

## **Precificação Automatizada de Imóveis com Random Forest e Regressão Linear**

Orlando Santos de Oliveira<sup>1</sup>; Igor Pinheiro de Araújo Costa<sup>2</sup>

<sup>1</sup> Cientista da Computação. Rua Munhoz de Melo, 182 – Jardim Paraíso; 07143-450 Guarulhos, SP, Brasil  
autor correspondente

<sup>2</sup> Centro de Análises de Sistemas Navais (CASNAV). Doutorando em Engenharia de Produção, com ênfase em  
Pesquisa Operacional e Ciência de Dados. Gerente de Projetos no CASNAV. Praça Barão de Ladário s/n - Ilha  
das Cobras - Rua da Ponte, Ed. 23 do AMRJ - Centro; 20091-000 Rio de Janeiro, RJ, Brasil

\*autor correspondente: [orlando23030@gmail.com](mailto:orlando23030@gmail.com)

## Precificação Automatizada de Imóveis com Machine Learning

### Resumo

O mercado imobiliário brasileiro apresenta elevada heterogeneidade de preços, influenciada por fatores estruturais, locacionais e socioeconômicos. Nesse contexto, a determinação precisa do valor dos imóveis constitui um desafio relevante para agentes do mercado, instituições financeiras e órgãos públicos, uma vez que avaliações imprecisas podem gerar distorções econômicas, riscos financeiros e impactos na arrecadação tributária. Tradicionalmente, a precificação imobiliária tem sido realizada por meio de métodos comparativos e modelos estatísticos convencionais, que podem apresentar limitações na identificação de relações complexas entre as variáveis envolvidas. Diante desse cenário, este trabalho tem como objetivo comparar modelos de precificação de imóveis utilizando técnicas de ciência de dados e “Machine Learning”, aplicadas ao município de Guarulhos, no estado de São Paulo. Para isso, foram integrados dados de anúncios imobiliários com indicadores demográficos provenientes do Censo Demográfico 2022 do Instituto Brasileiro de Geografia e Estatística [IBGE]. A metodologia adotada contemplou etapas de análise exploratória de dados, segmentação de mercado por meio do algoritmo K-Means e construção de modelos preditivos utilizando Regressão Linear e “Random Forest”. A pesquisa buscou analisar a contribuição de variáveis estruturais, locacionais e demográficas na formação dos preços dos imóveis, bem como comparar o desempenho das diferentes abordagens de modelagem. Os resultados revelam que atributos físicos (como área útil e vagas de garagem), combinados com o adensamento populacional urbano, são os principais impulsionadores do valor dos imóveis na região. Além disso, o estudo evidencia que abordagens flexíveis e não lineares lidam melhor com a assimetria típica desse mercado, gerando estimativas de preço mais precisas. Dessa forma, os achados entregam insights práticos que permitem a agentes financeiros, corretores e gestores públicos realizar avaliações mais robustas, mitigando riscos de crédito e otimizando a tomada de decisão estratégica.

**Palavras-chave:** Avaliação imobiliária; Ciência de dados; Machine Learning; Regressão Linear; Random Forest; Precificação de imóveis.

### Abstract

The Brazilian real estate market is characterized by high price heterogeneity, influenced by structural, locational, and socioeconomic factors. In this context, accurate property valuation represents a relevant challenge for market participants, financial institutions, and public agencies, as inaccurate assessments may lead to economic distortions, financial risks, and impacts on tax collection. Traditionally, real estate pricing has relied on comparative methods and conventional statistical models, which may present limitations in capturing complex relationships among the variables involved. Given this scenario, this study aims to compare property pricing models using data science and machine learning techniques applied to the municipality of Guarulhos, state of São Paulo. To this end, real estate listing data were integrated with demographic indicators from the 2022 Demographic Census conducted by the Brazilian Institute of Geography and Statistics [IBGE]. The adopted methodology included stages of exploratory data analysis, market segmentation using the K-Means algorithm, and the construction of predictive models using Linear Regression and Random Forest. The research sought to analyze the contribution of structural, locational, and demographic variables to property price formation, as well as to compare the performance of the different modeling approaches. The results reveal that physical attributes (such as usable area and parking spaces), combined with urban population density, are the main drivers of property values in the region. Furthermore, the study highlights that flexible and non-linear approaches better handle the typical asymmetry of this market, generating more accurate price estimates. Thus, the findings deliver practical insights that enable financial agents, real estate brokers, and

public managers to perform more robust valuations, mitigating credit risks and optimizing strategic decision-making.

**Keywords:** Real estate valuation; Data science; Machine learning; Linear Regression; Random Forest; Property pricing.

## Introdução

O mercado imobiliário brasileiro é caracterizado por uma elevada diversidade de preços, resultante da interação complexa entre fatores estruturais, características socioeconômicas e aspectos locacionais (Silveira, 2023).

A avaliação imobiliária imprecisa desencadeia impactos sistêmicos que transcendem a transação individual. Segundo Dantas (2003), falhas na estimativa de valor venal geram ineficiências de mercado que prejudicam a liquidez e distorcem a percepção de risco dos ativos. No âmbito bancário, a NBR 14653-2 (ABNT, 2011) reforça que a precisão avaliatória é um requisito de segurança para a constituição de garantias hipotecárias, uma vez que avaliações superestimadas elevam o risco de crédito das instituições. Adicionalmente, para o setor público, a divergência entre o valor de mercado e o valor avaliado compromete a justiça fiscal em tributos como o [IPTU] e o [ITBI], afetando a arrecadação e a gestão do patrimônio público (ABNT, 2011).

Historicamente, a avaliação imobiliária tem se baseado em métodos comparativos de mercado e em modelos estatísticos tradicionais. Embora amplamente utilizados, esses métodos podem apresentar limitações na representação de relações não lineares e de interações complexas entre variáveis presentes em mercados urbanos heterogêneos. Em particular, modelos lineares assumem relações proporcionais entre variáveis explicativas e a variável resposta, hipótese que nem sempre é compatível com a dinâmica do mercado imobiliário. Assim, diferenças locacionais e características específicas dos imóveis podem produzir efeitos distintos sobre os preços, reduzindo a capacidade explicativa desses modelos em determinados contextos (Fávero e Belfiore, 2024).

Alinhada às metodologias de “Data Science” e “Analytics” [DSA], esta pesquisa utilizou a Análise Exploratória de Dados, agrupamento (K-Means) e modelagem preditiva (“Random Forest”). As técnicas possibilitaram o processamento, de forma automatizada e simultânea, das informações dos imóveis e os dados demográficos. Com isso, é possível identificar os diferentes nichos de mercado e descobrir quais características realmente influenciam o preço. Dessa forma, o trabalho reduz a avaliação subjetiva humana e baseia as suas conclusões em evidências extraídas diretamente dos dados.

Apesar da eficácia demonstrada dessas tecnologias, observa-se ainda uma lacuna na aplicação integrada desses métodos no contexto das cidades brasileiras. O diferencial desta

pesquisa reside na integração de dados granulares de anúncios imobiliários (variáveis estruturais e locacionais) com indicadores demográficos do Censo 2022 do IBGE — base selecionada por representar o levantamento oficial mais recente e consolidado a nível de setor censitário disponível no país. Diante deste cenário, o objetivo central deste trabalho é desenvolver e comparar o desempenho de modelos preditivos baseados em "Machine Learning" com a abordagem tradicional de Regressão Linear na precificação de imóveis no município de Guarulhos. Ao unir as características do "marketplace" com o perfil socioespacial da cidade, busca-se verificar se a flexibilidade dessas ferramentas algorítmicas é capaz de estimar os valores com maior precisão, oferecendo assim um suporte mais qualificado à tomada de decisão e contribuindo para a modernização das avaliações imobiliárias locais.

## Material e Métodos

A pesquisa caracteriza-se como aplicada, de abordagem quantitativa, com foco em modelagem estatística e "Machine Learning". Inicialmente, foram utilizadas duas bases principais: uma base demográfica oriunda do IBGE, com indicadores territoriais e socioeconômicos do município de Guarulhos, e uma base de anúncios imobiliários obtida por extração automatizada de dados em marketplace público. Após o tratamento e a padronização das informações, ambas foram integradas para compor a base analítica utilizada nas etapas de exploração, segmentação e modelagem preditiva.

No que diz respeito à vertente demográfica dessa base analítica, a caracterização socioeconômica baseou-se nos dados do Censo Demográfico 2022, realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE, 2022). Utilizou-se especificamente a malha de setores censitários (BR\_setores\_CD2022), que representa a menor unidade territorial de coleta de dados no Brasil, permitindo uma análise geoespacial de alta granularidade. Esta base permite integrar dados de densidade populacional e volume de domicílios aos imóveis, identificando fatores territoriais que influenciam a formação dos preços locais. A Tabela 1 apresenta a estrutura de colunas e atributos selecionados para a composição do estudo.

Tabela 1. Estrutura de dados e descrição das variáveis da base BR\_setores\_CD2022.

| Coluna    | Descrição Técnica                                         | Tipo de Dado      |
|-----------|-----------------------------------------------------------|-------------------|
| CD_SETOR  | Identificador numérico único do Setor Censitário.         | Numérico (String) |
| NM_MUN    | Nome oficial do município (Guarulhos).                    | Texto             |
| NM_BAIRRO | Denominação do bairro onde o setor está localizado.       | Texto             |
| v0001     | População residente total em cada setor censitário.       | Inteiro           |
| v0002     | Total de domicílios particulares permanentes recenseados. | Inteiro           |

|           |                                                               |                     |
|-----------|---------------------------------------------------------------|---------------------|
| AREA_KM2  | Extensão territorial da unidade em quilômetros quadrados.     | Decimal (Float)     |
| DENSIDADE | Razão entre a população total e a área territorial (Hab/km²). | Decimal (Calculada) |

Fonte: Elaborado pelo autor (2026), com base em dados do IBGE (2022).

De forma complementar aos dados demográficos territoriais, a segunda fonte de informações utilizada nesta pesquisa compreende as características estruturais e financeiras de imóveis disponíveis para venda no município de Guarulhos. Para a composição desta base, empregou-se a técnica de extração automatizada de dados (“web scraping”), utilizando robôs (“crawlers”) para raspar informações públicas de anúncios ativos em um dos principais marketplaces imobiliários da internet. Essa abordagem permitiu a coleta de uma amostra ampla e atualizada da realidade do mercado local. Após a etapa de extração, o conjunto de dados consolidado apresentou um total de 2.885 observações (anúncios únicos), estruturadas em 18 variáveis (colunas) que englobam aspectos locacionais, características construtivas e valores listados das propriedades (Tabela 2).

Tabela 2. Estrutura de dados e descrição das variáveis de imóveis à venda em Guarulhos.

| Nome da Variável | Descrição Técnica                                        | Tipo de Dado             |
|------------------|----------------------------------------------------------|--------------------------|
| id               | Identificador alfanumérico único do anúncio no portal.   | Texto (String)           |
| modalidade       | Tipo de transação imobiliária (ex.: Venda).              | Texto (String)           |
| estado           | Unidade Federativa onde o imóvel está localizado.        | Texto (String)           |
| cidade           | Município de localização do imóvel (Guarulhos).          | Texto (String)           |
| bairro           | Denominação do bairro onde a propriedade está situada.   | Texto (String)           |
| tipo_imovel      | Categoria estrutural do imóvel (ex.: Casa, Apartamento). | Texto (String)           |
| data             | Registro temporal da coleta do anúncio.                  | Texto (String)           |
| area_m2          | Metragem de área útil da propriedade (m²).               | Numérico (Inteiro)       |
| preco            | Valor total de venda oferecido no anúncio (R\$).         | Numérico (Decimal/Float) |
| preco_metro_2    | Valor absoluto cobrado por metro quadrado útil (R\$/m²). | Numérico (Decimal/Float) |
| faixa_preco_m2   | Variável categórica agrupando o preço do metro quadrado. | Texto (String)           |
| faixa_preco      | Variável categórica agrupando o valor total do imóvel.   | Texto (String)           |

|                 |                                                          |                    |
|-----------------|----------------------------------------------------------|--------------------|
| quartos         | Quantidade total de dormitórios disponíveis.             | Numérico (Inteiro) |
| faixa_quartos   | Variável categórica agrupando a quantidade de quartos.   | Texto (String)     |
| banheiros       | Quantidade total de banheiros na propriedade.            | Numérico (Inteiro) |
| faixa_banheiros | Variável categórica agrupando a quantidade de banheiros. | Texto (String)     |
| vagas           | Quantidade de vagas de garagem disponíveis.              | Numérico (Inteiro) |
| faixa_vagas     | Variável categórica agrupando o número de vagas.         | Texto (String)     |

Fonte: Elaborado pelo autor (2026), com base na extração de dados dos marketplaces imobiliários da internet.

### Tratamento e Padronização da Base Demográfica (IBGE)

A etapa inicial de tratamento da base de setores censitários do IBGE consistiu na aplicação de um recorte territorial específico para o município de Guarulhos, utilizando como critério de filtragem o código municipal 3518800. Este procedimento garantiu que apenas os dados pertinentes à área de estudo fossem processados, eliminando ruídos de outras localidades.

Após a delimitação geográfica, procedeu-se à seleção criteriosa de variáveis, mantendo-se apenas os atributos com relevância direta para o contexto da precificação imobiliária e análise socioeconômica. Para conferir maior clareza técnica ao conjunto de dados e facilitar a manipulação computacional, as variáveis originais do IBGE foram submetidas a uma padronização semântica (renomeação). Este processo transformou códigos técnicos e nomenclaturas genéricas em identificadores autoexplicativos, conforme detalhado na Tabela 3.

Tabela 3. Mapeamento e tipagem das variáveis demográficas pós-tratamento.

| Variável Original | Variável Renomeada | Descrição do Conteúdo                         | Tipo de Dado    |
|-------------------|--------------------|-----------------------------------------------|-----------------|
| NM_MUN            | municipio          | Nome do município da observação (Guarulhos).  | Object (String) |
| NM_BAIRRO         | bairro             | Nome do bairro referente ao setor censitário. | Object (String) |
| AREA_KM2          | area_km2           | Extensão territorial da unidade em km².       | Float64         |
| v0001             | populacao_total    | Quantidade de pessoas residentes no setor.    | Float64         |

v0002                      total\_domicilios                      Total de domicílios particulares permanentes.                      Float64

Fonte: Elaborado pelo autor (2026), com base em dados do IBGE (2022).

## Tratamento e Padronização da Base de Anúncios Imobiliários

O tratamento da base de anúncios teve como objetivo preparar os dados para o cruzamento com as informações do IBGE. Para garantir que as duas bases se conectassem corretamente, a coluna de bairros foi padronizada, com o ajuste de letras maiúsculas e minúsculas e a remoção de espaços em branco desnecessários.

A principal etapa de limpeza foi o filtro geográfico: foram mantidos na base apenas os anúncios de imóveis localizados em bairros presentes no censo do IBGE. Esse corte reduziu o volume inicial de 13.792 anúncios para uma amostra final de 2.885 registros válidos. Apesar da redução expressiva, essa medida foi essencial para eliminar dados inconsistentes e garantir que o modelo preditivo seja treinado apenas com localizações validadas.

Por fim, as colunas que não influenciavam diretamente a formação do preço foram excluídas, resultando em uma base final composta por 14 variáveis. A Tabela 4 apresenta o dicionário de dados gerado após esse pré-processamento.

Tabela 4. Dicionário de variáveis e estrutura da base de anúncios pós-tratamento.

| Coluna          | Descrição Técnica                                  | Tipo de Dado       |
|-----------------|----------------------------------------------------|--------------------|
| tipo_imovel     | Categoria da propriedade (Casa, Apartamento, etc.) | Categórica (Texto) |
| data            | Registro da data de coleta do anúncio              | Temporal (Texto)   |
| area_m2         | Área útil total da unidade em metros quadrados     | Numérica (Inteiro) |
| preco           | Valor total de venda anunciado (R\$)               | Numérica (Decimal) |
| preco_metro_2   | Valor calculado do metro quadrado (R\$/m²)         | Numérica (Decimal) |
| faixa_preco_m2  | Agrupamento categórico por nível de valor do m²    | Categórica (Texto) |
| faixa_preco     | Agrupamento categórico por faixas de preço total   | Categórica (Texto) |
| quartos         | Quantidade de dormitórios na unidade               | Numérica (Inteiro) |
| faixa_quartos   | Agrupamento categórico pelo número de quartos      | Categórica (Texto) |
| banheiros       | Quantidade total de banheiros na unidade           | Numérica (Inteiro) |
| faixa_banheiros | Agrupamento categórico pelo número de banheiros    | Categórica (Texto) |
| vagas           | Quantidade de vagas de garagem disponíveis         | Numérica (Inteiro) |
| faixa_vagas     | Agrupamento categórico pelo número de vagas        | Categórica (Texto) |
| bairro          | Nome do bairro padronizado para o cruzamento       | Categórica (Texto) |

Fonte: Elaborado pelo autor (2026), com base na extração de dados dos marketplaces imobiliários da internet.

## **Unificação e Pré-processamento dos Dados**

Esta etapa focou na unificação e no refinamento das informações necessárias para a modelagem. De acordo com Endel e Piringer (2015), o processo de "Data Wrangling" envolve atividades de transformação, limpeza, integração e organização dos dados, sendo especialmente relevante quando diferentes fontes de informação precisam ser combinadas em uma base única e consistente. Sob essa perspectiva, o processo foi iniciado com a importação e o cruzamento de duas fontes principais: a base demográfica, contendo indicadores socioeconômicos de 34 bairros de Guarulhos, e a base do mercado imobiliário, composta por 2.885 anúncios detalhados. Para viabilizar essa integração, os dados do Censo do IBGE, originalmente estruturados em nível de setor censitário, foram espacialmente agregados para o nível de bairro, consolidando as medidas que representam o contexto socioeconômico da região.

Na sequência, a base unificada foi submetida a um processo de limpeza e padronização, visando garantir a integridade estatística do experimento. Inicialmente, procedeu-se à exclusão de registros duplicados e de anúncios com informações nulas, mitigando o risco de enviesamento durante o treinamento do algoritmo. Posteriormente, realizou-se a adequação da tipagem de dados, convertendo variáveis críticas para o cálculo financeiro — como o preço e a área — para o formato numérico decimal, o que assegura a precisão das operações matemáticas. Por fim, aplicou-se o tratamento de "outliers" mediante a remoção do 1% superior da amostra (percentil 0,99) referente ao preço e à metragem. Conforme advertem Hair et al. (2009), a exclusão de observações com valores extremos é uma etapa fundamental do pré-processamento, uma vez que tais anomalias podem influenciar desproporcionalmente os parâmetros estatísticos e enviesar a capacidade preditiva do modelo. Portanto, essa decisão metodológica evitou que imóveis com características extremamente atípicas distorcessem as estimativas, assegurando que o modelo seja treinado com padrões mais representativos do mercado local.

## **Análise Exploratória dos Dados e Estatísticas Descritivas**

A base inicial continha 2.885 anúncios. Após as etapas de limpeza, que incluíram remoção de registros duplicados, exclusão de observações com dados incompletos e tratamento de "outliers", a base final utilizada na modelagem passou a contar com 2.823 observações.

Em termos estruturais, nota-se uma predominância de imóveis de médio porte. A área útil possui uma média de 98,8 m<sup>2</sup>, sendo que a mediana (percentil 50%) indica que metade

das propriedades possuem até 70 m<sup>2</sup>. A configuração mais comum encontrada apresenta entre 2 e 3 dormitórios (média de 2,5), acompanhada de 2 banheiros e pelo menos 1 vaga de garagem.

Sob a ótica financeira, o valor médio exigido nas negociações é de aproximadamente R\$520.756,00, com o preço do metro quadrado girando em torno de R\$6.033,00. É importante notar que o corte prévio de “outliers” funcionou de forma eficaz, mantendo o valor máximo da amostra em R\$ 1.500.000,00 e o limite estrutural em 316 m<sup>2</sup>, o que representa valores mais próximos da realidade do mercado de moradias padrão, sem distorções por imóveis de alto padrão.

Por fim, os indicadores demográficos revelam o cenário de adensamento urbano de Guarulhos. A densidade média nos bairros avaliados é de 17.417 habitantes por km<sup>2</sup>, com uma ocupação média de 2,46 pessoas por domicílio. O resumo estatístico completo das variáveis numéricas que alimentarão o modelo preditivo está organizado na Tabela 5.

Tabela 5. Estatísticas descritivas das principais variáveis do modelo.

| Variável                         | Média   | Desvio Padrão | Mínimo | Mediana (50%) | Máximo    |
|----------------------------------|---------|---------------|--------|---------------|-----------|
| Área (m <sup>2</sup> )           | 98,8    | 63,4          | 21,0   | 70,0          | 316,0     |
| Preço Total (R\$)                | 520.756 | 263.891       | 20.000 | 459.000       | 1.500.000 |
| Preço por m <sup>2</sup> (R\$)   | 6.033   | 2.432         | 177    | 5.833         | 22.143    |
| Quartos                          | 2,5     | 0,7           | 1,0    | 2,0           | 10,0      |
| Banheiros                        | 1,7     | 0,9           | 1,0    | 2,0           | 5,0       |
| Vagas de Garagem                 | 1,6     | 1,0           | 1,0    | 1,0           | 11,0      |
| Densidade (hab/km <sup>2</sup> ) | 17.417  | 14.584        | 6,5    | 16.959        | 37.540    |
| Pessoas por Domicílio            | 2,4     | 0,6           | 1,0    | 2,7           | 3,2       |

Fonte: Dados originais da pesquisa

## **Análise Visual e Comportamento das Variáveis**

Para aprofundar o entendimento dos dados além dos resumos estatísticos, conduziu-se uma análise visual utilizando recursos gráficos. Como fundamentado por Tukey (1977), o pioneiro da Análise Exploratória de Dados, a inspeção visual é uma etapa crítica que permite identificar padrões de distribuição, concentrações de valor e a relação direta entre os atributos dos imóveis com uma clareza que os resumos numéricos não conseguem prover sozinhos.

A primeira análise visual focou na variável alvo do modelo o preço de venda. Através da plotagem de um histograma com curva de densidade, observou-se detalhadamente o comportamento financeiro da amostra. O gráfico na Figura 1 revela uma nítida assimetria positiva, característica estrutural típica de dados imobiliários. Nota-se uma concentração

expressiva de imóveis na faixa entre R\$300.000 e R\$500.000, o que representa o volume principal do mercado padrão na região de Guarulhos.

À medida que o valor avança, a quantidade de propriedades ofertadas diminui drasticamente, criando uma longa cauda à direita. Embora os “outliers” mais extremos já tenham sido tratados na etapa de limpeza, a amostra final ainda contempla imóveis de alto padrão que alcançam patamares próximos a R\$1,4 milhão, mantendo a realidade de um mercado heterogêneo.

A assimetria observada na distribuição dos preços, conforme Figura 1, sugere que a variável alvo apresenta comportamento não normal, o que pode limitar o desempenho de modelos estritamente lineares e justificar a avaliação de algoritmos capazes de capturar relações mais complexas, como o “Random Forest”.

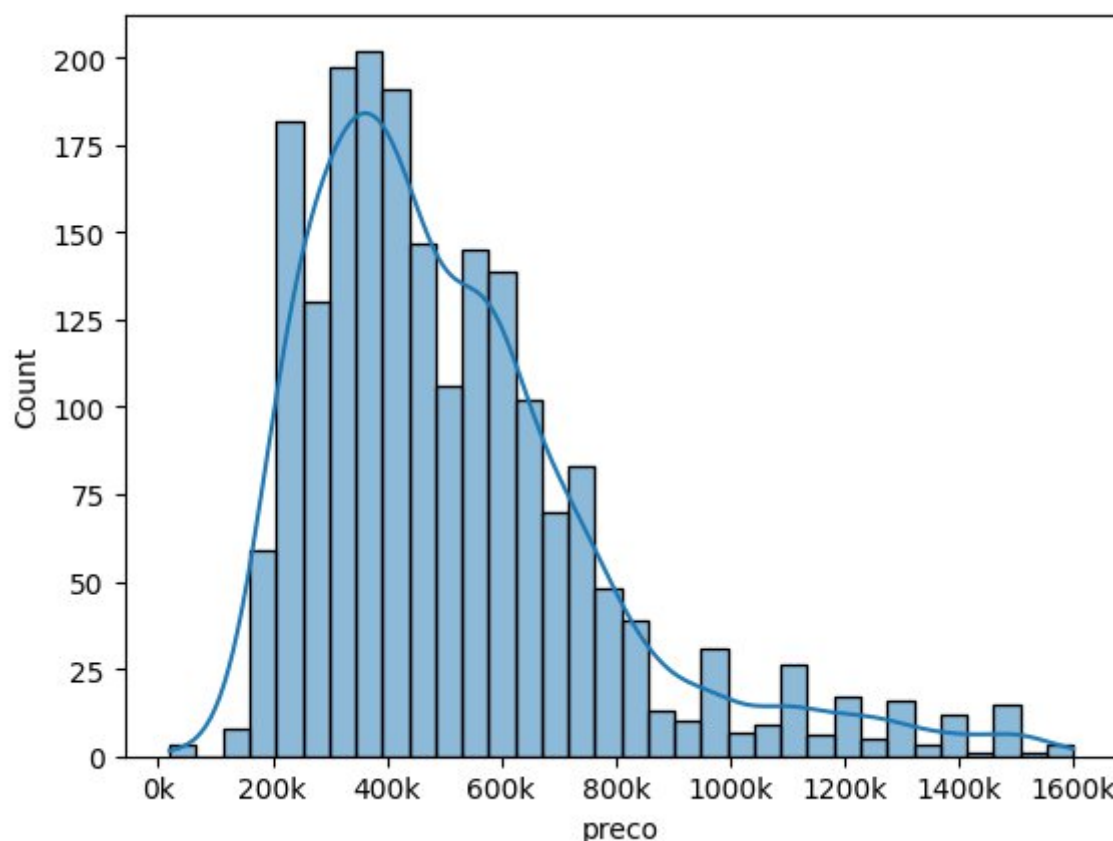


Figura 1. Distribuição de frequência dos preços dos imóveis em Guarulhos.

Fonte: Dados originais da pesquisa

A localização é um dos fatores que mais influenciam o valor dos imóveis. Por isso, a análise inicial focou na média de preços por bairro em Guarulhos. A Figura 2 apresenta os dez bairros com as maiores médias de preço. O bairro Água Azul lidera o ranking, com valor médio próximo a R\$1 milhão, acompanhado por regiões tradicionalmente valorizadas, como Itapegica, Vila Rio e Jardim Maia.

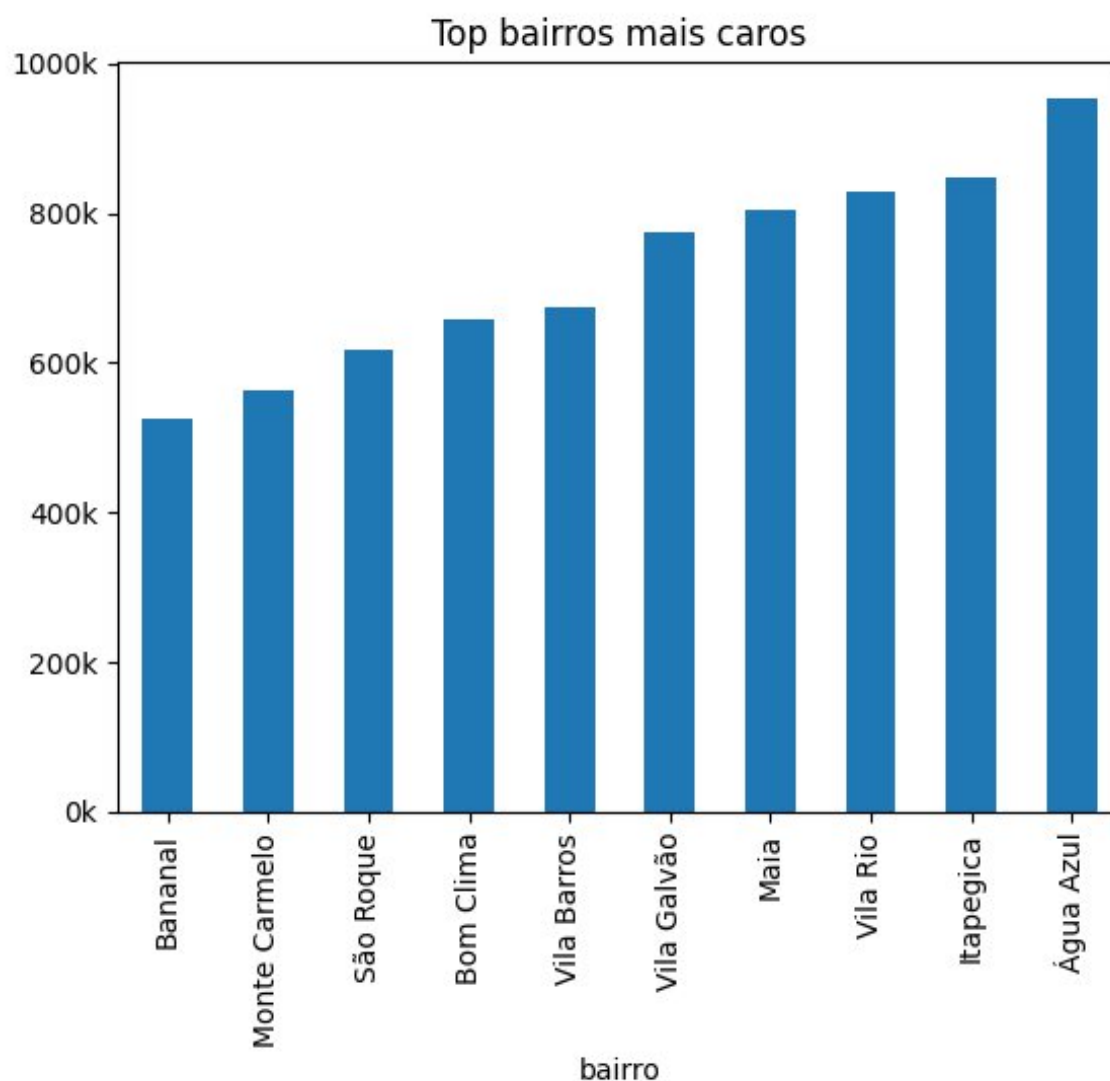


Figura 2. Top 10 bairros com o maior preço médio de oferta.

Fonte: Dados originais da pesquisa

No entanto, analisar apenas a média de preços pode levar a conclusões erradas se não considerarmos a quantidade de anúncios coletados. A Figura 3 mostra o número exato de imóveis disponíveis nessas mesmas dez regiões e revela uma distorção importante: o bairro Água Azul possui apenas 1 anúncio em toda a base.

Isso indica que a média da região foi puxada para cima por pouquíssimos imóveis de valor muito alto “outliers”. Logo, estatisticamente, não é correto afirmar que esse seja o bairro mais caro da cidade de forma geral. Por outro lado, bairros como Bom Clima e Vila Galvão possuem amostras grandes (88 e 173 anúncios), o que torna suas médias de preço muito mais confiáveis, conforme a Figura 3 abaixo.

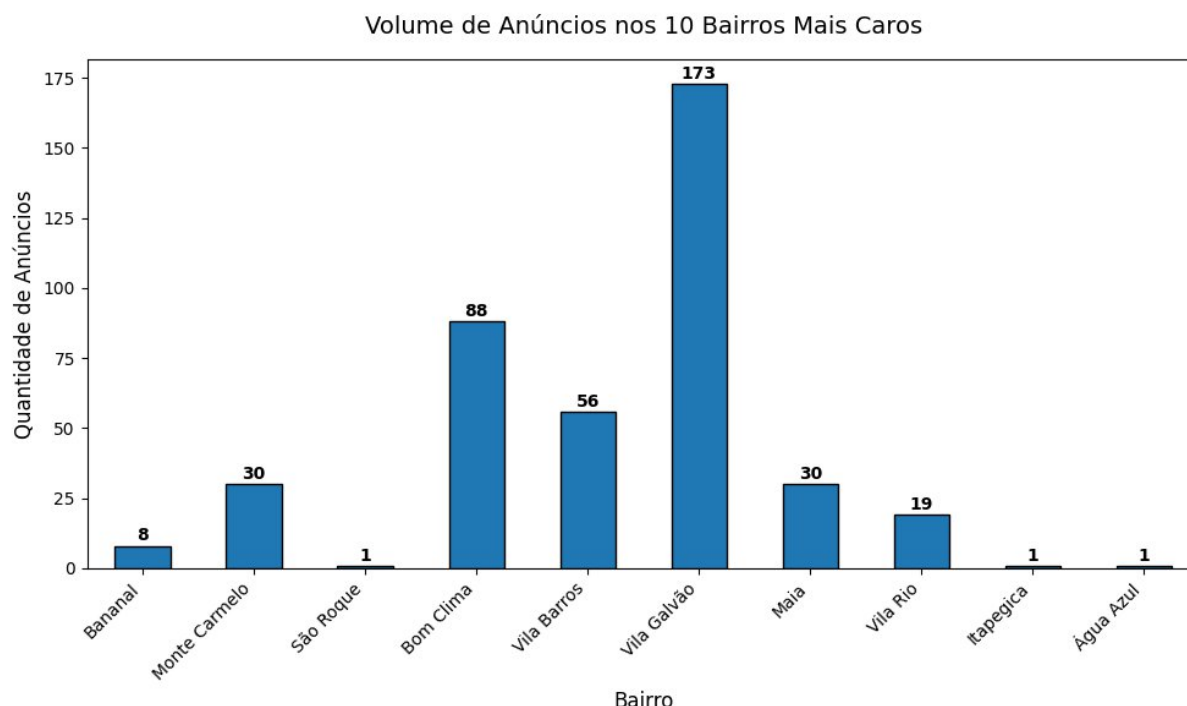


Figura 3. Volume de Anúncios nos 10 Bairros Mais Caros

Fonte: Dados originais da pesquisa

A descoberta dessa distorção nos dados justifica a escolha metodológica deste trabalho. Para evitar que o modelo seja enganado por locais com poucos anúncios ou preços fora do padrão, a previsão não se baseará apenas no nome do bairro. A solução adotada foi cruzar a localização com os dados censitários do IBGE (como densidade populacional e quantidade de domicílios). Essa técnica permite que o modelo de "Machine Learning" entenda a verdadeira estrutura socioeconômica da região, reduzindo o impacto de amostras pequenas e contribuindo para estimativas de preço mais consistentes.

A etapa final da exploração de dados consistiu na construção de uma Matriz de Correlação de Pearson (Figura 4), visualizada através de um mapa de calor. Utilizando uma escala gráfica de cores, mapeou-se a força e a direção da relação linear entre todas as variáveis numéricas da base. Esse mapeamento possui duas finalidades principais: identificar atributos com alto poder preditivo (diretamente ligados ao preço) e mitigar a multicolinearidade, mapeando variáveis explicativas que sejam redundantes ou excessivamente correlacionadas entre si, o que poderia desestabilizar a Regressão Linear nas próximas etapas.

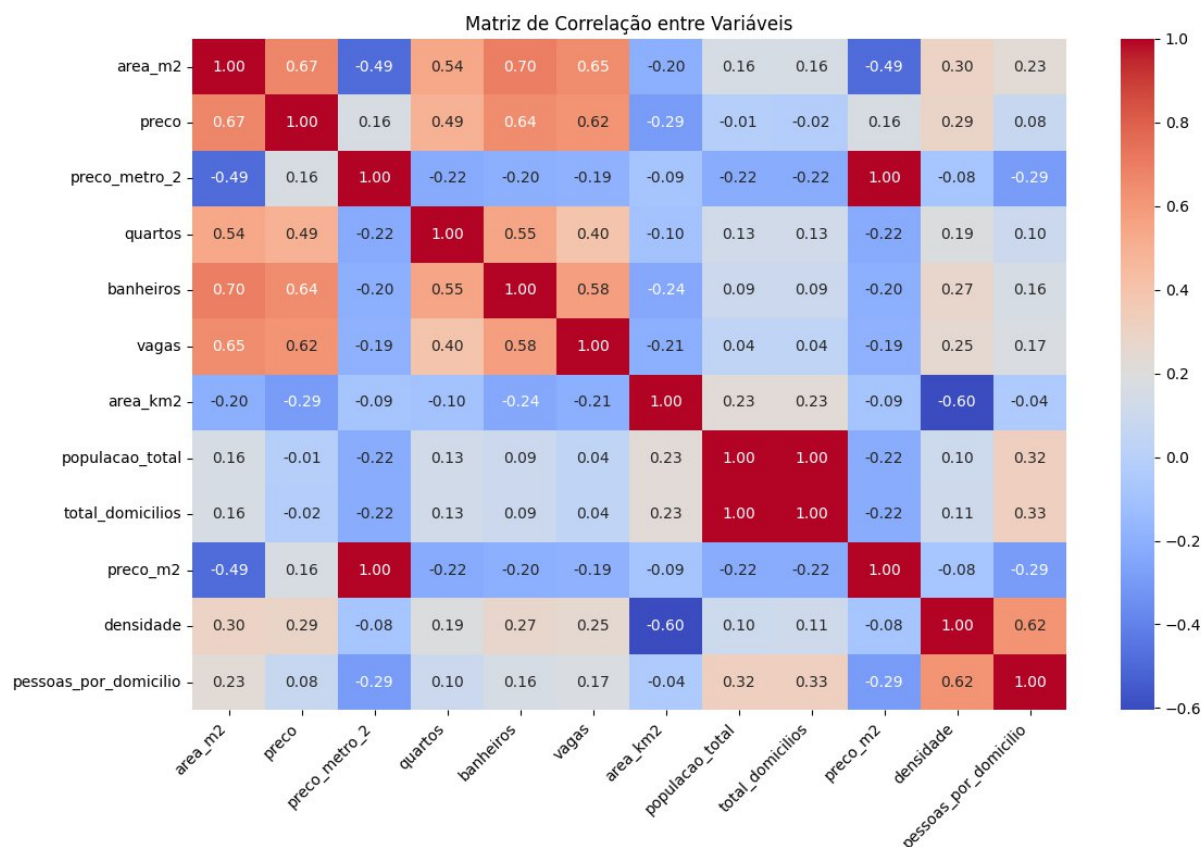


Figura 4. Matriz de correlação de Pearson entre as variáveis numéricas.

Fonte: Dados originais da pesquisa

A análise da matriz de correlação revelou padrões determinantes para a estruturação do modelo. Inicialmente, confirmou-se uma relação direta e expressiva entre o valor financeiro do imóvel e suas dimensões físicas, destacando-se a área útil (0,67), a quantidade de banheiros (0,64) e o número de vagas (0,62) como os atributos estruturais com maior correlação com a variável alvo. Esse comportamento valida a premissa econômica de que a expansão da metragem e o acréscimo de comodidades são os principais impulsionadores do valor de mercado. Além dos aspectos físicos, o cruzamento de dados evidenciou o impacto do efeito demográfico: a densidade populacional apresentou uma correlação positiva de 0,29 com o preço. Embora moderada, essa métrica indica que bairros mais adensados tendem a concentrar imóveis mais valorizados. Essa constatação ratifica a relevância do contexto urbano, reforçando a importância da inclusão dos dados do IBGE ao indicar que o entorno importa matematicamente para a formação do valor.

Por fim, no que tange à dinâmica espacial de Guarulhos, a matriz revelou relações inversas essenciais para o treinamento algorítmico, como a correlação negativa (-0,60) observada entre a extensão territorial do setor censitário e a densidade populacional. O mapeamento dessa particularidade é fundamental para a modelagem preditiva, pois permite

ao algoritmo compreender que áreas geograficamente mais extensas tendem a ser menos povoadas, ponderando o peso dessas variáveis de forma conjunta. Em decorrência disso, o modelo torna-se mais aderente à realidade local, evitando que o preço seja estimado de forma enviesada pela dimensão espacial da região, mas sim pela sua efetiva dinâmica de ocupação urbana.

### **Validação Exploratória Avançada (PCA e K-Means)**

O algoritmo K-Means foi empregado como técnica auxiliar de análise exploratória para a validação da estrutura dos dados. Fundamentando a abordagem da Análise de Componentes Principais (PCA), a Análise Fatorial é compreendida como uma técnica multivariada de natureza não preditiva, que visa simplificar a análise ao verificar a correlação entre as variáveis estudadas (Hair et al., 2009), o que contribui, segundo Fávero e Belfiore (2024), para avaliar a redundância e a consistência do conjunto de atributos. Paralelamente, utilizou-se a Análise de Cluster com o propósito de organizar as observações em subgrupos internamente homogêneos e externamente heterogêneos (James et al., 2023), permitindo a identificação de perfis distintos de imóveis presentes na amostra (Fávero e Belfiore, 2024). Embora tais técnicas não tenham sido incorporadas diretamente como entradas no modelo preditivo final, os seus resultados foram fundamentais para subsidiar a compreensão da heterogeneidade do mercado imobiliário analisado.

O PCA foi aplicado para avaliar a estrutura de correlação existente entre as variáveis e verificar se parte da informação presente na base poderia ser representada por um número reduzido de componentes principais, sem perda significativa de variabilidade. Como o estudo trabalha com sete variáveis diferentes (como área, quartos e densidade), existe sempre o risco de haver redundância. Os resultados do PCA, apresentados na Figura 5, indicam que grande parte da variabilidade dos dados pode ser representada por um número reduzido de componentes, sugerindo que as variáveis selecionadas apresentam estrutura informacional consistente e baixo grau de redundância.

A análise quantitativa dos componentes principais reforça essa interpretação. O primeiro componente principal (PC1) explicou 43,96% da variabilidade total dos dados, enquanto os dois primeiros componentes concentraram conjuntamente 63,97% da variância observada. Quando considerados os três primeiros componentes, a capacidade explicativa acumulada atingiu 78,14%, alcançando 86,95% com os quatro primeiros componentes. Esses resultados indicam que uma parcela significativa da informação presente nas variáveis originais pode ser representada por um número reduzido de componentes, evidenciando a

consistência estrutural da base de dados e a existência de correlações entre alguns atributos analisados.

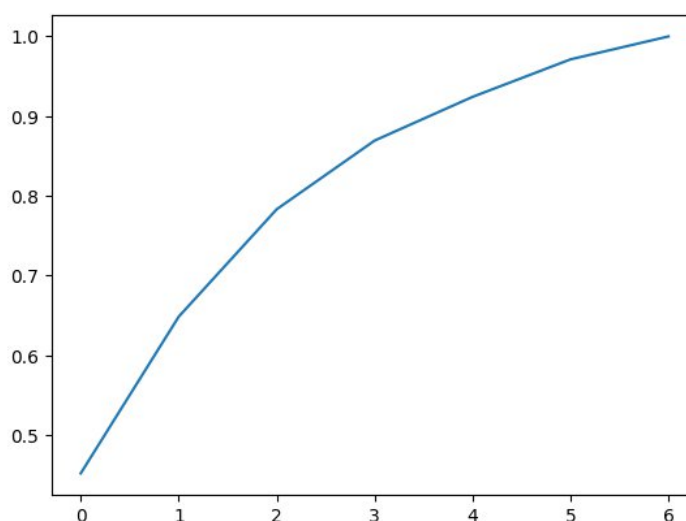


Figura 5. Variância explicada acumulada pelos componentes do PCA.

Fonte: Dados originais da pesquisa

Tabela 6. Variância explicada pelos componentes principais

| Componente Principal | Variância Explicada (%) | Variância Acumulada (%) |
|----------------------|-------------------------|-------------------------|
| PC1                  | 43,96                   | 43,96                   |
| PC2                  | 20,01                   | 63,97                   |
| PC3                  | 14,17                   | 78,14                   |
| PC4                  | 8,81                    | 86,95                   |
| PC5                  | 5,38                    | 92,33                   |
| PC6                  | 4,89                    | 97,22                   |
| PC7                  | 2,79                    | 100,00                  |

Fonte: Dados originais da pesquisa.

Para identificar os diferentes nichos do mercado, foi aplicado o algoritmo K-Means. No entanto, antes de agrupar os imóveis, foi necessário definir o número ideal de grupos através do Método do Cotovelo, analisando a Inércia.

A inércia mede o quão compactos estão os grupos: quanto menor o valor, mais parecidos são os imóveis dentro de cada grupo. Ao observar o gráfico Figura 6, percebe-se que após 3 clusters, a queda na inércia torna-se menos acentuada. Esse cotovelo na curva indica que dividir o mercado em três perfis é a estratégia mais eficiente para equilibrar a simplicidade do modelo com a precisão da segmentação.

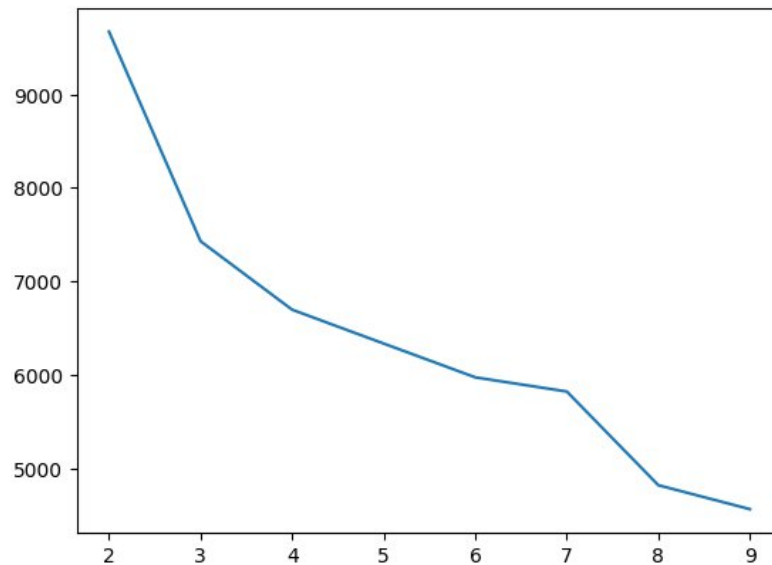


Figura 6. Método do Cotovelo para definição do número ideal de clusters (k).

Fonte: Dados originais da pesquisa

A análise da distribuição dos dados, quando observada sob a ótica das variáveis de Preço e Área útil, sugere a formação de grupos com comportamentos distintos dentro do mercado de Guarulhos. Conforme demonstrado na Figura 7, o processamento realizado pelo algoritmo de agrupamento resultou na identificação de três principais segmentos, delimitados a partir das semelhanças encontradas entre a dimensão física e o valor de mercado dos ativos analisados.

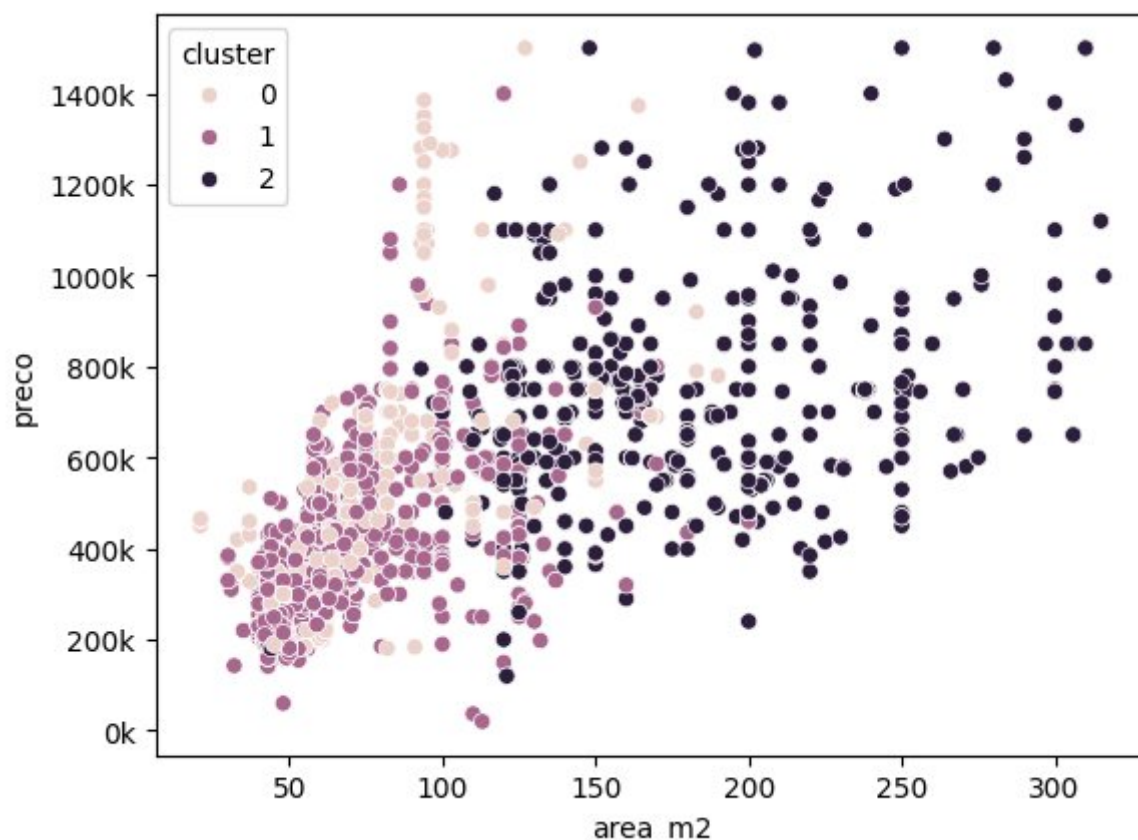


Figura 7. Segmentação do mercado imobiliário de Guarulhos em três perfis distintos.

Fonte: Dados originais da pesquisa

A partir do mapeamento apresentado na Figura 7, verifica-se que o Cluster 0 concentra o maior volume da amostra, caracterizando o segmento de médio padrão da cidade por apresentar uma relação equilibrada entre metragem e preço. Em uma faixa distinta, o Cluster 1 evidencia o perfil de entrada, composto por imóveis mais compactos e com valores mais acessíveis. Por fim, o Cluster 2 representa o segmento de alto padrão, abrangendo unidades com amplas áreas e preços elevados. Nota-se que este último grupo apresenta uma dispersão superior, o que sugere que, para imóveis de maior valor, variáveis além da metragem podem exercer influência mais significativa na composição do preço final.

O algoritmo K-Means apresentou um índice de Silhouette de 0,325. Embora esse valor não represente uma separação fortemente definida entre os grupos, ele indica uma estrutura de agrupamento aceitável para o contexto analisado. Em bases imobiliárias é comum que exista sobreposição entre segmentos de mercado, uma vez que imóveis com características semelhantes podem apresentar valores distintos em função de fatores locais, socioeconômicos e de características não observadas diretamente na base de dados.

Dessa forma, o resultado obtido sugere que os agrupamentos identificados possuem consistência estatística adequada e são capazes de representar diferentes perfis de imóveis

presentes no mercado imobiliário de Guarulhos, contribuindo para a validação exploratória da estrutura dos dados utilizada nesta pesquisa.

### **Escolha Metodológica: O Valor da Interpretabilidade**

Embora o PCA e o K-Means tenham sido fundamentais para validar a qualidade da base de dados, optou-se por manter as variáveis originais como metros quadrados e número de vagas para o treinamento final dos modelos de Regressão e "Random Forest".

Essa decisão prioriza a interpretabilidade, é muito mais transparente entender uma previsão de preço baseada em atributos físicos reais do que em componentes matemáticos abstratos. Tal escolha favorece a aplicabilidade dos resultados e a interlocução com os usuários finais.

### **Preparação dos Dados e Estratégia de Modelagem**

Nesta etapa, os dados consolidados foram organizados e segmentados para a construção dos modelos preditivos, visando a identificação de padrões com precisão e rigor estatístico. Conforme detalhado na Tabela 7, as decisões metodológicas adotadas — que incluem a seleção de atributos, a integração de bases e as técnicas de validação — fundamentam a robustez do experimento. A arquitetura da modelagem baseou-se na estruturação de variáveis preditoras (X), compostas por atributos estruturais dos imóveis integrados a indicadores demográficos do Censo 2022, permitindo que o sistema capture tanto as características físicas do ativo quanto o contexto socioeconômico da região, tendo como variável alvo (y) o preço de venda dos imóveis.

Visando garantir a capacidade de generalização do modelo e mitigar o risco de “overfitting” (sobreajuste), adotou-se a técnica de divisão amostral (“train-test split”). Sob essa configuração, 80% da base foi destinada ao ajuste dos parâmetros dos algoritmos, enquanto os 20% remanescentes foram reservados exclusivamente para a validação de dados inéditos. Estrategicamente, a proporção de teste foi fixada em 0,2 (`test_size=0.2`) e a divisão foi controlada por uma semente aleatória fixa (`random_state=42`), procedimento que, além de assegurar a transparência, garante a reprodutibilidade integral do experimento em futuras execuções da pesquisa.

Tabela 7. Justificativas metodológicas para a etapa de modelagem.

| Procedimento         | Decisão Adotada |           | Justificativa para a Pesquisa |                               |
|----------------------|-----------------|-----------|-------------------------------|-------------------------------|
| Seleção de Atributos | Variáveis       | Originais | Garante a                     | interpretabilidade do modelo, |

|                     |                                         |                                                                                                                               |
|---------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
|                     | (Físicas)                               | permitindo que o usuário entenda quais características físicas definem o preço.                                               |
| Integração de Dados | Base Híbrida<br>(Imóveis + IBGE)        | Permite que o modelo capture o impacto do entorno socioeconômico (densidade), indo além da área construída.                   |
| Validação do Modelo | Divisão<br>Treino/Teste (80/20)         | Garante a capacidade de generalização, provando que o modelo funciona para novos anúncios e não apenas para os dados antigos. |
| Confiabilidade      | Semente Fixa<br>( <i>Random State</i> ) | Assegura a reprodutibilidade, permitindo que outros pesquisadores cheguem aos mesmos resultados usando sua base.              |

Fonte: Elaborado pelo autor (2026)

## Aplicação e Resultados do Modelo de Regressão Linear

A etapa de modelagem preditiva iniciou-se com a aplicação do algoritmo de Regressão Linear. No contexto do "Machine Learning", este modelo clássico foi utilizado com o objetivo de estabelecer uma linha de base para a pesquisa, permitindo avaliar o comportamento inicial dos dados frente a uma abordagem linear.

Após o treinamento do algoritmo com a base de dados de treino (80%), as predições foram realizadas no conjunto de teste (20%) e o desempenho do modelo foi mensurado através de duas métricas estatísticas principais: o Coeficiente de Determinação [ $R^2$ ] e o Erro Médio Absoluto [MAE].

O resultado obtido para o  $R^2$  foi de 0,5565. Estatisticamente, isso indica que aproximadamente 55,6% da variabilidade dos preços dos imóveis em Guarulhos consegue ser explicada pelo modelo através das variáveis independentes selecionadas (área, dormitórios, vagas, densidade populacional, entre outras). Para um modelo de Regressão Linear aplicado ao mercado imobiliário, um setor complexo e influenciado por fatores subjetivos, como estado de conservação e acabamento, esse resultado indica capacidade explicativa moderada, pois indica que as variáveis estruturais e demográficas possuem poder preditivo real sobre o valor venal.

A segunda métrica avaliada foi o MAE, que registrou o valor de R\$119.783,35. Essa métrica traduz a margem de erro para valores monetários reais, indicando que, em média, as previsões geradas pelo modelo linear divergem do preço real ofertado no mercado em cerca de R\$119,7 mil. Considerando que a média de preço dos imóveis estudados gira em torno de R\$520 mil, esse erro representa um desvio de aproximadamente 23%, o que aponta limitações práticas para o uso isolado deste algoritmo em uma ferramenta de precisão.

Para complementar a análise estatística, elaborou-se o gráfico de dispersão entre os valores reais e as predições do modelo Figura 8. A linha diagonal tracejada representa o

cenário de (predição perfeita), onde o valor previsto seria exatamente igual ao valor real de mercado.

Ao observar a distribuição dos pontos na Figura 8, nota-se uma dispersão considerável ao longo da reta, especialmente conforme o preço dos imóveis aumenta. O modelo apresenta dificuldade em acompanhar a valorização de imóveis de alto padrão (acima de R\$800 mil), onde os pontos se afastam significativamente da linha de referência. Esse comportamento sugere que, para imóveis mais caros, o preço não cresce de forma linear com a metragem ou número de quartos, sendo influenciado por outros fatores que a Regressão Linear não consegue capturar com precisão.

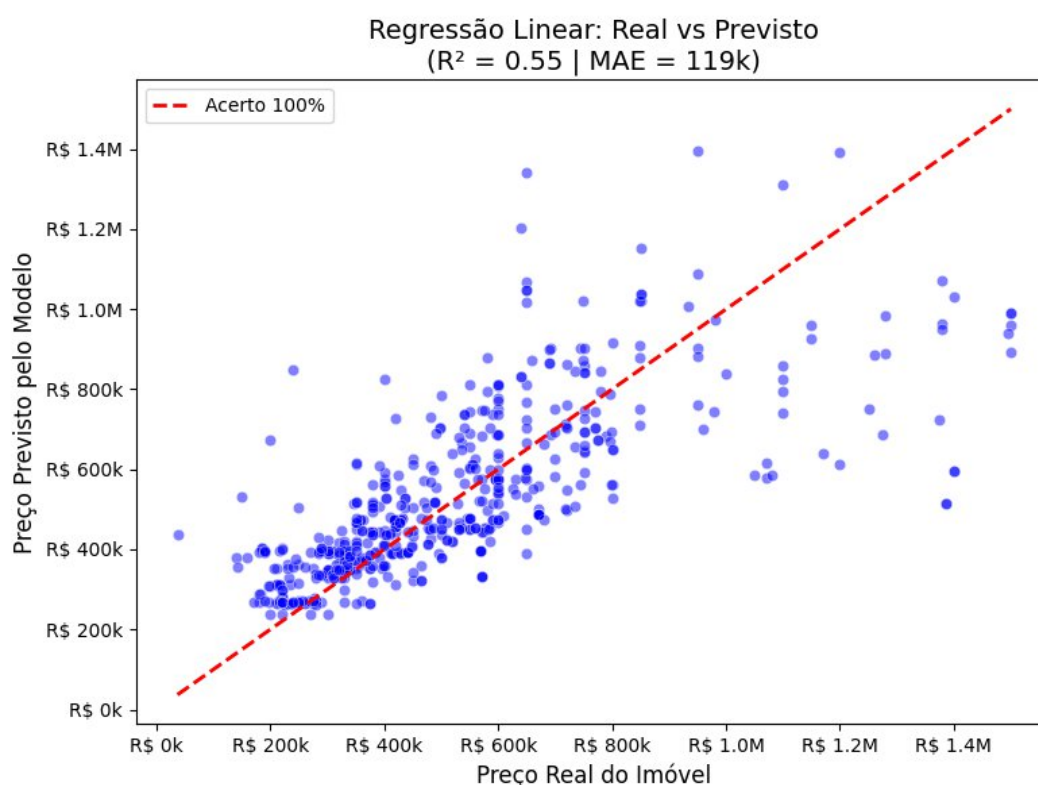


Figura 8. Dispersão entre Preços Reais e Previsões da Regressão Linear.

Fonte: Dados originais da pesquisa

Os resultados desta etapa sugerem que a precificação imobiliária em Guarulhos não pode ser adequadamente representada por um modelo estritamente linear. A dispersão visual observada e o valor do MAE justificam metodologicamente o avanço da pesquisa para a aplicação de algoritmos de aprendizado de máquina capazes de representar relações mais complexas entre as variáveis. Essa interpretação encontra respaldo na literatura de modelagem estatística e aprendizado de máquina. Segundo Fávero e Belfiore (2024), modelos lineares podem apresentar limitações quando os fenômenos analisados são influenciados por interações complexas e comportamentos não lineares entre as variáveis

explicativas. Nesse contexto, a relação entre características estruturais, fatores locais e aspectos socioeconômicos do mercado imobiliário tende a desafiar os pressupostos de linearidade assumidos pelos modelos de regressão tradicionais.

### **Aplicação e Desempenho do Algoritmo "Random Forest"**

Para superar as limitações de linearidade identificadas no modelo anterior, utilizado como linha de base, avançou-se para a aplicação do "Random Forest". Este algoritmo de "Machine Learning", que opera através da construção simultânea de múltiplas árvores de decisão, é amplamente reconhecido por sua capacidade de representar relações não lineares e interações complexas entre variáveis, características que modelos lineares nem sempre conseguem capturar adequadamente. Esse comportamento é compatível com os fundamentos teóricos dos métodos baseados em árvores de decisão. De acordo com James et al. (2023), algoritmos como o "Random Forest" possuem elevada flexibilidade para modelar relações não lineares e interações entre variáveis sem a necessidade de especificação prévia da forma funcional dessas relações. Essa característica torna tais algoritmos particularmente adequados para problemas preditivos complexos, nos quais múltiplos fatores atuam simultaneamente sobre a variável de interesse. A implementação deste modelo utilizou seis variáveis preditoras: área útil, dormitórios, banheiros, vagas de garagem, densidade populacional e densidade de moradores por domicílio.

Em termos de configuração metodológica, o modelo foi inicialmente parametrizado com 100 árvores de decisão, adotando-se o critério baseado no erro quadrático médio e uma semente aleatória fixa para garantir a reprodutibilidade. Nesta versão preliminar, optou-se por uma configuração inicial do algoritmo, sem um ajuste extensivo de hiperparâmetros, procedimento que poderá ser aprofundado em etapa posterior para a verificação de ganhos adicionais de desempenho. Ademais, a avaliação do modelo foi realizada por meio da estratégia "holdout", com a divisão da base em conjuntos de treinamento e teste. Embora esse procedimento permite uma primeira aferição do desempenho preditivo, reconhece-se que análises adicionais, como a validação cruzada, podem ser incorporadas na versão final da pesquisa para ampliar a robustez da avaliação.

Como resultado, o desempenho preditivo do modelo "Random Forest" apresentou uma melhora observada em relação à Regressão Linear. O Coeficiente de Determinação  $R^2$  atingiu a marca de 0,8566, o que significa que o algoritmo é capaz de explicar, com elevado poder explicativo, aproximadamente 85,6% da variabilidade dos preços em Guarulhos. O aumento do coeficiente de determinação ( $R^2$ ), de 55,65% na Regressão Linear para 85,66%

no modelo "Random Forest", fornece evidências de que a precificação imobiliária analisada é influenciada por relações não lineares entre as variáveis explicativas.

Esse resultado sugere que algoritmos baseados em árvores de decisão apresentam maior capacidade para modelar essas relações complexas, produzindo estimativas mais aderentes aos dados observados. Consequentemente, essa melhoria na capacidade de explicação refletiu-se na redução do erro financeiro. O Erro Médio Absoluto (MAE) caiu de R\$119.783,35 para R\$52.442,04. Considerando a complexidade do setor — onde fatores subjetivos, como o padrão de acabamento e o estado de conservação, influenciam expressivamente o valor —, um erro médio nesta faixa é considerado estatisticamente satisfatório. Esse resultado indica que o modelo alcançou um desempenho preditivo elevado, evidenciando seu potencial de aplicação como uma ferramenta prática de apoio à decisão.

Resultados semelhantes têm sido reportados em estudos empíricos que compararam modelos lineares e algoritmos de aprendizado de máquina em tarefas de avaliação imobiliária. Čeh et al. (2018), ao analisarem a previsão de preços de apartamentos na Eslovênia, observaram desempenho superior do "Random Forest" em relação à regressão múltipla tradicional, atribuindo esse resultado à capacidade do algoritmo de capturar padrões complexos presentes nos dados. De forma semelhante, Hong, Choi e Kim (2020) verificaram que modelos baseados em "Random Forest" produziram estimativas mais precisas em sistemas de avaliação em massa de imóveis residenciais na Coreia do Sul. Embora os mercados analisados possuam características distintas do contexto brasileiro, tais evidências reforçam a interpretação de que métodos não lineares tendem a oferecer vantagens em cenários de elevada heterogeneidade imobiliária.

A superioridade do "Random Forest" é evidenciada visualmente na Figura 9. Ao contrário da dispersão acentuada vista na Regressão Linear, os pontos neste gráfico apresentam uma forte concentração em torno da linha diagonal tracejada (o alvo de predição perfeita). Visualmente, nota-se maior proximidade entre os valores reais e previstos no modelo "Random Forest" em comparação à regressão linear, o que reforça o melhor ajuste já evidenciado pelas métricas quantitativas de desempenho.

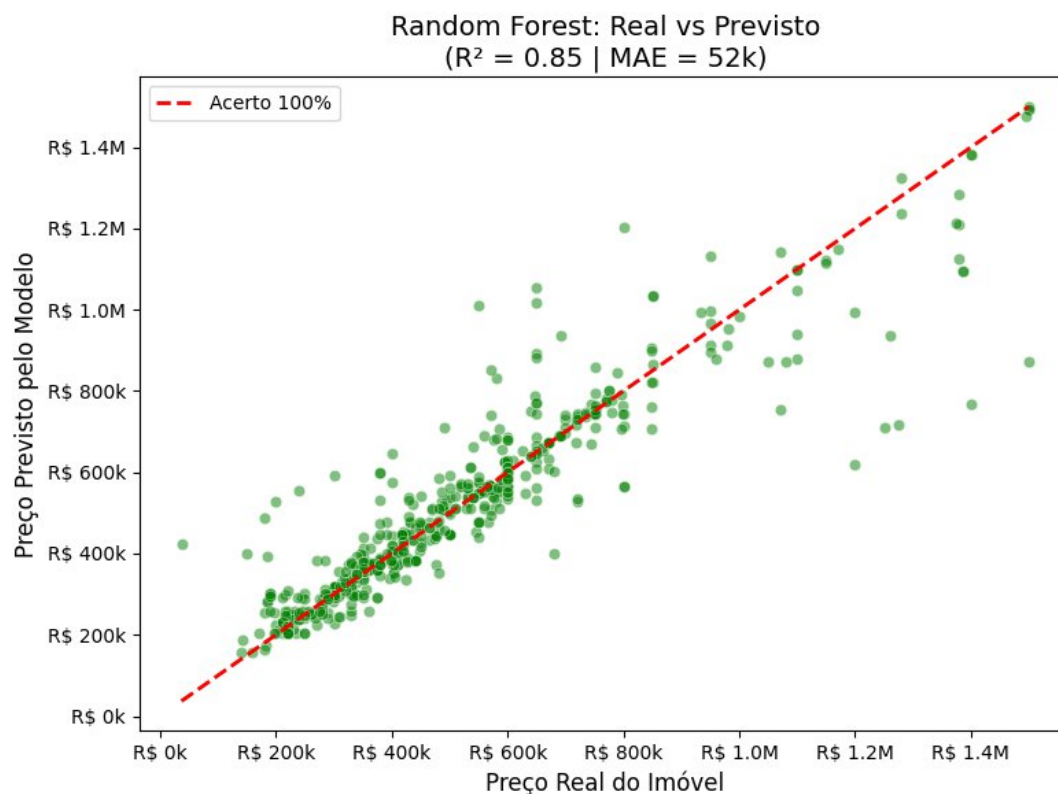


Figura 9. Dispersão entre Preços Reais e Previsões do "Random Forest"

Fonte: Dados originais da pesquisa

Sob uma perspectiva técnica, a eficácia do "Random Forest" justifica-se por dois pilares fundamentais de sua arquitetura. O primeiro é a sua capacidade de mapeamento não-linear, que permite ao algoritmo adaptar-se a diferentes curvas de valorização espacial, capturando que o preço por metro quadrado não é uma constante linear e varia significativamente entre os bairros. O segundo pilar é a sua elevada resiliência a "outliers". Como a estimativa final do modelo resulta do consenso de múltiplas árvores de decisão, essa abordagem diminui a influência de valores extremos nas previsões do algoritmo, garantindo maior estabilidade às previsões.

### Comparação de Desempenho e Escolha do Modelo Final

A Tabela 8 sintetiza o desempenho comparativo entre os modelos de Regressão Linear e "Random Forest", consolidando as métricas necessárias para a validação da hipótese central desta pesquisa.

Tabela 8. Comparação de performance preditiva entre os modelos avaliados.

| Modelo Algorítmico | $R^2$ (Poder de Explicação) | MAE (Erro Médio Absoluto) |
|--------------------|-----------------------------|---------------------------|
|--------------------|-----------------------------|---------------------------|

|                           |        |                |
|---------------------------|--------|----------------|
| Regressão Linear Clássica | 55,65% | R\$ 119.783,35 |
| "Random Forest" Regressor | 85,66% | R\$ 52.442,04  |

Fonte: Dados originais da pesquisa

Os resultados obtidos possuem implicações práticas relevantes para diferentes agentes do mercado imobiliário. Para imobiliárias e corretores, o modelo pode atuar como ferramenta de apoio à definição de preços mais aderentes às condições de mercado, contribuindo para reduzir a subjetividade presente nos processos de avaliação. Para investidores, a abordagem proposta pode auxiliar na identificação de ativos potencialmente subavaliados ou superavaliados, apoiando a tomada de decisão em operações de compra e venda. Adicionalmente, instituições financeiras podem utilizar modelos preditivos dessa natureza como mecanismo complementar nos processos de avaliação de garantias imobiliárias e análise de risco de crédito. Dessa forma, os resultados sugerem que a aplicação de técnicas de aprendizado de máquina possui potencial para ampliar a eficiência e a consistência dos processos de precificação imobiliária.

O algoritmo "Random Forest" ( $R^2 = 85,66\%$ ) apresentou desempenho preditivo superior ao da Regressão Linear Clássica ( $R^2 = 55,65\%$ ). O ganho de aproximadamente 30 pontos percentuais na explicação da variância sugere que relações não lineares desempenham papel relevante na precificação analisada imobiliária. Isso justifica a aplicação de algoritmos baseados em árvores de decisão, que se mostram mais adequados para mapear as interações complexas entre atributos físicos e locais do que as abordagens estatísticas clássicas.

A análise do Erro Médio Absoluto (MAE) destaca o impacto real da escolha metodológica. Enquanto a Regressão Linear registrou um Erro Médio Absoluto (MAE) de R\$119 mil, o "Random Forest" reduziu esse valor para R\$52,4 mil — uma queda de 56,2%. Considerando que o preço médio dos imóveis da amostra é de aproximadamente R\$500 mil, esse nível de erro limita a precisão das estimativas produzidas pelo modelo linear. Por outro lado, o desempenho do "Random Forest" mostra-se aderente às complexidades do setor, oferecendo uma margem de erro competitiva frente às subjetividades intrínsecas ao mercado imobiliário local.

Complementarmente, a análise de importância das variáveis permitiu identificar quais atributos exerceram maior influência sobre as previsões realizadas pelo modelo "Random Forest". Segundo Breiman (2001), a importância das variáveis nesse algoritmo pode ser estimada a partir da contribuição de cada atributo para a redução da impureza nos nós das árvores de decisão que compõem a floresta. Dessa forma, a métrica fornece uma indicação da relevância relativa de cada variável para o desempenho preditivo do modelo. Os resultados indicaram que a área do imóvel foi o fator mais relevante, representando aproximadamente

60,4% da importância total do modelo. Em seguida destacaram-se o número de vagas de garagem (13,4%), a densidade populacional da região (8,6%), o número médio de pessoas por domicílio (6,5%), a quantidade de banheiros (5,9%) e o número de quartos (5,2%).

Esses resultados sugerem que as características físicas dos imóveis exercem papel predominante na formação dos preços observados em Guarulhos. Entretanto, a contribuição das variáveis socioeconômicas demonstra que fatores relacionados ao contexto urbano e demográfico também influenciam o comportamento do mercado imobiliário local. Tal evidência reforça a importância da integração de diferentes fontes de dados na construção de modelos preditivos para avaliação imobiliária.

Tabela 9. Importância das Variáveis no Modelo "Random Forest"

| Variável               | Importância (%) |
|------------------------|-----------------|
| Área (m²)              | 60,44           |
| Vagas                  | 13,41           |
| Densidade Populacional | 8,64            |
| Pessoas por Domicílio  | 6,49            |
| Banheiros              | 5,86            |
| Quartos                | 5,17            |

Fonte: Dados originais da pesquisa.

De modo geral, os resultados obtidos indicam que os objetivos propostos foram alcançados, desde a estruturação da base de dados até o desenvolvimento dos modelos preditivos. A estratégia de integração dos anúncios imobiliários com a matriz censitária do IBGE permitiu a construção de uma base de dados robusta, na qual o processo de pré-processamento — especialmente a remoção de valores extremos por meio do corte em quantil (percentil 0,99) — preservou aproximadamente 98% da amostra original (2.823 imóveis), garantindo volume amostral adequado para o treinamento dos algoritmos. Os resultados obtidos sugerem a existência de padrões não lineares na precificação imobiliária de Guarulhos, uma vez que o modelo "Random Forest" apresentou desempenho superior ao modelo linear adotado como referência. Nesse contexto, enquanto a Regressão Linear estabeleceu uma linha de base estatística para comparação, a performance alcançada pelo "Random Forest" evidencia seu potencial como ferramenta de apoio à estimativa de preços no mercado imobiliário regional.

## **Contribuições Gerenciais e Implicações Práticas**

A presente pesquisa apresenta contribuições relevantes para diferentes agentes do mercado imobiliário ao demonstrar o potencial da aplicação de técnicas de aprendizado de máquina na precificação automatizada de imóveis. Os resultados obtidos evidenciam que modelos baseados em "Random Forest" podem constituir ferramentas de apoio à tomada de decisão, fornecendo estimativas de preços mais consistentes e reduzindo a subjetividade inerente aos processos tradicionais de avaliação imobiliária.

Sob a perspectiva gerencial, o modelo desenvolvido pode auxiliar imobiliárias, corretores e avaliadores na definição de preços mais aderentes às condições de mercado, contribuindo para negociações fundamentadas em evidências quantitativas. Da mesma forma, investidores podem utilizar as estimativas produzidas pelo modelo para identificar ativos potencialmente subavaliados ou superavaliados, apoiando decisões relacionadas à aquisição, venda ou composição de portfólios imobiliários.

Os resultados também possuem implicações para instituições financeiras, especialmente em atividades relacionadas à avaliação de garantias imobiliárias e à análise de risco de crédito. Embora o modelo não substitua a avaliação técnica realizada por profissionais especializados, sua utilização como ferramenta complementar pode aumentar a padronização das análises e reduzir a influência de fatores subjetivos na estimativa do valor dos imóveis.

Além disso, a análise de importância das variáveis permitiu identificar que atributos estruturais, como área útil e número de vagas de garagem, exercem maior influência sobre a formação dos preços, sem desconsiderar a contribuição de fatores demográficos, como densidade populacional e número médio de moradores por domicílio. Esses achados podem subsidiar construtoras, incorporadoras e gestores públicos na definição de estratégias de investimento, planejamento urbano e desenvolvimento de novos empreendimentos.

Por fim, destaca-se que a aplicação prática do modelo depende da disponibilidade de bases de dados atualizadas e representativas do mercado imobiliário. Dessa forma, recomenda-se que futuras implementações contem com processos automatizados de coleta, atualização e retreinamento periódico dos modelos, de modo a preservar sua capacidade preditiva diante das mudanças econômicas e das dinâmicas do mercado imobiliário.

## **Considerações Finais**

Este estudo teve como objetivo desenvolver um modelo de "Machine Learning" capaz de prever preços de imóveis em Guarulhos, integrando características estruturais e indicadores demográficos. Ao concluir a pesquisa, constata-se que os resultados indicam que os objetivos propostos foram alcançados, demonstrando que a inclusão de dados do IBGE

(como densidade populacional e perfil do domicílio) agrega uma camada essencial de contexto geográfico.

A comparação entre os algoritmos evidenciou que métodos baseados em árvores de decisão, como o "Random Forest", apresentaram desempenho superior na base de dados analisada para o setor imobiliário do que regressões lineares clássicas. A redução de 56% na margem de erro financeiro sugere que a tecnologia pode contribuir para a redução da subjetividade nas avaliações, oferecendo suporte técnico valioso para corretores, investidores e compradores.

Como limitações e oportunidades para trabalhos futuros, sugere-se a inclusão de variáveis qualitativas adicionais, como o padrão de acabamento interno e o estado de conservação das unidades, além da expansão da metodologia para outros municípios da Região Metropolitana de São Paulo. Este estudo utilizou a estratégia ("Hold-out") para a separação dos conjuntos de treinamento e teste, reservando 20% da amostra para validação. Embora essa estratégia seja amplamente empregada em aplicações práticas de aprendizado de máquina, ela pode apresentar sensibilidade à partição dos dados. Nesse contexto, estudos futuros poderão empregar técnicas de validação cruzada, como o K-Fold Cross Validation, a fim de reduzir a variabilidade das estimativas e obter avaliações mais robustas da capacidade de generalização dos modelos desenvolvidos. Em suma, o trabalho apresenta uma metodologia que alia rigor estatístico à aplicabilidade prática no contexto do mercado imobiliário.

## **Agradecimentos**

Agradeço à minha esposa, Michele Azevedo Rocha de Oliveira, pelo apoio, compreensão e incentivo ao longo desta jornada. Sua parceria foi fundamental para que eu pudesse concluir mais esta etapa da minha formação.

Aos meus filhos, Nicolas Rocha de Oliveira e Davi Rocha de Oliveira, por serem minha maior motivação para buscar crescimento pessoal e profissional. Cada conquista tem um significado ainda maior por poder compartilhá-la com vocês.

Ao meu orientador, Prof. Igor Pinheiro de Araújo Costa, agradeço pela orientação, disponibilidade e pelos conhecimentos compartilhados durante o desenvolvimento deste trabalho.

Por fim, agradeço a todos os professores, colegas e familiares que contribuíram, direta ou indiretamente, para a conclusão deste curso e para minha formação em Data Science e Analytics.

## Referências

Fávero, L.P.; Belfiore, P. 2024. Manual de Análise de Dados: Estatística e Machine Learning com Excel®, SPSS®, Stata®, R® e Python®. 2ed. GEN LTC, São Paulo, SP, Brasil.

Instituto Brasileiro de Geografia e Estatística [IBGE]. Censo Demográfico 2022: Malha de Setores Censitários. Rio de Janeiro: IBGE, 2022. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=41851>. Acesso em: 20 mar. 2026.

Silveira, U.C. 2023. Estudo de caso para avaliação de um imóvel urbano. Revista Científica Multidisciplinar Núcleo do Conhecimento 8(4): 22-35.

Associação Brasileira de Normas Técnicas [ABNT]. NBR 14653-2: Avaliação de bens — Parte 2: Imóveis urbanos. Rio de Janeiro: ABNT, 2011.

Dantas, R.A. 2003. Engenharia de avaliações: uma introdução à metodologia científica. Pini, São Paulo, SP, Brasil.

Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E.; Tatham, R.L. 2009. Análise Multivariada de Dados. 6. ed, Bookman, Porto Alegre, RS, Brasil. Disponível em: <[https://archive.org/details/kupdf.net\\_hair-j-f-anaacutelise-multivariada-de-dados-6ordf-ediccedilatildeopdf/page/n1/mode/1up?view=theater](https://archive.org/details/kupdf.net_hair-j-f-anaacutelise-multivariada-de-dados-6ordf-ediccedilatildeopdf/page/n1/mode/1up?view=theater)>. Acesso em: 8. 2026.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. 2023. An Introduction to Statistical Learning: with Applications in Python. Springer, Cham, Suíça.

Endel, F.; Piringer, H. Data Wrangling: making data useful again. 2015. IFAC-PapersOnLine 48(1): 111-112.

Tukey, J.W. 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA, EUA.

Čeh, M.; Kilibarda, M.; Lisec, A.; Bajat, B. 2018. Estimating the performance of random forest versus multiple regression for predicting prices of apartments. ISPRS International Journal of Geo-Information 7(5): 168.

Hong, J.; Choi, H.; Kim, W. 2020. A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management* 24(3): 140-152.

Breiman, L. 2001. Random Forests. *Machine Learning* 45(1): 5-32. DOI: 10.1023/A:1010933404324.