dos dados profunda, e a escolha das variáveis mais relevantes para a construção do modelo.

Como resultado, o conjunto de dados foi preparado pronto a ser usado em algoritmos de ML, isto é, para a divisão do conjunto de dados em subconjuntos de 80% de dados para o treino e 20% de dados para o teste dos algoritmos, isto é, em *x_treino, x_teste, y_treino, y_teste*. Foi feita a construção dos modelos e submetidos para treino e avaliação de forma a selecionarmos o melhor modelo após a análise e interpretação de resultados mais adiante.

Em suma, estas tarefas foram fundamentais para a criação com sucesso de um modelo de ML mais simples, fácil de interpretar, capaz de fazer generalizações em novos dados com elevado nível de precisão utilizando os algoritmos de ML *Random Forest* (RF) e a regressão LASSO, através da aprendizagem supervisionada.

CAPÍTULO IV

Análise e Interpretação de Resultados

4.1. Resultados por modelos

Para esta seção faremos uma análise e interpretação de resultados sobre o trabalho, baseada nos resultados obtidos pelos modelos de *Machine Learning* (ML) construídos para a previsão do preço da habitação em Singapura, através dos algoritmos *Random Forest* e a regressão *LASSO*. Será feita a análise dos modelos de acordo com os resultados das métricas calculadas para avaliar as suas *performances*, conforme a Tabela 6 sobre os resultados dos modelos vistos na secção anterior.

Para uma melhor compreensão, primeiramente de forma comparativa vamos analisar os resultados dos modelos (modelos 1 e 2) criados para cada algoritmo de ML de acordo com os resultados do RMSPE, MAPE, e R^2, e de seguida uma comparação dos melhores modelos de cada algoritmo para finalmente escolher-se o modelo final, isto é, o modelo com a melhor *performance* ao prever o preço da habitação com elevada precisão.

4.1.1. Resultados modelo *Random Forest*

Os modelos de ML construídos com o algoritmo RF apresentaram bons resultados de acordo com as métricas calculadas tanto para o conjunto de dados de teste e para o conjunto de dados de treino, visto que, a abordagem mais direta para identificar *overfitting* ou *underfitting* em modelos de ML, é comparar as métricas de desempenho do modelo nos conjuntos de treino e teste.

Quanto ao modelo RF 1, pelas métricas calculadas do RMSPE, os valores em treino (0,3301) e teste (0,3555) indicam que o modelo faz previsões com erros percentuais relativamente pequenos, tanto no conjunto de treinamento quanto no conjunto de teste. Isso sugere que o modelo está a capturar bem a relação entre as variáveis e não existem indícios de *overfitting*.

Os resultados do R^2 (coeficiente de determinação) no treino 0,9772 e no teste 0,9738, indicam a proporção da variabilidade dos preços das casas que é explicada pelo modelo. Os valores do R^2 para o modelo RF 1 são muito próximos de 1, tanto no treino quanto no teste, o que indica que o modelo é capaz de explicar cerca de 97,72% da variação dos preços nos dados de treino e 97,38% da variação nos dados de teste.

O R^2 é ligeiramente mais alto no conjunto de treino (0,9772) em comparação com o conjunto de teste (0,9738). Essa diferença mínima ainda sugere que o modelo está a generalizar bem para novos dados e não está apenas ajustado aos dados de treino. Se os preços reais das casas variam muito devido a diferentes características e localizações, o modelo consegue explicar 97,38% dessa variabilidade, isso indica que o modelo está muito bem ajustado aos dados e captura a maior parte das tendências dos preços.

O MAPE, erro percentual absoluto médio no teste é de 23.70%, levemente maior que o valor de 22.04% no treino, mas ainda assim é considerado consistente. Esses valores são razoáveis, especialmente no contexto imobiliário. Entretanto, as métricas indicam que o modelo de RF 1 tem um desempenho excecionalmente bom, tanto no conjunto de treino quanto no de teste, o que significa que o modelo generaliza bem para dados que não viu durante o treinamento.

Quanto ao modelo RF 2, a segunda versão do modelo criado com os parâmetros mais ajustados, os resultados calculados mostram que, o RMSPE para o conjunto de teste (0,8159) é significativamente maior do que o do treino (0.3301). Isso sugere que o modelo pode não estar a generalizar bem e pode estar a sofrer de *overfitting*, onde se ajusta bem aos dados de treino, mas não se ajusta bem aos dados de teste.

Com um R^2 de aproximadamente 0,86, tanto no treino (0,8591) quanto no teste (0,8600), o modelo RF 2 foi capaz de explicar 86% da variação nos preços das habitações com base nas variáveis fornecidas em ambos modelos (ano, tamanho em metros quadrado, cidade, número de quartos, lote por andar e ano de construção). Isso sugere que o modelo é bom em explicar a variabilidade dos dados, mas menos competitivo do que o primeiro modelo (que tem um R^2 de cerca de 0,97).

Igualmente para o MAPE, o modelo RF 2 apresenta resultados altos em treino (0,6183) e teste (0,6182), indicando que as previsões estão a uma média de 61,83% dos preços reais. Isso é considerado um valor alto e sugere que o modelo pode não ser tão preciso nas suas previsões. O modelo RF 2, embora tenha um bom R^2, apresenta sinais de *overfitting*, especialmente refletidos no RMSPE e MAPE do teste (0,8159 e 0,6183), parece não capturar muito bem a variabilidade dos dados. A precisão das suas previsões é significativamente menor em comparação ao RF 1.

Assim, pelas métricas utilizadas, o modelo RF 1 demonstra um excelente desempenho tanto em treino quanto em teste, com alta capacidade de explicar a variabilidade dos dados e fazer previsões consistentes. O modelo RF 1 tem desempenho significativamente superior em termos de R^2, MAPE, e RMSPE tanto no treino quanto no teste em relação ao modelo RF 2, mostrando-se mais consistente e com melhor capacidade preditiva. Isso o torna uma escolha confiável para prever o preço da habitação.

4.1.2. Resultados modelo LASSO

Quanto aos resultados das métricas RMSPE, MAPE e R^2 calculadas para os modelos LASSO 1 e 2, diferentes dos resultados obtidos com os modelos RF 1 e 2 construídos com o algoritmo *Random Forest*, os modelos *LASSO* apresentaram aproximadamente os mesmos resultados como se vê na Tabela 6 sobre os resultados do modelo *LASSO e RF* e em Anexo A13, mesmo depois de ajustarmos alguns parâmetros como também se verifica em Anexo A11 ao construirmos a segunda versão do modelo com a esperança de obtermos maior precisão.

O modelo LASSO 1 mostra uma boa consistência entre o desempenho no treino e no teste. O R^2 de aproximadamente 0.7275 significa que o modelo é capaz de explicar cerca de 72,75% da variação dos

dados, o que indica um ajuste moderado ou bom. Os valores do RMSPE e MAPE no treino (1,1500) e no teste (1,1600) são muito próximos, sugerindo que o modelo não está sofrendo de *overfitting*. Contudo, tanto o RMSPE quanto o MAPE indicam que há uma margem considerável de erro percentual nos resultados das previsões.

De igual modo, o modelo LASSO 2 também apresenta resultados muito semelhantes entre treino e teste. O R^2 de 0,7275 no teste e 0,7272 no treino, indica que este modelo tem praticamente o mesmo poder preditivo que o LASSO 1, explicando 72,75% da variação dos dados. Os valores de RMSPE (1,1515) e MAPE (0,8200) são ligeiramente superiores aos do LASSO 1, mas as diferenças são mínimas e provavelmente não muito significativas.

Ambos os modelos LASSO 1 e 2 diferem no número de *folds* usados na validação cruzada *(cv = 20 no modelo LASSO 1 e cv = 50 no LASSO 2)*, assim como no número de iterações (1000 para 2000) que permitissem ao algoritmo encontrar a solução ótima para a previsão do preço da habitação. A validação cruzada com mais *folds* (50) normalmente tende a fornecer uma estimativa mais robusta da *performance* do modelo, pois cada *fold* representa uma amostra menor do conjunto de dados. No entanto, como as métricas de desempenho são idênticas entre os modelos, isso indica que aumentar o número de *folds* de 20 para 50 e o número de iterações, não trouxe benefícios percetíveis na precisão do modelo LASSO 2.

Ambos os modelos (LASSO 1 e 2) têm um R^2 praticamente idênticos, sugerindo que ambos explicam a mesma quantidade de variabilidade nos dados. As métricas de erro percentual, RMSPE e MAPE, são muito próximas entre os dois modelos, com o LASSO 1 com ligeiras diferenças que podem não ter um impacto relevante no desempenho prático.

Assim, como as métricas de desempenho (RMSE, R^2, MAE) são idênticas para ambos os modelos, tanto no conjunto de treino quanto no conjunto de teste. Isso sugere que os modelos têm a mesma capacidade de predição e generalização. Embora os resultados sejam os mesmos, o modelo LASSO 2 com *cv* = 50, exige mais recursos computacionais (tempo de processamento) devido ao número maior de validações cruzadas realizadas e o número de iterações.

Para o nosso caso, essa mudança não resultou em melhoria no desempenho preditivo, e sugere que o modelo LASSO 1 já estava suficientemente robusto com a configuração anterior, razão pela qual além de produzirem resultados idênticos, se levarmos em conta a eficiência computacional, o modelo LASSO 1 tende a ser preferível, já que oferece a mesma performance com menor custo computacional.

Quanto aos impactos dos atributos ao prever o preço da habitação em Singapura, outro aspeto interessante que despertou a nossa atenção quando extraímos os melhores preditores, isto é, a ordem de importância dos atributos para o algoritmo RF e LASSO, embora já tivéssemos usado a regressão LASSO como uma técnica para a seleção de variável. O algoritmo RF também identificou aproximadamente as mesmas variáveis como os melhores preditores para prever o preço da habitação como se verifica nas Figuras 7 e 8 abaixo, o que reforça a precisão do algoritmo LASSO como um bom seletor de variáveis quando queremos criar modelos de ML mais simples, interpretável e com elevada

capacidade de generalização como se verifica com os resultados do modelo RF construído com os preditores ou variáveis selecionados através do modelo LASSO.

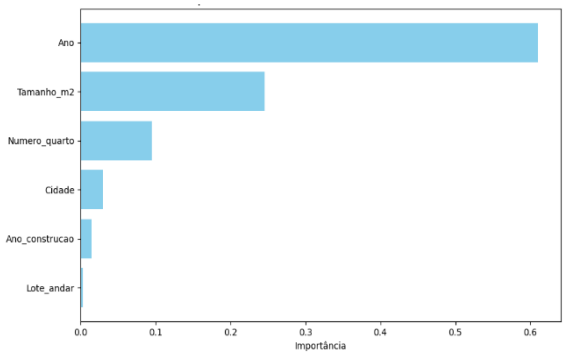


Figura 7 - Ordem de importância e impacto dos atributos na previsão do preço de habitação para o modelo RF.

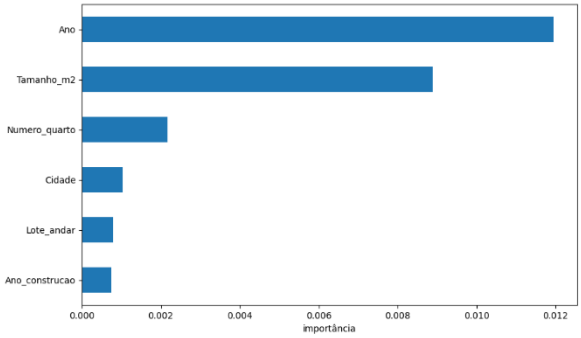


Figura 6 - Ordem de importância e impacto dos atributos na previsão do preço de habitação para o modelo LASSO.

4.1.3. Seleção do melhor modelo

De acordo com Rocha & Figueira, (2017), quando trabalhamos com modelos em analise preditiva, uma tarefa comum passa pela escolha ou seleção do melhor modelo e algoritmo aplicado em um determinado conjunto de dados. Essa escolha passa a representar a melhor solução ao compararmos os resultados das métricas obtidas pelos modelos de acordo com os algoritmos. Para o nosso estudo, a fim de encontrar e selecionar o melhor modelo, foi feita uma análise e comparação entre os resultados das métricas do melhor modelo de cada algoritmo RF e LASSO.

Os resultados mostram que, pela métrica do RMSPE, o modelo RF 1 apresenta um resultado significativamente menor (0,35) do que o modelo LASSO 1 (1,16) como se pode ver na Tabela 6 e na Figura 9 abaixo com os resultados dos modelos na fase de avaliação ao indicar que o modelo RF tem previsões mais precisas e com menos erros.

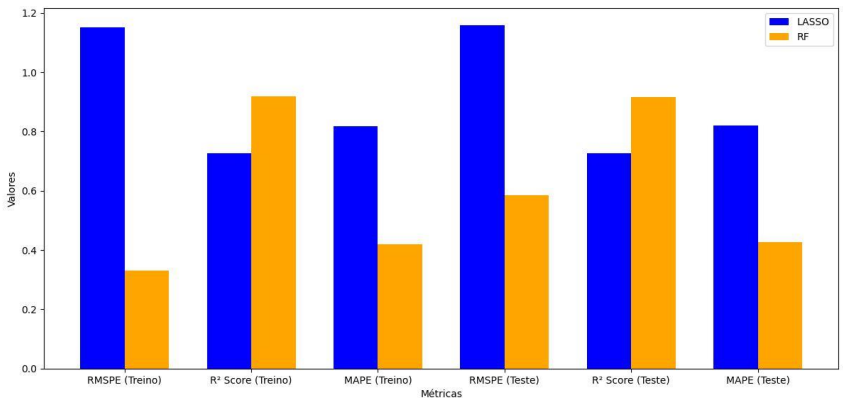


Figura 8 - Resultados das métricas dos melhores modelos Random Forest e LASSO.

De igual modo o modelo baseado no algoritmo *Random Forest* apresentou um R^2 muito mais alto (97,38%), que significa que é capaz de explicar uma maior proporção da variabilidade dos dados em comparação com o modelo baseado na regressão LASSO cujo valor foi inferior (72,75%). Isso sugere que o modelo RF 1 está a capturar melhor a estrutura de padrões subjacente dos dados de formas a realizar melhores previsões.

Deste modo, pelas métricas utilizadas (RMSE, R^2 e MAE), o modelo e algoritmo RF apresentou melhor performance tanto em treino como em teste em relação ao LASSO tendo em conta o nosso conjunto de dados, o que faz do modelo RF 1 a melhor escolha de modelo para prever o preço da habitação em Singapura com alta precisão de acordo com as variáveis selecionadas.

CAPÍTULO V

Conclusão

5.1. Principais conclusões

Este trabalho teve por tema a aplicação de modelos de *Machine Learning* para a previsão do preço da habitação em Singapura entre 1990 e 2023, encontrando um conjunto de resultados interessantes sobre os determinantes do preço da habitação e a aplicação de *Machine Learning* à sua previsão.

A habitação sempre constituiu uma das necessidades mais essenciais da vida humana, assim como a comida, água, etc., a ponto de ser consagrado na declaração universal dos direitos humanos como um direito.

A habitação tende a ser o ativo mais importante na composição da riqueza das famílias, motivo pelo qual, flutuações no preço da habitação sempre constituem preocupações para os agentes económicos.

O setor imobiliário é considerado um dos mercados mais importantes das economias desenvolvidas pelos níveis de investimentos que neles se realizam e como fonte de garantia para os empréstimos, e relaciona-se com as principais variáveis macroeconómicas como a inflação, o PIB e a taxa de juro.

Tendências recentes em previsões económicas têm enfatizado o uso de métodos e de técnicas de *Machine Learning* principalmente em cenários com muitos preditores e grande volume de dados. Ao aplicarmos esta abordagem no nosso estudo, foi possível construir um modelo capaz de prever o preço da habitação em Singapura com elevada precisão, onde o modelo de ML construído com o algoritmo *Random Forest* apresentou melhor desempenho em termos de resultados (RMSE, R^2 e MAE) em comparação com o modelo LASSO.

O *Random Forest* é um algoritmo de ML que ao construir modelos preditivos, combina várias árvores de decisão, o que frequentemente leva a uma maior capacidade de modelagem e menos *overfitting* comparado a um único modelo linear com regularização (LASSO).

Os resultados obtidos com o modelo RF foram satisfatórios, sugerindo-se que o modelo é altamente eficaz para prever preços de habitação com precisão. A baixa margem de erro e o R^2 alto fazem deste modelo uma ferramenta confiável para estimativas de preços, permitindo obter informações que auxiliam na formulação de políticas (sobre a habitação), e na tomada de decisões mais informadas na compra, venda ou avaliação de imóveis.

De acordo com os resultados do estudo, as variáveis mais importantes que determinam o preço das casas em Singapura ao terem maior impacto são: o ano de transação, e as variáveis tamanho em metros quadrado, número de quarto e cidade. Esses fatores foram confirmados tanto pelo algoritmo *Random*

Forest (RF) quanto pela regressão LASSO, ambos identificando esses atributos como os principais preditores do preço das casas.

Os resultados fornecem informações valiosas para políticas de habitação, especialmente no controle e regulação de preços, como: o monitoramento de fatores macroeconómicos, impostos sobre propriedades com base em tamanho, e a regulação do tamanho e uso do solo.

Como o ano de transação foi identificado como o principal preditor, políticas que acompanhem de perto as condições económicas (como taxas de juro, inflação, políticas fiscais) podem ser úteis para prever e mitigar os impactos nos preços da habitação. O governo poderia ajustar políticas monetárias e fiscais para controlar a inflação imobiliária em períodos de expansão económica.

No caso de Singapura, por exemplo, o governo pode ajustar as taxas de juros conforme a condição do mercado imobiliário, se os preços das habitações começarem a subir de forma descontrolada devido ao crescimento económico. O Banco Central poderia aumentar as taxas de juro para desacelerar o mercado, isso tornaria os empréstimos imobiliários mais caros, o que por sua vez, reduziria a procura por compra de casas, ajudando a estabilizar os preços.

Implementar ou ajustar impostos progressivos sobre propriedades maiores pode ajudar a controlar os preços elevados, tornando o mercado imobiliário mais acessível para a classe média e baixa. Isso poderia limitar o impacto de grandes aumentos de preços em imóveis maiores.

Considerando que o tamanho em metros quadrados e o número de quartos são fatores determinantes para o preço da habitação, políticas que incentivem a construção de moradias mais compactas e acessíveis poderiam ser implementadas, assim como incentivos fiscais ou subsídios para desenvolvedores que criem habitações menores e mais acessíveis ajudariam equilibrar a oferta e conter os aumentos de preços causados pela escassez de moradias mais baratas.

Implementar uma reforma no uso do solo para permitir a construção de edifícios de maior densidade em áreas urbanas centrais, onde os preços de terrenos são mais elevados. Esse tipo de reforma facilita a criação de mais habitações, aumentando a oferta e reduzindo a pressão sobre os preços.

Através dos resultados alcançados com este estudo, também podemos saber que medidas políticas importantes voltadas para regular as variáveis identificadas como principais determinantes do preço da habitação (o ano de transação, tamanho da casa, número de quartos, e cidade), poderiam ajudar a controlar as subidas dos preços e mencionamos as seguintes:

Em momentos de crise económica ou desaceleração do mercado imobiliário, o governo pode *introduzir subsídios ou deduções fiscais para compradores de primeira casa*, reduzindo o custo de entrada no mercado imobiliário para pessoas que pretendem comprar o seu primeiro imóvel. Isso não só ajudaria a controlar o impacto de uma desaceleração económica, mas também ajudaria a manter o mercado de habitação acessível para novos compradores.

Em épocas de rápido crescimento económico, o governo poderia adotar *medidas macroprudenciais*, como limitar a quantidade de empréstimos que os bancos podem conceder para a compra de imóveis. Isso ajudaria a evitar bolhas imobiliárias causadas por um excesso de crédito fácil

para imóveis de alto padrão, estabilizando os preços no segmento de luxo e, indiretamente, no mercado imobiliário de forma geral.

Política habitacional ativa, ao incentivar a construção de habitações de pequeno e médio porte, além de criar subsídios para imóveis mais acessíveis, pode reduzir o impacto de fatores de valorização como o tamanho e o número de quartos.

Singapura pode expandir seus programas de habitação pública subsidiada, com o HDB (*Housing Development Board*), para garantir que uma maior parte da população tenha acesso a moradias a preços acessíveis. O governo pode aumentar a oferta de casas do HDB com base nas procuras do mercado, evitando que a escassez de moradias acessíveis leve a uma elevação generalizada dos preços das propriedades.

Criar uma política de *incentivos para remodelação e renovação de casas antigas* em áreas urbanas centrais. Essas propriedades, uma vez renovadas, podem ser colocadas no mercado a preços competitivos, de modo a ajudar a aumentar a oferta de habitações em áreas de alta demanda e controlar o aumento dos preços.

Impostos e regulação do mercado, ajustar as taxas sobre transações imobiliárias, especialmente para propriedades de alto valor, pode desincentivar a especulação e ajudar a moderar o aumento excessivo de preços das casas.

Uma destas medidas mais especifica seria, impor limites à compra de imóveis por estrangeiros ou aumentar os impostos para compradores estrangeiros que não sejam residentes permanentes, especialmente em áreas onde a demanda local é alta. Isso pode ajudar a reduzir a pressão sobre o aumento dos preços causada pela demanda externa, e tornar as propriedades mais acessíveis para os cidadãos.

O governo também pode aumentar o imposto sobre o ganho de capital para transações imobiliárias de curto prazo, isto é, propriedades que são compradas e revendidas em um período de menos de três anos podem ser taxadas com uma taxa mais alta, de formas a desencorajar especuladores e, assim, ajudar a estabilizar o mercado imobiliário a longo prazo.

Essas medidas, baseadas nos resultados do modelo de previsão, podem ajudar a controlar as subidas de preços das casas em Singapura, tornando o mercado mais acessível e equilibrado para os compradores locais e não só. A implementação dessas políticas ajudaria a mitigar os efeitos de bolhas imobiliárias e garantir uma oferta sustentável de habitação no longo prazo.

Podemos concluir que, com os métodos e técnicas escolhidas associadas a qualidade dos dados e da fonte de dados selecionada, bem como do tratamento dado aos dados durante a análise exploratória, tornou possível a elaboração deste trabalho com um nível de sucesso esperado ao construirmos um modelo de ML capaz de prever o preço da habitação com elevada precisão em Singapura, e com o qual podemos tomar decisões de caracter socioeconómica importantes.

Referências Bibliográficas

Abreu, M., Afonso, A., Escária, V., & Ferreira, C. (2018). *Economia Monetária e Financeira* (J. Costa, Ed.; 3rd ed.). Escolar Editora.

Agliari, E., Alemanno, F., Aquaro, M., & Fachechi, A. (2024). Regularization, early-stopping and dreaming: A Hopfield-like setup to address generalization and overfitting. *Neural Networks*, *177*. https://doi.org/10.1016/j.neunet.2024.106389

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies*, *9*(3). https://doi.org/10.3390/technologies9030052

Akansu, A. N., Kulkarni, S. R., & Malioutov, D. M. (2016). *Financial Signal Processing and Machine Learning* (First). John Wiley & Sons Ltd, The Atrium Southem Gate, Chicester, West Sussex, PO19 8SQ, United Kingdom.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. *Applied Sciences (Switzerland)*, *8*(11). https://doi.org/10.3390/app8112321

Basu, R., & Ferreira, J. (2020). Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*, *48*, 1674–1693. https://doi.org/10.1016/j.trpro.2020.08.207

Bian, T., Chen, J., Feng, Q., & Li, J. (2019). *Comparing Econometric Analyses with Machine Learning Approaches: A Study on Singapore Private Property Market*.

Blanchard, O., Amighini, A., & Giavazzi, F. (2021). *Macroeconomic: A European Perspective* (Fourth Edition). Pearson Education Limited, Harlow CM17 9NA.

Borup, D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*(2), 841–868. https://doi.org/10.1016/j.ijforecast.2022.02.010

Bostrom, N. (2014). *Superinteligência: Caminhos, Perigos, Estratégias* (1st ed.). Relógio D´água Editores, Rua Sylvio Roberto, nº15, 1000-282.

Bris, A., Cabolis, C., Caballero, J., Hediger, M., Jobin, C., Milner, W., Pistis, M., & Zargari, M. (2020). *Imd World Digital Competitiveness Ranking 2020*. https://www.imd.org/centers/wcc/world-competitiveness-center/

Camões, F., & Vale, S. (2018). Housing Valuation, Wealth Perception, and Homeowners' Portfolio Composition. *Journal of Family and Economic Issues*, *39*(3), 494–508. https://doi.org/10.1007/s10834-018-9570-y

Case, K. E., Quigley, J. M., & Shiller, R. J. (2005). Comparing Wealth Effects: The Stock Market versus the Housing Market. *Topics in Macroeconomics*, *5*(1). https://doi.org/10.2202/1534-6013.1235

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to Algorithms Third Edition Introduction to Algorithms* (M. I. of Technology, Ed.; Third Edition). The MIT Press Cambridge, Massachusetts London, England.

Cortinhas, C., & Black, K. (2012). *Statistics for Business and Economics*. John Wiley & Sons, Ltd.

Costa, E., & Simões, A. (2004). *Inteligência Artificial: Fundamentos e Aplicações* (F.-E. de I. Lda, Ed.). FCA - Editora de Informática, Lda.

Cui, S., Hero III, A. O., Luo, Z.-Q., & Moura, J. M. F. (2016). *Big Data: Over Networks* (C. U. Press, Ed.). Cambridge University Press.

Czajkowski, M., & Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. *Expert Systems with Applications*, *137*, 392–404. https://doi.org/10.1016/j.eswa.2019.07.019

Domingos, P. (2017). *A Revolução do Algoritmo Mestre: Como a Aprendizagem Automática Está a Mudar o Mundo* (2nd ed.). Letras & Diálogos, Estrada das Palmeiras, 59 Queluz de Baixo 2730-132 Barcarena.

Dornbusch, R., Fischer, S., & Startz, R. (2013). *Macroeconomics* (A. J. Affonso, V. R. Nepomuceno, & C. L. Silva, Eds.; 11th ed.). McGraw-Hill Companies, Inc, Inc., New York, New York 10020.

Eldionara, M., Machado, R., Ceretta, P. S., & Mendes Vieira, K. (2014, December 1). A Relação Entre As Variáveis Macroeconômicas E A Concessão De Crédito No Mercado Imobiliário Brasileiro. *Universidade Federal de Santa Maria (UFSM)*.

Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. In *Machine Learning and Knowledge Extraction* (Vol. 1, Issue 1, pp. 359–383). MDPI. https://doi.org/10.3390/make1010021

Favilukis, J., Ludvigson, S. C., & Nieuwerburgh, S. Van. (2017). The Macroeconomic Effects of Housing Wealth, Housing Finance, and Limited Risk Sharing in General Equilibrium. *Journal of Political Economy*, *125*, 140–223.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* (1st ed.). Cambridge University Press, The Edinburgh Building, Cambridge CB 8RU, UK.

Fonti, V., & Belitser, E. (2017). *Feature Selection using LASSO*.

Foundation, P. S. (2001). *Applications for Python*. Python Software Foundation. https://www.python.org/about/apps/

Foundation, T. R. (2024, April 7). *What is R?* The R Foundation. https://www.r-project.org/about.html

Gama, J., Carvalho, A. P. de L., Faceli, K., Lorena, A. C., & OLiveira, M. (2012). *Extração de Conhecimento de Dados: Data Maning* (M. Robalo, Ed.; 1st ed.). Edições Sílabo, Lda - R. Cidade de Manchester, 2 1170 - 100.

Ghojogh, B., & Crowley, M. (2019). *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial*. http://arxiv.org/abs/1905.12787

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*(1), 40–55. https://doi.org/10.1038/s41580-021-00407-0

Guyon, X., & Yao, J.-F. (1999). On the Underfitting and Overfitting Sets of Models Chosen by Order Selection Criteria. In *Journal of Multivariate Analysis* (Vol. 70). www.idealibrary.com

HDB, H. & D. B. (2022, March 8). *About Us*. Housing & Development Board. https://www.hdb.gov.sg/cs/infoweb/about-us

He, S., Wu, J., Wang, D., & He, X. (2022). Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere*, *290*. https://doi.org/10.1016/j.chemosphere.2021.133388

INE, I. N. de E. (2024, July 11). *Estatísticas Construção e habitação*. Https://Www.Ine.Pt/Xportal/Xmain?Xpgid=ine_tema&xpid=INE&tema_cod=1610&xlang=pt. https://www.ine.pt/xportal/xmain?xpgid=ine_tema&xpid=INE&tema_cod=1610&xlang=pt

Kaffash, S., Nguyen, A. T., & Zhu, J. (2021). Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, *231*. https://doi.org/10.1016/j.ijpe.2020.107868

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2020). *Understanding House Price Appreciation using Multi-source Big Geo-data and Machine Learning*.

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, 609–629. https://doi.org/10.1016/j.ins.2021.11.036

Kelleher, J. D., Namee, B. Mac, & D´Arcy, A. (2015). *Fundamentals of Machine Learning For Predictive Data Analytics: Algorithms, Worked Exemples, And Case Studies* (1st ed.). Massachusetts Institute of Tecnology.

Khajavi, H., & Rastgoo, A. (2023). Predicting the carbon dioxide emission caused by road transport using a Random Forest (RF) model combined by Meta-Heuristic Algorithms. *Sustainable Cities and Society*, *93*. https://doi.org/10.1016/j.scs.2023.104503

Khandani, A. E., Lo, A. W., & Merton, R. C. (2013). Systemic risk and the refinancing ratchet effect. *Journal of Financial Economics* , *108*, 29–45.

Kohler, K., Tippet, B., & Stockhammer, E. (2023). House price cycles, housing systems, and growth models. *European Journal of Economics and Economic Policies: Intervention*, *20*(3), 461–490. https://doi.org/10.4337/ejeep.2023.0121

Lagoa, S., Leão, E., & Santos, J. (2004, July 4). Sistema Bancário: Evolução Recente E Seu Papel No Ajustamento Da Economia Portuguesa. *Departamento de Prospectiva e Planeamento*.

Leão, E. R., Leão, P. R., & Lagoa, S. C. (2019). *Política Monetária e Mercados Financeiros* (M. Robalo, Ed.; 3rd ed.). Edições Sílabo, Lda.

Lee, J. H., Shi, Z., & Gao, Z. (2022). On LASSO for predictive regression. *Journal of Econometrics*, *229*(2), 322–349. https://doi.org/10.1016/j.jeconom.2021.02.002

Lewandowska, G., Taracha, M., & Maciuk, K. (2023). Socio-economic factors associated with house prices. Evidence based on key macroeconomic aggregates globally. *Budownictwo i Architektura*, *22*(3), 45–58. https://doi.org/10.35784/bud-arch.3635

Li, K.-C., Jiang, H., Yang, L. T., & Cuzzocrea, A. (2015). *Big Data: Algorithms, Analytics, And Applications*. Tylor & Francis Group.

Lin-Heng, L. (2020). *Public Housing In Singapore: A Success Story In Sustainable Development* [National University of Singapore]. http://law.nus.edu.sg/apcel/wps.htmlElectroniccopyavailableat:https://ssrn.com/abstract=3595956

Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Mankiw, N. G. (1998). *Principles of Economics* (E. Barrosse, Ed.). The Dryden Press, 6277 Sea Harbor Drive, Orlando, FL 32887-6777.

Mankiw, N. G. (2015). *Macroeconomia* (8 ed.). LTC — Livros Técnicos e Científicos Editora Ltda, Rio de Janeiro, Travessa do Ouvidor, 11.

Marsland, S. (2009). *Machine Learning: An Algorithmic Perspective* (1st ed.). Chapman & Hall / CRC, Taylor & Francisc Group, 6000 Broken Sound Parkway NW, Suite 300.

Mei, Z., & Shi, Z. (2024). On LASSO for high dimensional predictive regression. *Journal of Econometrics*, *242*(2). https://doi.org/10.1016/j.jeconom.2024.105809

Mendonça, A., Magriço, V., Vale, S., Abreu, A., & Cunha, V. (2021). *Lições de Macroeconomia: Uma introdução* (M. Robalo, Ed.; 1st ed.). Edições Sílabo, R. Cidade de Manchester, 2 1170-100.

Mian, A., Rao, K., & Sufi, A. (2013). Household Balance Sheets, Consumption, and the Economic Slump. *Quarterly Journal of Economics*, *128*, 1687–1726.

Moturi, S. (2020). Classification Model for Prediction of Heart Disease using Correlation Coefficient Technique. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2116-– 2123. https://doi.org/10.30534/ijatcse/2020/185922020

Musarat, M. A., Alaloul, W. S., & Liew, M. S. (2021). Impact of inflation rate on construction projects budget: A review. In *Ain Shams Engineering Journal* (Vol. 12, Issue 1, pp. 407–414). Ain Shams University. https://doi.org/10.1016/j.asej.2020.04.009

Obschonka, M., & Audretsch, D. B. (2020). Artificial intelligence and big data in entrepreneurship: a new era has begun. *Small Business Economics*, *55*(3), 529–539. https://doi.org/10.1007/s11187-019-00202-4

ONS, O. for N. S., & Registry, H. L. (2024). *Housing: Property price, private rent and household statistics*. Office for National Statistics. https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads#full-publication-update-history

ONU. (1948). *Declaração Universal Dos Direitos Humanos* (Nações Unidas, Ed.). Nações Unidas. https://unric.org/pt/declaracao-universal-dos-direitos-humanos/

ONU. (1966, December 19). *Pacto Internacional dos Direitos Econômicos, Sociais e Culturais*. https://www.unicef.org/brazil/pacto-internacional-dos-direitos-econ%C3%B4micos-sociais-e-culturais

Ozsahin, D. U., Taiwo Mustapha, M., Mubarak, A. S., Said Ameen, Z., & Uzun, B. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. *Proceedings - 2022 International Conference on Artificial Intelligence in Everything, AIE 2022*, 87–94. https://doi.org/10.1109/AIE57029.2022.00024

Pinheiro, J. L. (2012). *Mercado de Capitais: Fundamentos e Técnicas* (6th ed.). Editora Atlas S.A.

Poole, D. L., & Mackworth, A. K. (2010). *Artificial Inteligence: Fundations Of Computational Agents* (1st ed.). Cambridge University Press, 32 Avenue of the Americas, New York, NY 10013-2473, USA.

Prüfer, J., & Schottmüller, C. (2021). Competing with Big Data*. *Journal of Industrial Economics*, *69*(4), 967–1008. https://doi.org/10.1111/joie.12259
Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*. https://doi.org/10.3389/fnano.2022.972421
Quang, T., Minh, N., Hy, D., & Bo, M. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111
Rącka, I., & Rehman, S. K. ur. (2018). Housing Market in Capital Cities – the Case of Poland and Portugal. *Geomatics and Environmental Engineering*, *12*(3), 75. https://doi.org/10.7494/geom.2018.12.3.75
Ranstam, J., & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895
Rawool, A. G., Rogye, D. V, Rane, S. G., Vinayk, D. R., & Bharadi, A. (2021). House Price Prediction Using Machine Learning. In *IRE Journals* | (Vol. 4).
Roberts, S., & Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics and Data Analysis*, *70*, 198–211. https://doi.org/10.1016/j.csda.2013.09.008
Rocha, M., & Figueira, P. G. (2017). *Análise e Exploração de Dados em R* (F.-E. de I. Lda, Ed.).
Sachs, J. D., & Larrain B., F. (1995). *Macroeconomia*. Editora McGraw-Hill ltda, Rua Tabapuã, 1105, Itaim-Bibi CEP 04533-905.
Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM JOURNAL*.
Schwab, K. (2017). *The Global Competitiveness Report 2017–2018*.
Schwab, K. (2019). *The Global Competitiveness Report 2019*.
Shafiee, S., Lied, L. M., Burud, I., Dieseth, J. A., Alsheikh, M., & Lillemo, M. (2021). Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery. *Computers and Electronics in Agriculture*, *183*. https://doi.org/10.1016/j.compag.2021.106036
Shi, Y. (2022). Advances in Big Data Analytics: Theory, Algorithms and Practices. In *Advances in Big Data Analytics: Theory, Algorithms and Practices*. Springer Science+Business Media. https://doi.org/10.1007/978-981-16-3607-3
Siegel, A. F. (2016). *Pratical Business Statistics*. Elsevier Inc.
Silva, A. C. (2023, December 4). *Antevisão da Economia e Políticas 2024*. IPPS-Iscte, Avenida das Forças Armadas,1649-026, Auditório do Edifício 4, do Iscte (Iscte CVTT). https://ipps.iscte-iul.pt/index.php/pt/divulgacao/conferencias/politicas-publicas/2024
Simons, G. L. (1986). *Introdução à Inteligência Artificial* (1st ed.). Clássica Editora, R. Glória, 10, r/c 1298.
Sloman, J., & Wride, A. (2009). *Economics* (Seventh Edition). Pearson Education Limited, Edinburgh Gate, Harlow Essex CM20 2JE.
Snyder, T. C., & Vale, S. (2022). House prices and household credit in the Eurozone: A single monetary policy with dissonant transmission mechanisms. *Quarterly Review of Economics and Finance*, *84*, 243–256. https://doi.org/10.1016/j.qref.2022.01.018
Tchakoucht, T. A., Elkari, B., Chaibi, Y., & Kousksou, T. (2024). Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems. *Energy Reports*, *12*, 988–999. https://doi.org/10.1016/j.egyr.2024.07.002
Teixeira, M. C. C., Superior Agrária, E., Maria Caridad Ocerin, J., & Ceular Villamandos, N. (2010, November 17). Redes Neuronais Artificiais para estimar o preço da habitação em Portugal. *International Meeting on Regional Science*.
Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering and Electronic Business*, *12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03

Thara, T. D. K., Prema, P. S., & Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, *128*, 544–550. https://doi.org/10.1016/j.patrec.2019.10.029

The Economist. (2017). Why 80% of Singaporeans live in government-built flats. *The Economist*. https://www.economist.com/asia/2017/07/06/why-80-of-singaporeans-live-in-government-built-flats

The Economist. (2019). After half a century of success, the Asian tigers must reinvent themselves. *The Economist*. https://www.economist.com/special-report/2019/12/05/after-half-a-century-of-success-the-asian-tigers-must-reinvent-themselves

Theodoridis, S. (2015). *Machine Learnng: A Bayesian and Optimization Perspective* (Elsevier Ltd. & C. Kent, Eds.; 1st ed.). Elsevier, 125 London Wall, London, EC2Y 5AS, UK.

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik*, *125*(3), 1439–1443. https://doi.org/10.1016/j.ijleo.2013.09.017

Wickens, M. (2011). *Macroeconomic Theory: A Dynamic General Equilibrium Approach* (Second Edition). Princeton Universy Press, 41 William Street New Jersey 08540.

World Bank. (2019, April 9). *The World Bank In Singapore*. The World Bank. https://www.worldbank.org/en/country/singapore/overview

Wu, Y., & Lux, N. (2018). U.K. House Prices: Bubbles or Market Efficiency? Evidence from Regional Analysis. *Journal of Risk and Financial Management*, *11*(3), 54. https://doi.org/10.3390/jrfm11030054

Yaseen, H. K., & Mahdi Obaid, A. (2020). Big Data: Definition, Architecture & Applications. *International Journal On Informatics Visualization*, *4*.

Yi, C., Kah, C., Ong, Y., & Xian, L. (2022). *The Determinants Of Housing Price In Selected Developed And Developing Economies*. Universiti Tunku Abdul Rahman.

Zaman, U., Waqar, M., & Zaman, A. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. In *Soft Computing and Machine Intelligence Journal* (Issue 1). www.kaggle.com

Zhang, C., Vinyals, O., Munos, R., & Bengio, S. (2018). *A Study on Overfitting in Deep Reinforcement Learning*. http://arxiv.org/abs/1804.06893

Zhang, S., Wu, J., Jia, Y., Wang, Y. G., Zhang, Y., & Duan, Q. (2021). A temporal LASSO regression model for the emergency forecasting of the suspended sediment concentrations in coastal oceans: Accuracy and interpretability. *Engineering Applications of Artificial Intelligence*, *100*. https://doi.org/10.1016/j.engappai.2021.104206

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, *12*(1), 469–477. https://doi.org/10.1016/j.gsf.2020.03.007

Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: A survey of literature. *International Journal of Modern Education and Computer Science*, *12*(6), 46–54. https://doi.org/10.5815/ijmecs.2020.06.04

Anexos

Anexo A1 - Resultado do conjunto de dados com coluna mês separados em Ano e mês e tradução dos atributos para português.

	Ano	Mes	Cidade	Numero_quarto	Bloco	Nome_da_rua	Lote_andar	Tamanho_m2	Modelo	Ano_construcao	Preco_venda
1	1990	1	ANG MO KIO	1	309	ANG MO KIO AVE 1	10 TO 12	31	IMPROVED	1977	9000
2	1990	1	ANG MO KIO	1	309	ANG MO KIO AVE 1	04 TO 06	31	IMPROVED	1977	6000
3	1990	1	ANG MO KIO	1	309	ANG MO KIO AVE 1	10 TO 12	31	IMPROVED	1977	8000
4	1990	1	ANG MO KIO	1	309	ANG MO KIO AVE 1	07 TO 09	31	IMPROVED	1977	6000
5	1990	1	ANG MO KIO	3	216	ANG MO KIO AVE 1	04 TO 06	73	NEW GENERATION	1976	47200
6	1990	1	ANG MO KIO	3	211	ANG MO KIO AVE 3	01 TO 03	67	NEW GENERATION	1977	46000
7	1990	1	ANG MO KIO	3	202	ANG MO KIO AVE 3	07 TO 09	67	NEW GENERATION	1977	42000
8	1990	1	ANG MO KIO	3	235	ANG MO KIO AVE 3	10 TO 12	67	NEW GENERATION	1977	38000
9	1990	1	ANG MO KIO	3	235	ANG MO KIO AVE 3	04 TO 06	67	NEW GENERATION	1977	40000
10	1990	1	ANG MO KIO	3	232	ANG MO KIO AVE 3	01 TO 03	67	NEW GENERATION	1977	47000

Showing 1 to 10 of 915,374 entries, 11 total columns

Anexo A2 - Verificação de valores ausentes NA na coluna (variável) Número de quarto

	Ano	Mes	Cidade	Numero_quarto	Bloco	Nome_da_rua	Lote_andar	Tamanho_m2	Modelo	Ano_construcao	Preco_venda
334	1990	1	BEDOK	*NA*	716	BEDOK RESERVOIR RD	07 TO 09	143	APARTMENT	1984	200000
335	1990	1	BEDOK	*NA*	725	BEDOK RESERVOIR RD	01 TO 03	151	MAISONETTE	1984	195000
395	1990	1	BUKIT BATOK	*NA*	223	BT BATOK EAST AVE 3	10 TO 12	141	APARTMENT	1985	182000
396	1990	1	BUKIT BATOK	*NA*	223	BT BATOK EAST AVE 3	07 TO 09	146	APARTMENT	1985	183000
397	1990	1	BUKIT BATOK	*NA*	221	BT BATOK EAST AVE 3	07 TO 09	141	APARTMENT	1985	210000
398	1990	1	BUKIT BATOK	*NA*	221	BT BATOK EAST AVE 3	04 TO 06	141	APARTMENT	1985	180000
399	1990	1	BUKIT BATOK	*NA*	214	BT BATOK ST 21	01 TO 03	146	APARTMENT	1984	161200
400	1990	1	BUKIT BATOK	*NA*	120	BT BATOK CTRL	07 TO 09	148	APARTMENT	1985	196000
401	1990	1	BUKIT BATOK	*NA*	121	BT BATOK CTRL	04 TO 06	145	APARTMENT	1985	196000
402	1990	1	BUKIT BATOK	*NA*	142	BT BATOK ST 11	10 TO 12	145	APARTMENT	1984	188000

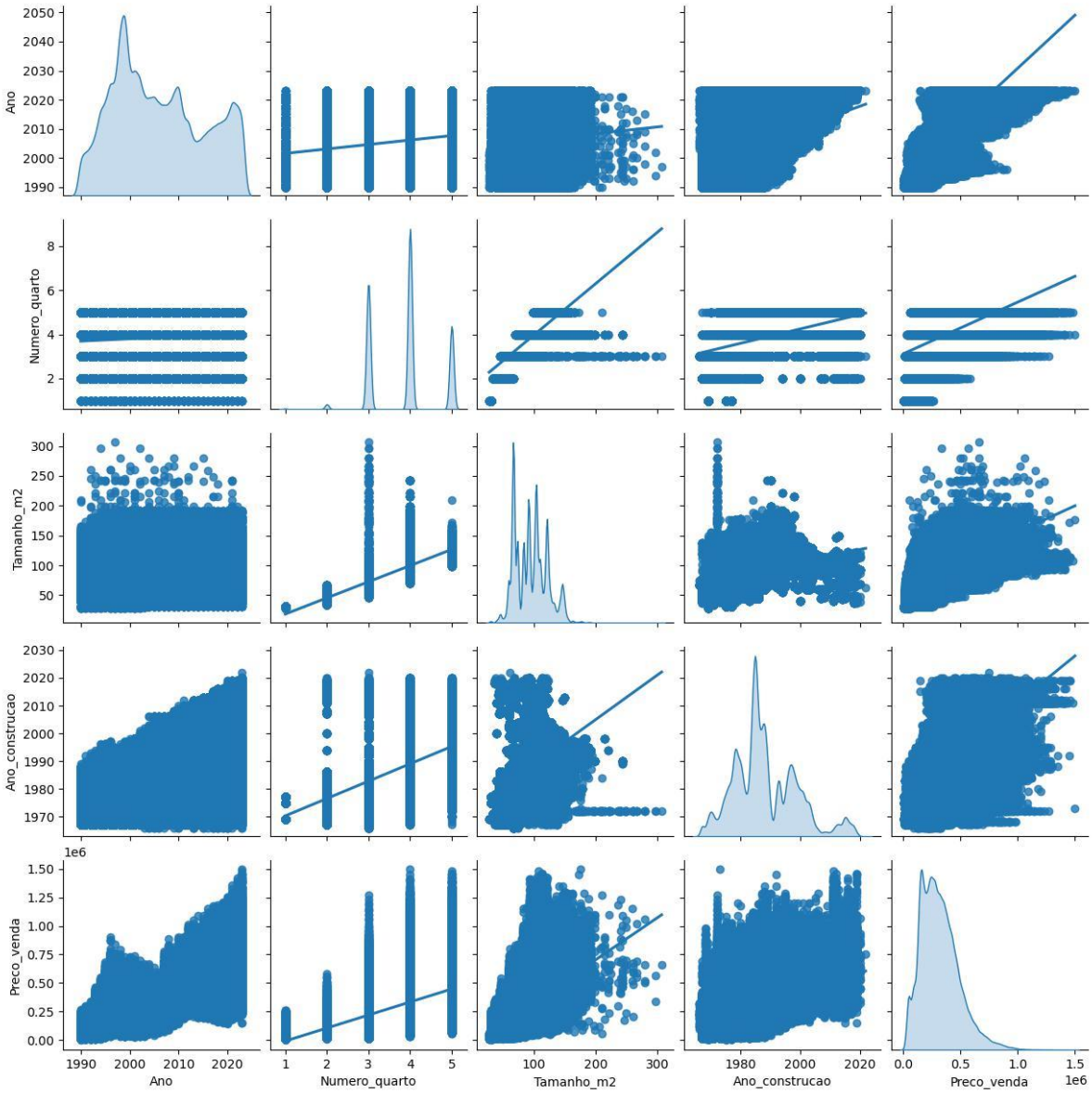
Showing 1 to 10 of 69,467 entries, 11 total columns

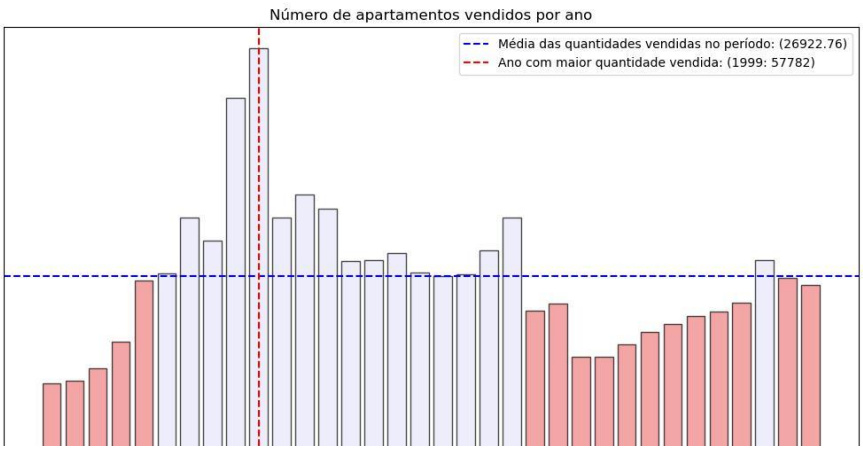
Anexo A3 - Resultado do teste de normalidade de Shapiro-wilk para a variável preço de venda.

```
        Shapiro-Wilk normality test

data:  sample(dados$Preco_venda, size = 5000)
W = 0.95044, p-value < 2.2e-16
```

Anexo A4 - Representação gráfica da correlação entre as variáveis numéricas.





Anexo A5 - Quantidade media de apartamentos vendidos por ano e o ano em que a maior quantidade foi vendida entre o período de 1990 a 2023.

Anexo A6 - Conjunto de dados considerados outliers.

	Ano	Mes	Cidade	Numero_quarto	Bloco	Nome_da_rua	Lote_andar	Tamanho_m2	Modelo	Ano_construcao	Preco_venda
1	1996	6	YISHUN	4	652	YISHUN AVE 4	01 TO 03	187	APARTMENT	1992	883000
2	1996	8	BISHAN	4	117	BISHAN ST 12	22 TO 24	163	MAISONETTE	1987	870000
3	1996	10	BISHAN	4	285	BISHAN ST 22	19 TO 21	172	MAISONETTE	1992	868000
4	1996	11	BISHAN	4	102	BISHAN ST 12	22 TO 24	163	MAISONETTE	1987	900000
5	1997	1	BISHAN	4	173	BISHAN ST 13	19 TO 21	163	MAISONETTE	1987	840000
6	2008	3	QUEENSTOWN	4	150	MEI LING ST	19 TO 21	150	APARTMENT	1995	890000
7	2008	5	QUEENSTOWN	4	148	MEI LING ST	19 TO 21	149	APARTMENT	1995	832000
8	2008	6	QUEENSTOWN	4	150	MEI LING ST	19 TO 21	148	APARTMENT	1995	830000
9	2008	7	QUEENSTOWN	4	150	MEI LING ST	10 TO 12	150	APARTMENT	1995	820000
10	2010	7	BISHAN	4	286	BISHAN ST 24	22 TO 24	172	MAISONETTE	1992	900000

Showing 1 to 10 of 10,322 entries, 11 total columns

Anexo A7 - Resultado do conjunto de dados com as variáveis categóricas transformadas para o tipo factor em R.

	Ano	Mes	Cidade	Numero_quarto	Bloco	Nome_da_rua	Lote_andar	Tamanho_m2	Modelo	Ano_construcao	Preco_venda
188942	1998	4	3	4	396	282	3	89	5	1986	250000
134058	1996	8	19	5	26	185	1	138	3	1985	492000
124022	1996	5	11	3	354	120	4	76	3	1984	189000
685285	2013	7	9	3	600	92	2	67	2	1980	350000
226318	1998	12	10	3	407	272	2	65	1	1977	128000
365209	2002	2	18	3	248	183	4	67	2	1984	131000
648795	2011	10	2	4	80	14	4	92	3	2006	480000

Showing 1 to 7 of 10,000 entries, 11 total columns

Anexo A8 - Resultado do conjunto de dado padronizado.

```
df_padronizado
```

	Ano	Mes	Cidade	Numero_quarto	Bloco	Nome_da_rua	Lote_andar	Tamanho_m2	Modelo	Ano_construcao	Preco_venda
1	-1.738271	-1.641042	-1.624825	-3.767717	-0.985841	-1.494502	-1.034467	-2.502374	-0.764665	-1.056852	0.792459
2	-1.738271	-1.641042	-1.624825	-3.767717	-0.985841	-1.494502	-0.518254	-2.502374	-0.764665	-1.056852	0.771620
3	-1.738271	-1.641042	-1.624825	-3.767717	-0.985841	-1.494502	-1.034467	-2.502374	-0.764665	-1.056852	0.786547
4	-1.738271	-1.641042	-1.624825	-3.767717	-0.985841	-1.494502	-0.002040	-2.502374	-0.764665	-1.056852	0.771620
5	-1.738271	-1.641042	-1.624825	-1.127606	-0.983869	-1.494502	-0.518254	-0.870892	-0.452741	-1.152971	0.865433
...	...	...	...	...	...	...	...	...	...	...	...
905048	1.897467	1.586392	1.066764	0.192450	-0.405835	-0.105177	-0.002040	1.964778	1.418801	0.000466	0.958082
905049	1.897467	1.586392	1.066764	0.192450	-0.664273	0.334564	-0.002040	1.809399	1.106877	-0.095654	0.958767
905050	1.897467	1.586392	1.066764	0.192450	-0.851690	0.334564	-1.034467	1.809399	1.106877	-0.095654	0.957525
905051	1.897467	1.586392	1.066764	0.192450	0.734450	0.442906	0.514173	1.809399	1.106877	0.000466	0.958447
905052	1.897467	1.586392	1.066764	0.192450	0.738395	0.442906	0.514173	1.809399	1.106877	0.000466	0.958628

905052 rows × 11 columns

Anexo A9 - Conjunto de dados com os melhores preditores para a construção do modelo de ML.

Filter

	Ano	Cidade	Numero_quarto	Lote_andar	Tamanho_m2	Ano_construcao	Preco_venda
1	-1.738271	-1.624825	-3.767717	-1.034467169	-2.5023740	-1.0568515	0.7924591
2	-1.738271	-1.624825	-3.767717	-0.518253689	-2.5023740	-1.0568515	0.7716197
3	-1.738271	-1.624825	-3.767717	-1.034467169	-2.5023740	-1.0568515	0.7865469
4	-1.738271	-1.624825	-3.767717	-0.002040209	-2.5023740	-1.0568515	0.7716197
5	-1.738271	-1.624825	-1.127606	-0.518253689	-0.8708924	-1.1529713	0.8654332
6	-1.738271	-1.624825	-1.127606	0.514173270	-1.1039612	-1.0568515	0.8644244
7	-1.738271	-1.624825	-1.127606	-0.002040209	-1.1039612	-1.0568515	0.8608331
8	-1.738271	-1.624825	-1.127606	-1.034467169	-1.1039612	-1.0568515	0.8568312
9	-1.738271	-1.624825	-1.127606	-0.518253689	-1.1039612	-1.0568515	0.8588889

Showing 1 to 9 of 905,052 entries, 7 total columns

Anexo A10 - Pcotes python utilizados para a construição dos modelos de ML.

In [156]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
from scipy import stats
from sklearn.preprocessing import LabelEncoder, MinMaxScaler, StandardScaler
from tabulate import tabulate
import warnings
from sklearn.linear_model import LinearRegression, LassoCV, Lasso
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score, mean_absolute_percentage_error
from sklearn.ensemble import RandomForestRegressor
warnings.filterwarnings("ignore")
```

In [2]:
```
from platform import python_version
print('Versão da Linguagem Python utilizada para este trabalho foi Jupyter Notebook:', python_version())
```

```
Versão da Linguagem Python utilizada para este trabalho foi Jupyter Notebook: 3.9.13
```

Anexo A11 - Parâmetros definidos para a construção do modelo LASSO 1 e 2.

	Parâmetro	Modelo 1 (modelo_lasso_1)	Modelo 2 (modelo_lasso_2)
1	alphas	[0, 1, 0.1, 0.001, 0.0001, 0.0005]	[0, 1, 0.1, 0.001, 0.0001, 0.0005]
2	cv	20	50
3	random_state	42	42
4	fit_intercept	True	True
5	max_iter	1000	2000
6	tol	0.0001	0.0001
7	selection	cyclic	cyclic

Anexo A12 - Parâmetros definidos para a construção do modelo Random Forest 1 e 2.

	Parâmetro	Modelo_RF_1	Modelo_RF_2
1	n_estimators	100	300
2	min_samples_leaf	10	100
3	random_state	123	42
4	max_depth		5
5	min_samples_split	10	20
6	max_features	auto	sqrt
7	criterion	squared_error	squared_error
8	bootstrap	True	True
9	max_leaf_nodes		
10	min_weight_fraction_leaf	0.0	0.0
11	n_jobs		
12	oob_score	False	False
13	verbose	0	0
14	warm_start	False	False
15	ccp_alpha	0.0	0.0
16	max_samples		

Anexo A13 - Métricas bases calculadas para o modelo LASSO.

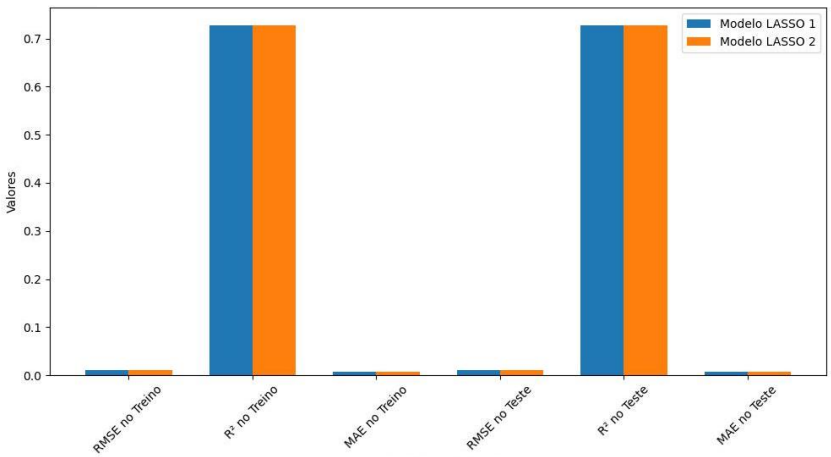


Figura 9 - Representação gráfica do resultado do modelo LASSO 1 e 2

Anexo 14 - Gráfico comparativo dos resultados dos modelos RF 1 e LASSO 1.