UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Guilherme Lourenção Brandt de Faria

Previsão de Preços Imobiliários com Machine Learning

São José do Rio Preto
2025

Guilherme Lourenção Brandt de Faria

Previsão de Preços Imobiliários com Machine Learning

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística (DCCE), do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Orientador: Profº. Drº. Rodrigo Capobianco Guido

São José do Rio Preto

2025

Faria, Guilherme
Previsão de Preços Imobiliários com Machine Learning / Guilherme Lourenção Brandt de Faria. -- São José do Rio Preto, 2025
61 p.
Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Rodrigo Capobianco Guido

1. Inteligência Artificial. 2. Mercado Imobiliário. 3. Previsão de Preços. I. Previsão de Preços Imobiliários com Machine Learning

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Guilherme Lourenção Brandt de Faria

Previsão de preços imobiliários com Machine Learning

Trabalho de Conclusão de Curso (TCC) apresentado como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação, junto ao Departamento de Ciências de Computação e Estatística (DCCE), do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Câmpus de São José do Rio Preto.

Comissão Examinadora

Prof. Dr. Rodrigo Capobianco Guido
UNESP – Câmpus de São José do Rio Preto
Orientador

Profª. Drª. Carina Alexandra Rondini
UNESP – Câmpus de São José do Rio Preto

Profª. Drª. Marilaine Colnago
UNESP – Câmpus de São José do Rio Preto

São José do Rio Preto
4 de Dezembro de 2025

Agradecimentos

Quero expressar a minha gratidão com todos que me apoiaram ao longo desses anos, minha mãe Alessandra e meus avós, Lindaura e Narcizo, sempre apoiaram meus estudos me incentivando sempre.

Agradeço a minha namorada, Maria Laura, por me ajudar no dia a dia e sempre estar do meu lado colaborando para que consigamos manter equilíbrio entre trabalho, estudos e lar.

Agradeço a meus amigos de graduação que me ajudaram nos estudos e em projetos, sempre compartilhando conhecimento.

Agradeço a todos os professores do curso, o conhecimento que obtive nesse período foi imensurável.

Obrigado aos meus 4 gatos: Keats, Lola, Charlie e Vaquinha por todo carinho e suporte emocional nesses últimos anos.

"Aprender sem pensar é trabalho perdido;
pensar sem aprender é perigoso."
Confúcio, Analectos 2:15

RESUMO

O início do século XXI foi um período marcado na história por grandes feitos na ciência, protagonizados por avanços significativos no campo da computação, com a evolução em diversos segmentos: processadores, GPUs, armazenamento, dados, entre outros. Esta evolução permitiu que alguns entraves no final do século passado fossem resolvidos, abrindo espaço para um novo cenário: os dados que antes eram escassos, se tornaram abundantes na era pós Big Data. Os dados em alta disponibilidade, somados com o conhecimento de negócio, provenientes da inteligência natural humana e o uso de inteligência artificial, gera predições de maior qualidade, reforçando a afirmação clássica na área de ciência de dados: “Dados de qualidade geram predições de qualidade”. Este trabalho traz a síntese dessa afirmação como propósito, utilizando técnicas de regressão para previsão de preços de imóveis na cidade de São José do Rio Preto usando atributos modernos como distância a pontos de interesse, uma abordagem que utiliza técnicas espaciais como modelagem de dados e a comparação de diferentes estratégias de remoção de outliers por clusters. Foi concluído que a melhor estratégia de remoção de outliers é a segmentação por latitude-longitude com 3 clusters e que modelos baseados em árvores (Random Forest e XGBoost) tiveram melhores performances frente a regressão linear, devido a natureza não-linear presente na correlação do preço com as principais variáveis preditoras. Também foi analisada a importância dos atributos nos 3 melhores resultados, apesar da dominância das características estruturais na estimativa dos preços, como número de suítes, banheiros e área útil, há também uma leve influência da distância a pontos de interesses nesse cálculo, especialmente de Shoppings, Hospitais, Farmácias, Faculdades e Escolas Particulares ou Públicas.

Palavras–chave: Inteligência Artificial, Ciência de Dados, Previsão de Preços, Aprendizado de Máquina, Clusterização, Mercado Imobiliário, Econometria Espacial.

ABSTRACT

The beginning of the 21st century was a period marked in history by great achievements in science, led by significant advances in the field of computing, with evolution in several segments: processors, GPUs, storage, data, among others. This evolution allowed some obstacles from the end of the last century to be solved, paving the way for a new scenario: data that was once scarce became abundant in the post-Big Data era. Readily available data, combined with business knowledge from natural human intelligence and the use of artificial intelligence, generates higher quality predictions, reinforcing the classic statement in the field of data science: "Quality data generates quality predictions". This work synthesizes this statement as its purpose, using regression techniques to predict real estate prices in the city of São José do Rio Preto using distance to points of interest, an approach that uses spatial techniques such as data modeling and the comparison of different outlier removal strategies by clusters. It was concluded that the best strategy for outlier removal is segmentation by latitude-longitude with 3 clusters and that tree-based models (Random Forest and XGBoost) had better performance compared to linear regression, due to the non-linear nature present in the correlation of price with the main predictor variables. The importance of attributes in the top 3 results was also analyzed; despite the dominance of structural characteristics in price estimation, such as the number of suites, bathrooms, and usable area, there is also a slight influence of the distance to points of interest in this calculation, especially to Shopping Malls, Hospitals, Pharmacies, Colleges, and Private or Public Schools.

Keywords: Artificial Intelligence, Data Science, Price Prediction, Machine Learning, Clustering, Real Estate Market, Spatial Econometrics.

Lista de Ilustrações

Lista de Abreviaturas e Siglas

IA – Inteligência Artificial

ML – Machine Learning / Aprendizado de Máquina

EDA / AED –Exploratory Data Analysis / Análise Exploratória dos Dados

KDD – Knowledge Discovery in Databases / Descoberta de Conhecimento em Banco de Dados

CRISP-DM – Cross-Industry Standard Process for Data Mining / Processo Padrão Inter-Indústrias para Mineração de Dados

IQR – Interquartile Range / Intervalo Interquartil

RMSE – Root Mean Squared Error/ Raiz do Erro Quadrático Médio

GPU – Graphics Processing Unit / Unidade de Processamento Gráfico

SAR – Spatial Autoregressive Model / Modelo Autoregressivo Espacial

SEM – Spatial Error Model / Modelo de Erro Espacial

MAE – Mean Absolute Error / Erro Médio Absoluto

MAPE – Mean Absolute Percentage Error / Erro Percentual Médio Absoluto

R^2 – Coeficiente de Determinação

IBGE – Instituto Brasileiro de Geografia e Estatística

IFDM – Índice FIRJAN de Desenvolvimento Municipal

URL – Uniform Resource Locator/ Alocador de Recursos Uniforme

Sumário

1. Introdução

O termo Inteligência Artificial, proposto por John McCarthy em 1956, define-se por máquinas com capacidade de realizar funções cognitivas, como o aprendizado e a solução de problemas que se assemelham à mente humana. Ou seja, a capacidade de aprender e associando padrões entre os dados e resultados. Contudo com as limitações da época, faltavam dados e capacidade de processamento, sendo assim a IA ficou até o final do século XX restrita a centros de pesquisas e grandes empresas de tecnologia.

Seu renascimento no início do século XXI foi possibilitado por uma junção de fatores: o aumento drástico do poder de processamento (Lei de Moore), o desenvolvimento de algoritmos mais eficientes (como o Deep Learning) e, não menos importante, a explosão na disponibilidade de dados, dando início ao big data.

Com o aumento significativo do volume e diversidade das fontes de dados, presenciamos o até então auge da era do big data, dispositivos que antes cumpriam somente funções primárias agora possuem funções em segundo plano no intuito de gerar dados para análise. Uma máquina de vender refrigerantes não só vende bebidas, ela também gera dados de estoque e vendas em tempo real, permitindo analisar os principais pontos de venda (hotspots), prever as máquinas que necessitam reabastecimento e traçar um trajeto otimizado para os repositores minimizarem o uso de combustível.

O principal desafio do Big Data reside não apenas no armazenamento, mas na extração de significado dessa massa de informação. Para transformar dados brutos em conhecimento útil, emprega-se o processo de Knowledge Discovery in Databases (KDD). O KDD é um processo não trivial de descoberta de padrões válidos, novos e potencialmente úteis a partir dos dados, por meio de técnicas que incluem mineração de dados, análise estatística e aprendizado de máquina.

1.1 Motivação e Escopo

Assim como todo o mercado, o segmento imobiliário passou por uma transformação nos últimos anos, através de mudanças intensas na integração de tecnologia, desde a etapa da obra ao planejamento. Construtoras e incorporadoras que movimentam dezenas bilhões anualmente estão buscando aprimorar seus processos com o setor de Business Intelligence.

No caso de São José do Rio Preto, uma cidade de tamanho médio que emerge como polo econômico, com alto crescimento populacional dos anos 2000 para 2010 cresceu sua população em 17,68%, de 2010 para 2022 cresceu 17%, com estimativa para atingir aproximadamente 504 mil habitantes (IBGE). Uma cidade média que cresceu expressivamente nos últimos 25 anos, da população de 357 mil em 2000 para estimados 504 mil em 2025, cresce aproximadamente 146 mil habitantes em 25 anos.

Esse crescimento populacional impacta diretamente a dinâmica urbana e o mercado imobiliário local, ampliando a demanda por habitação, infraestrutura e serviços. Estudos recentes apontam que, em diversas cidades brasileiras, o ritmo de crescimento das construções supera o crescimento populacional, fenômeno associado à financeirização do espaço urbano e à intensificação da especulação imobiliária, com a produção de imóveis voltada não apenas à ocupação, mas também à valorização patrimonial (WRI BRASIL, 2023). Nesse contexto, cidades médias em expansão, como São José do Rio Preto, tornam-se ambientes propícios à valorização imobiliária e à atração de investimentos, reforçando tais dinâmicas.

Além disso a cidade tem o 50º melhor IDH do brasil (IBGE, 2010) 0,797 um valor considerado alto, com tendência de melhoria nos últimos 15 anos, se analisarmos índices de desenvolvimentos mais atualizados como o IFDM - Índice FIRJAN de Desenvolvimento Municipal está em 8º lugar no país, com IFDM geral de 0,875.

A união de fatores como uma cidade com grande potencial e a versatilidade da inteligência artificial são as principais motivações.

1.2 Objetivo

A abordagem utilizada neste trabalho tem como objetivo prever os preços imobiliários da cidade de São José do Rio Preto com utilização de Inteligência Artificial e técnicas para modelagem de dados espaciais, trazendo o conhecimento de negócio sobre o setor imobiliário, a modelagem de dados proposta traz um esquema misto juntando dados hedônicos a proximidade de pontos de interesse (supermercados, shoppings, escolas, hospitais, entre outros).

Em resumo, busca-se coletar, tratar e integrar um conjunto de dados, aplicar e comparar o desempenho de algoritmos de aprendizado de máquina na tarefa de regressão.

1.3 Metodologia

A metodologia deste trabalho foi estruturada em etapas que refletem o fluxo completo de construção de um modelo preditivo aplicado ao mercado imobiliário. Inicialmente, realizou-se a coleta dos dados por meio de web scraping, contendo atributos estruturais dos imóveis e informações necessárias para geocodificação.

Em seguida, foi conduzido o entendimento dos dados, a verificação de nulos, padronização das escalas numéricas e remoção de outliers. Essa limpeza foi realizada aplicando o método do IQR dentro de clusters definidos pelo algoritmo K-Means, garantindo que valores extremos fossem detectados respeitando as particularidades de cada região da cidade. Na sequência, executou-se a feature engineering, obtendo informações espaciais sobre cada imóvel como frequência de pontos de interesses em raios específicos e a distância de pontos mais próximos.

Posteriormente, na etapa de modelagem foram utilizados os algoritmos Regressão Linear, Random Forest e XGBoost. Cada modelo foi avaliado por meio de validação cruzada e métricas como MAE, MAPE e R^2, permitindo identificar aqueles com melhor capacidade de generalização. Por fim, foi realizada uma análise complementar da importância das variáveis e da performance por cluster, possibilitando entender como características estruturais e espaciais contribuíram para a formação dos preços imobiliários.

1.4 Organização da Monografia

Este primeiro capítulo possui um teor introdutório, trazendo a motivação e designando um objetivo claro a ser alcançado com a realização do trabalho. Os demais capítulos desta monografia são compostos pelos seguintes conteúdos:

a) Capítulo 2 - Revisão Bibliográfica: São revisados conceitos de ciência de dados, Machine Learning e métodos aplicados ao setor imobiliário, bem como conceitos de modelagem de dados espaciais e técnicas de regressão. Também são apresentadas as definições das métricas de avaliação de desempenho dos modelos.

b) Capítulo 3 - Metodologia e Desenvolvimento: É apresentada a descrição do processo usado no desenvolvimento do trabalho, incluindo a base de dados, o pré-processamento e a implementação dos modelos.

c) Capítulo 4 - Validação: São apresentados os resultados obtidos com a aplicação dos modelos, a comparação de seu desempenho e a análise do impacto das variáveis espaciais na previsão.

d) Capítulo 5 - Conclusão: São expostas as conclusões obtidas a partir dos resultados e sugestões para trabalhos futuros.

2. Fundamentação Teórica

Neste capítulo, primeiramente iremos partir de uma visão macro sobre dados, envolvendo Big Data, o processo de KDD (Knowledge Discovery in Databases) e a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) para gerenciamento de projetos na ciência de dados, seguido por um aprofundamento progressivo em cada etapa dessa metodologia, abordando as técnicas específicas usadas por este experimento.

2.1 Big Data

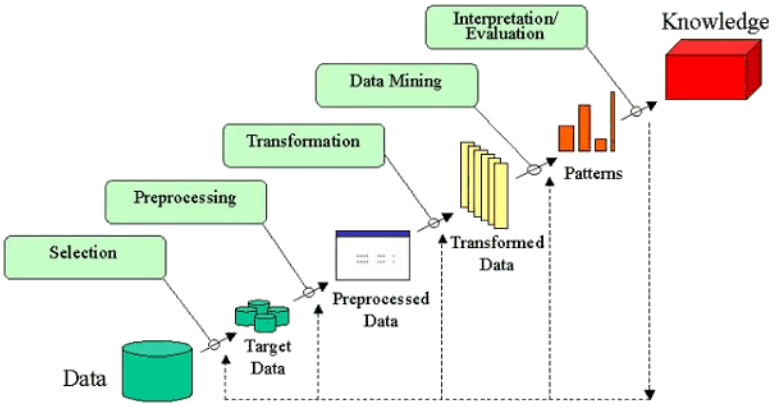
O termo big data foi idealizado em 1997 e se popularizou na primeira década dos anos 2000, pois este foi um período onde houve um aumento significativo no volume de dados disponíveis em sistemas de tecnologia. Atualmente, quase três décadas após a criação do termo Big Data, vivemos o atual ápice dessa tendência, que foi reforçada por maiores capacidades de memória e processamento.Um conceito amplamente difundido sobre Big Data são os 3 Vs do Big Data:

1. **Volume**: Enormes quantidades de informações disponíveis, como feeds de redes sociais (por exemplo, Instagram), registros de cliques em sites e aplicativos móveis, ou ainda dados de milhões de transações imobiliárias.
2. **Velocidade**: Equipamentos conectados online frequentemente operam em tempo real, com baixa latência, representando como a velocidade está presente nesta era.
3. **Variedade**: como dados estruturados (área, quartos) e não estruturados (imagens das propriedades e descrições em texto livre).

Diante desse contexto, foi necessária a criação de metodologias que extraem conhecimento válido dentro desse grande volume de dados, como Knowledge Discovery in Databases (KDD).

2.2 Knowledge Discovery in Databases (KDD)

Diagrama 1: Processo KDD



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

O KDD se refere ao processo em que é feito a detecção de padrões a partir dos dados, pode ser por meio de mineração de dados, análises estatísticas ou aprendizado de máquina. Por exemplo, na mineração de dados, algoritmos podem descobrir que propriedades com proximidade aos shoppings têm correlação com preços 15% mais altos.

As etapas que compõem o processo de KDD são as seguintes:

Seleção de dados: A etapa inicial do processo de KDD é a seleção de dados. Nessa etapa, poderá haver a necessidade de encontrar fontes de dados apropriadas, a depender de quão brutos os dados são, pode requerer engenharia de dados para garantir uma modelagem ideal nos dados.

Pré-processamento: Após a seleção dos dados, eles devem passar por um pré-processamento que garanta os dados estarem preparados para análises gráficas, estatísticas ou ML. Sendo necessário tratar valores ausentes, lidar com outliers, normalizar os dados, ajustar formatos de data e remover duplicatas.

Transformação de dados: Passando pelo pré-processamento, entramos na etapa de transformação onde ocorre a reestruturação dos dados, como agregação, redução de dimensionalidade e codificação das variáveis categóricas.

Mineração dos dados: A mineração é a etapa onde são aplicados os algoritmos para a descoberta dos padrões.

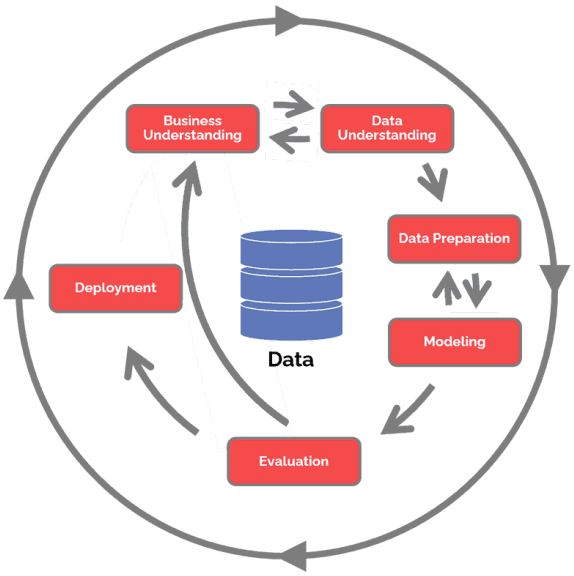
Avaliação e interpretação dos resultados: Etapa final em que os padrões e insights obtidos são avaliados e interpretados, a medição de significância estatística e o impacto da análise para resolução de problemas.

Embora o KDD estabeleça os fundamentos conceituais para descoberta de conhecimento em bases de dados, na prática adotou-se amplamente o CRISP-DM como metodologia operacional. Enquanto o KDD define o que é descoberta de conhecimento, o CRISP-DM detalha como implementar esse processo de forma estruturada e orientada a negócios.

2.3 Cross-Industry Standard Process for Data Mining (CRISP-DM)

A metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) tem se destacado por ser um método de gerenciamento de projetos para abordagens envolvendo ciência de dados, podendo também ser adaptado e utilizado na análise de dados.

Diagrama 2: Processo CRISP-DM



Fonte: Adaptado de Shearer (2000)

Ela inclui seis fases para direcionar os cientistas de dados durante todo o processo de descoberta a partir da base de dados. Estas etapas são:

Entendimento do negócio: Etapa inicial do projeto, ao invés de ir direto a parte prática é preciso entender o contexto, as dores e o objetivo do negócio, dominar as etapas dos processos que compõem o problema. Etapa em que é definido as metas e alinhá-las a necessidade estratégica do negócio

Entendimento dos dados: Estudar e coletar dados relevantes é crucial para o sucesso do projeto. Nesta fase entendemos os dados disponíveis, identificando lacunas e potenciais problemas, avaliando a qualidade dos dados.

Preparação dos dados: Os dados brutos raramente estão prontos para serem analisados. Nessa etapa é realizada a limpeza dos dados, tratamento de valores ausentes e integração entre fontes de dados diferentes, com o intuito de criar um conjunto de dados sólido para a resolução do problema.

Modelagem: Fase em que ocorre a aplicação das técnicas e algoritmos de aprendizado de máquina, selecionando a técnica mais adequada ao problema , como classificação, regressão ou clusterização.

Avaliação: Etapa onde os modelos são avaliados, buscando se melhorar as métricas para melhorar a qualidade dos modelos, com base nesta avaliação pode ser feito um ajuste de parâmetros para melhorar a performance

Implementação: Na fase final, o modelo é posto em ambiente de produção e integrado em um sistema onde gerará resultados consumíveis pelos usuários da aplicação, neste trabalho essa fase será adaptada em forma de análises dos resultados.

Os próximos tópicos seguirão a ordem das etapas do CRISP-DM, em cada um deles iremos aprofundar as teorias que embasaram este trabalho. Ele será iniciado com o entendimento de negócio e irá se concluir após o tópico de implementação.

2.4 Entendimento do negócio

2.4.1 Econometria Espacial na Precificação de Imóveis

A fundamentação teórica para previsão de preços de imóveis apoia-se historicamente no modelo hedônico, formalizado por Sherwin Rosen em 1974 em seu trabalho "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". O artigo afirma que o preço de um bem heterogêneo, como um imóvel, não é determinado por uma utilidade monolítica, mas sim pela soma dos valores implícitos de seus atributos individuais, que podem ser diferenciados em três categorias principais:

1. **Características Estruturais**: Metragem (área útil), número de quartos e banheiros, vagas de garagem, idade do imóvel, presença de comodidades como piscina ou academia.

2. **Características de Localização**: Bairro, proximidade a centros de negócios, shoppings, parques, estações de metrô/ônibus e qualidade de escolas.

3. **Características Ambientais**: Níveis de poluição sonora e atmosférica, índice de criminalidade e paisagismo urbano.

A aplicação prática do modelo hedônico é essencialmente econométrica. A forma funcional básica, geralmente linear, é expressa como:

$$P_i = \beta_0 + \beta_1 S_{1i} + \beta_2 S_{2i} + \ldots + \beta_n L_{ni} + \epsilon_i \quad (1)$$

Fonte: Adaptado de Rosen (1974)

Onde P_i é o preço do imóvel i, β_0 é o intercepto, S_{1i} e S_{2i} são as características estruturais, L_{ni} são as características de localização, β_1, β_2, …, β_n são os coeficientes a serem estimados, representando o preço hedônico marginal de cada atributo e ϵi é o termo de erro estocástico.

Contudo havia limitações neste modelo, especialmente em termos espaciais. Para contornar essa limitação, a econometria espacial desenvolveu um conjunto de modelos que expressam a dependência espacial. Dois modelos são considerados populares nessa área:

Modelo Autoregressivo Espacial (Spatial Autoregressive Model - SAR): Este modelo considera que o preço de um imóvel é diretamente influenciado pelos preços dos imóveis vizinhos. Sua formulação é:

$$P_i = \rho W P_j + \beta X_i + \epsilon_i \quad (2)$$

Fonte: Adaptado de Anselin (1988)

Onde ρ é o coeficiente de autocorrelação espacial , W é uma matriz de pesos espaciais que define a estrutura de vizinhança, WP_j representa o preço médio dos vizinhos. Um ρ positivo e significativo evidencia a existência de efeito de transbordamento (spillover) espacial nos preços.

Modelo de Erro Espacial (Spatial Error Model - SEM): Este modelo captura a dependência espacial através dos termos de erro, indicando que variáveis omitidas ou fatores não observados são espacialmente correlacionados.

$$P_i = \beta X_i + u_i, \qquad u_i = \lambda W u_j + \epsilon_i \quad (3)$$

Fonte: Adaptado de Anselin (1988)

No Modelo de Erro Espacial (SEM), a variável P é explicada pelas variáveis X por meio dos coeficientes β, enquanto o termo de erro u_i apresenta dependência espacial. Esse erro é modelado como $u_i = \lambda W u_j + \epsilon_i$, em que λ representa o grau de autocorrelação espacial, W é a matriz de pesos espaciais que define a vizinhança entre as unidades analisadas, e ε é o erro aleatório não correlacionado. Assim, o

SEM ajusta a regressão tradicional para capturar fatores não observados que se propagam espacialmente entre regiões próximas.

No contexto deste trabalho, a econometria espacial fornece a base teórica para entender por que características de localização como proximidade a shoppings, supermercados, hospitais e escolas devem ser incorporadas ao modelo de forma estruturada. A matriz de pesos espaciais (W) nos modelos SAR e SEM representa formalmente a mesma idéia que motiva a criação de variáveis baseadas em distâncias: que imóveis vizinhos compartilham características de localização semelhantes e, portanto, têm preços correlacionados, conceitos que serão explorados neste trabalho

2.4.2 Estado da Arte na Previsão de preços imobiliários

Com a disponibilidade de dados geoespaciais nas últimas décadas houve uma melhoria de qualidade na prática dos conceitos de econometria espacial. Pesquisas mais recentes passaram a incorporar variáveis de localização com uma granularidade e riqueza sem precedentes, superando as limitações dos modelos tradicionais como "bairro" ou "distância ao centro".

Zhao et al. (2022) propuseram o modelo PATE (Property, Amenities, Traffic and Emotions) para previsão de preços imobiliários em Pequim, incorporando variáveis espaciais que incluem a quantidade de comodidades urbanas em um raio de 1 km e a distância média até diferentes tipos de pontos de interesse, como transporte, educação, saúde e lazer. Essa abordagem demonstrou que a proximidade e densidade de comodidades urbanas exercem influência significativa sobre o valor dos imóveis.

De forma semelhante, Trindade Neves et al. (2024) analisaram dados abertos e aplicaram técnicas explicáveis de IA para modelar preços de imóveis considerando o número de pontos de interesse em raios de 75 m, 250 m e 500 m, além da distância até o ponto mais próximo em diversas categorias (comércio, cultura, saúde, parques, transporte, universidades, entre outros). O estudo reforça a importância de combinar variáveis espaciais em diversas escalas com métodos de aprendizado de máquina para aumentar a precisão das previsões.

Outros trabalhos também reforçam a importância da densidade e proximidade de pontos de interesse na explicação dos preços imobiliários. Cellmer (2023), ao analisar a influência de diferentes categorias de pontos de interesse sobre valores imobiliários, observou que áreas com maior densidade de comodidades urbanas tendem a apresentar preços significativamente mais altos, embora o impacto varia conforme o tipo de ponto de interesse.

Já o estudo de Cellmer, Bełej e Trojanek (2024), conduzido em três cidades polonesas, confirma empiricamente que tanto a densidade quanto a distância a pontos de interesse representam determinantes estatisticamente significativos dos preços de moradias. Os autores destacam que modelos que incorporam variáveis espaciais detalhadas apresentam melhor desempenho preditivo que regressões hedônicas tradicionais baseadas apenas em características internas dos imóveis.

Em conjunto, esses estudos demonstram que para além do simples uso da geolocalização, a composição da localização, entendida como o conjunto de relações espaciais entre o imóvel e o ambiente urbano, constituem elementos importantes na previsão de preços imobiliários. Abordagens que integram variáveis espaciais multiescalares, densidade de pontos de interesse e distâncias específicas tendem a produzir modelos mais precisos, interpretáveis e alinhados com a complexidade urbana contemporânea.

2.5 Entendimento dos dados

2.5.1 Fonte

O web scraping, ou coleta automatizada de dados da web, é uma metodologia consolidada para aquisição de dados em larga escala quando fontes estruturadas tradicionalmente não se encontram disponíveis ao público. Na literatura aplicada ao mercado imobiliário, esta técnica tem sido empregada de forma sistemática para a construção de bases de dados em estudo sobre a precificação de imóveis urbanos.

Entretanto, a utilização desta abordagem apresenta limitações que devem ser consideradas na interpretação dos resultados. A qualidade dos dados coletados está diretamente vinculada à estabilidade e estrutura dos websites fonte, sendo

suscetível a alterações no layout das páginas que podem invalidar scripts de coleta previamente desenvolvidos.

Outra limitação significativa reside nas restrições técnicas impostas pelos provedores de conteúdo, incluindo limites de taxa de requisições e mecanismos de proteção contra acesso automatizado.

Apesar dessas limitações, quando executado com os devidos cuidados técnicos e éticos, o web scraping mantém-se como uma abordagem metodológica válida e amplamente adotada na literatura para a construção de bases de dado, permitindo a coleta sistemática de informações que seriam praticamente inacessíveis em escala por meio de métodos manuais.

2.5.2 Análise Exploratória dos Dados

A análise exploratória dos dados (AED) é uma etapa inicial essencial em qualquer projeto de ciência de dados, cujo objetivo é entender a estrutura, a qualidade e os padrões presentes no conjunto de dados antes da modelagem. Nesse processo, investigam-se aspectos como distribuições das variáveis, presença de valores ausentes, outliers, correlações e possíveis relações entre atributos. A AED também auxilia na identificação de inconsistências, na necessidade de transformações e na seleção preliminar de variáveis relevantes. Assim, funciona como uma etapa diagnóstica que orienta decisões antes do pré-processamento, feature engineering e escolha de modelos, garantindo maior coerência às análises posteriores.

2.6 Preparação de Dados

A qualidade e a consistência das informações utilizadas impactam diretamente o desempenho dos modelos, por isso se deve à essencialidade dessa etapa. Nessa fase, aplicam-se técnicas destinadas a tratar inconsistências, padronizar escalas e estruturar as variáveis de modo que os algoritmos de aprendizado de máquina possam extrair padrões de forma mais eficiente.

2.6.1 Tratamento de valores ausentes

Um primeiro passo consiste na remoção de valores nulos, que ocorrem quando determinadas observações não possuem registros em uma ou mais variáveis. Esses valores ausentes podem distorcer a análise e comprometer a aprendizagem do modelo. As estratégias de tratamento incluem tanto a exclusão dos registros incompletos, quando sua proporção é pequena, quanto a substituição dos valores faltantes por médias, medianas ou estimativas baseadas em correlações com outras variáveis.

2.6.2 Padronização dos dados

Após o tratamento de nulos, realiza-se a padronização dos dados por meio do método Standard Scaler, que transforma as variáveis numéricas para uma distribuição com média zero e desvio-padrão igual a um. Essa transformação é importante para algoritmos baseados em distância ou gradiente, pois evita que atributos com escalas maiores dominem o processo de aprendizagem.

2.6.3 Remoção de Outliers

2.6.3.1 Intervalo Interquartil (IQR)

Outra etapa relevante é a limpeza de outliers, ou valores extremos, que podem distorcer a distribuição das variáveis e afetar negativamente a performance dos modelos. Para essa tarefa, adota-se o método do intervalo interquartil (IQR – Interquartile Range), que identifica observações fora do intervalo definido por $[Q1-1.5\times IQR, Q3+1.5\times IQR]$ onde Q1 e Q3 representam os quartis inferior e superior da variável e $IQR = Q3 - Q1$.

2.6.3.2 K-Means na Remoção de Outliers

O K-Means é um método de aprendizado de máquina não supervisionado que busca particionar o conjunto de dados em k grupos, minimizando a soma das distâncias quadráticas entre os pontos e o centróide do grupo correspondente. Matematicamente, o objetivo é minimizar a função objetivo:

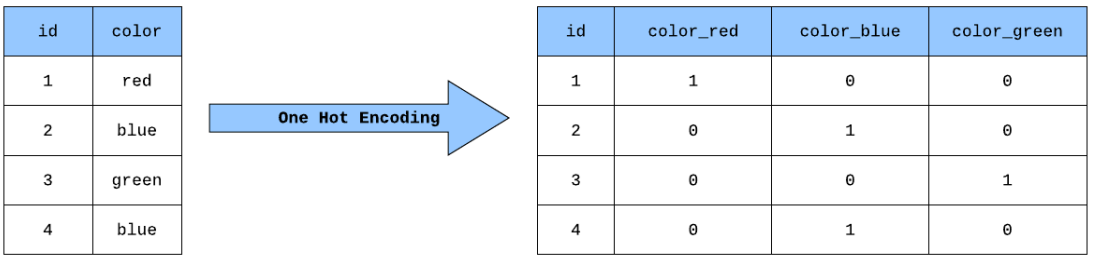
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (4)$$

Fonte: Adaptado de MacQueen (1967).

Onde $x_i^{(j)}$ representa o ponto de uma amostra i associada aquele cluster e c_j é o centróide do cluster j. Assim, o algoritmo permite agrupar imóveis de características semelhantes (como localização e preço) e, dentro de cada cluster, eliminar outliers de forma mais precisa e contextualizada, assim como será feito neste trabalho, ele foi usado exclusivamente na etapa de pré-processamento.

2.6.4 One Hot Encoder

Figura 1: One Hot Encoder



Fonte: Adaptado de scikit-learn (2010)

O One Hot Encoder, técnica de codificação de variáveis categóricas, como tipo de imóvel ou status da propriedade, em variáveis binárias (0 e 1). Essa transformação é importante para permitir que modelos de regressão e aprendizado de máquina tratem informações categóricas sem introduzir relações ordinais inexistentes entre as categorias.

2.7 Modelagem

Para a previsão dos preços imobiliários, foram testados três modelos: Regressão Linear, Random Forest e XGBoost. A escolha dessas abordagens visa comparar métodos lineares e não lineares, avaliando tanto a capacidade explicativa quanto o desempenho preditivo de cada um.

2.7.1 Regressão Linear

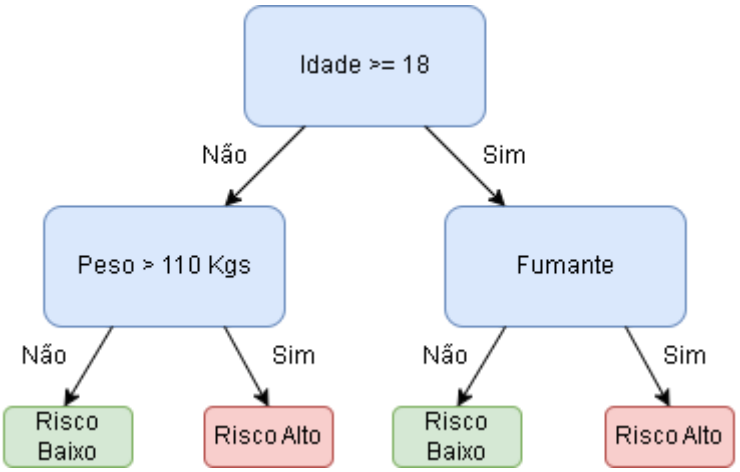
A Regressão Linear é um modelo estatístico que busca estabelecer uma relação linear entre uma variável dependente Y (no caso, o preço do imóvel) e um conjunto de variáveis independentes X_n, como número de dormitórios, área útil e localização. Seu objetivo é encontrar os coeficientes (β) que minimizam o erro quadrático médio entre os valores reais e os valores previstos e β_0 o intercepto. Esse modelo serve como uma referência inicial para avaliar ganhos de performance em algoritmos mais complexos.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \quad (5)$$

Fonte: Adaptado de Wooldridge (2010).

2.7.2 Árvore de Decisão

Figura 2: Estrutura de Árvore de Decisão



Fonte: Adaptado de Breiman et al. (1984).

O modelo de Árvore de Decisão é baseado em uma estrutura hierárquica de decisões, onde o conjunto de dados é recursivamente dividido em nós conforme critérios que maximizam a pureza das divisões (como o índice de Gini para classificação ou a redução da variância para regressão).

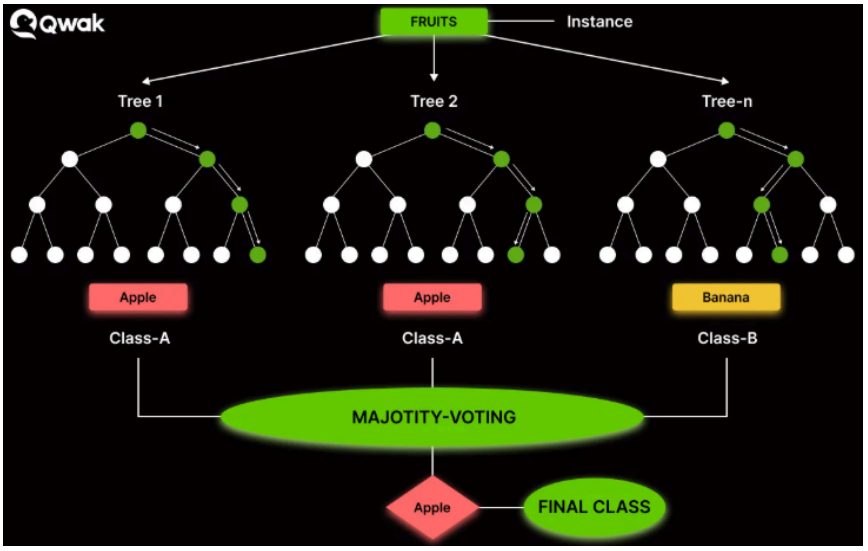
A árvore é composta por nós internos (que aplicam regras de divisão) e nós folha (que fornecem as predições finais). Cada nó interno divide os dados em nós

filhos com base em uma condição, criando caminhos específicos desde a raiz (primeiro nó) até as folhas (nós finais).

Por exemplo, uma árvore pode começar dividindo os imóveis pelo número de dormitórios, em seguida pela área útil e, por fim, pela presença de piscina, até atingir folhas que representam intervalos de preços estimados. Essa estrutura torna o modelo facilmente interpretável, pois permite visualizar o processo de decisão de forma sequencial.

2.7.3 Random Forest

Figura 3:Visão Geral do Algoritmo Random Forest.



Fonte:Adaptado de Breiman (2001).

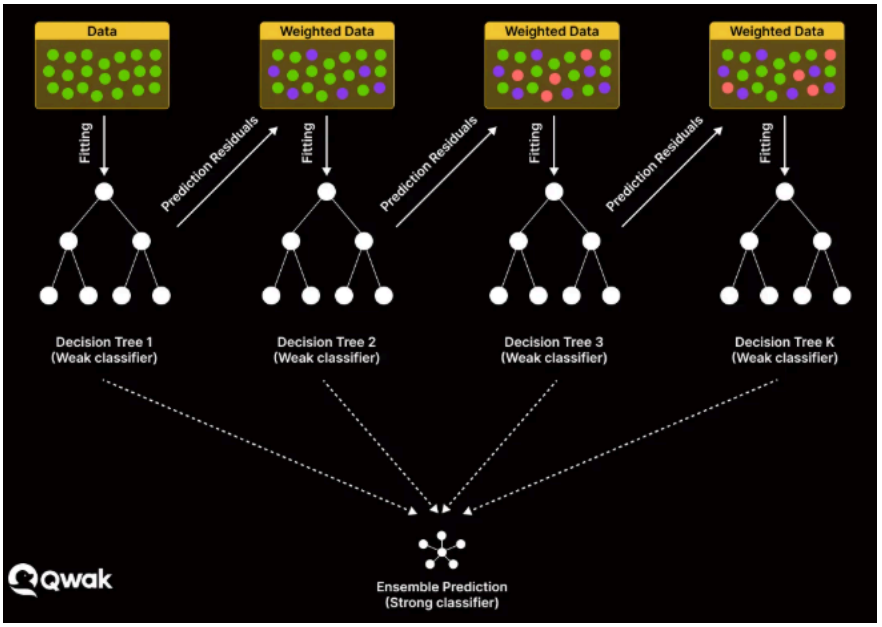
O Random Forest é um método de aprendizagem baseado em um conjunto de árvores de decisão. Cada árvore é treinada sobre uma amostra do dataset e com um subconjunto aleatório de variáveis em cada divisão, aumentando a diversidade entre elas. Após todas as árvores serem treinadas, suas previsões são combinadas por meio de uma votação, na tarefa de regressão, calcula-se a média das previsões individuais (como na equação 6), todas com o mesmo peso, já na classificação, utiliza-se a votação majoritária. Essa etapa de agregação ocorre somente após todas as árvores gerarem suas previsões, garantindo que o modelo final seja mais estável e apresente menor variância.

$$\hat{y}_{RF} = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t \qquad (6)$$

Fonte:Adaptado de Breiman (2001).

2.7.4 XGBoost (Extreme Gradient Boosting)

Figura 4: Visão Geral do Algoritmo XGBoost



Fonte: Adaptado de Chen e Guestrin (2016)

O XGBoost é um algoritmo de aprendizado de máquina baseado no método de boosting, no qual várias árvores de decisão são construídas de forma sequencial, cada uma voltada para corrigir os erros cometidos pelas árvores anteriores. Seu nome, Extreme Gradient Boosting, vem do fato de que o algoritmo utiliza uma forma avançada de gradiente descendente para ajustar sucessivamente as árvores, minimizando o erro de previsão a cada iteração. Esse processo transforma o aprendizado das árvores em uma otimização contínua: cada nova árvore é treinada para seguir a direção do gradiente, isto é, a direção em que o erro diminui mais rapidamente.

Ele se destaca por utilizar informações não apenas do gradiente, mas também da curvatura do erro, o que torna o ajuste mais estável e eficiente. Em vez de corrigir apenas a direção do erro, ele considera também o quanto esse erro está variando, o

que ajuda a construir modelos mais precisos e menos sensíveis a ruídos. Essa abordagem faz com que o treinamento avance em passos mais informados, aproximando-se da solução ótima com maior rapidez.

Outro ponto fundamental do XGBoost é a presença de regularização explícita. O algoritmo controla o tamanho e a complexidade das árvores e evita que as folhas assumam valores excessivamente altos, reduzindo assim o risco de overfitting. Essa regularização impede que o modelo se torne complexo demais e garanta melhor capacidade de generalização.

2.8 Validação

A fase de validação tem como objetivo principal avaliar a capacidade de generalização dos modelos desenvolvidos, assegurando que seu desempenho preditivo seja mantido em dados não vistos durante o treinamento. Para quantificar e comparar o desempenho dos algoritmos de regressão implementados foram empregadas três métricas fundamentais: Mean Absolute Error (MAE) e Mean Absolute Percentage Error (MAPE) e R2. Esta abordagem multidimensional permite uma análise complementar dos erros de predição sob diferentes perspectivas.

2.8.1 Validação Cruzada k-Fold

A validação cruzada k-fold é uma técnica utilizada para estimar a capacidade de generalização de um modelo, reduzindo o risco de overfitting. O conjunto de dados é dividido em k partes de tamanho similar; em cada iteração, uma delas é usada como conjunto de teste, enquanto as demais servem para treinamento. Ao final das k execuções, as métricas são agregadas, produzindo uma estimativa. Essa abordagem garante que todas as observações sejam testadas exatamente uma vez, proporcionando uma avaliação estável e menos sensível à uma divisão específica dos dados do que o treino–teste tradicional.

2.8.2 Erro Médio Absoluto (MAE)

O Erro Médio Absoluto representa a média das diferenças absolutas entre os valores previstos e observados, fornecendo uma medida direta e interpretável do erro médio do modelo. Sua formulação é expressa por:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right| \quad (7)$$

Fonte: Adaptado de Hyndman e Athanasopoulos (2018).

Onde y_i é o valor real, $\widehat{y}_i$ o valor previsto e n o número de observações. A principal vantagem do MAE está em sua interpretabilidade intuitiva, representando o erro médio em unidades monetárias (Reais), o que facilita sua compreensão dentro do contexto imobiliário. Diferentemente de métricas quadráticas como RMSE, o MAE é menos sensível a outliers, característica desejável em dados imobiliários onde valores extremos podem ocorrer naturalmente.

2.8.3 Erro Percentual Médio Absoluto (MAPE)

O Erro Percentual Médio Absoluto complementa a análise ao expressar o erro médio em termos percentuais, oferecendo uma perspectiva relativa do desempenho do modelo. É calculado conforme:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \quad (8)$$

Fonte: Adaptado de Hyndman e Athanasopoulos (2018).

Esta métrica é perfeita para o contexto imobiliário, pois permite avaliar a magnitude do erro em relação ao valor do imóvel, facilitando comparações entre diferentes faixas de preço e regiões. Por exemplo, um MAPE de 10%

indica que, em média, as previsões divergem em 10% dos valores reais, proporcionando uma medida intuitiva da acurácia do modelo.

2.8.4 Coeficiente de Determinação (R^2)

O coeficiente de determinação, ou R^2, mede a proporção da variância da variável dependente que é explicada pelo modelo. Ele varia entre 0 e 1, onde valores mais próximos de 1 indicam maior capacidade explicativa. Embora útil para avaliar o ajuste de modelos de regressão, o R^2 não indica diretamente a qualidade das previsões e pode ser enganoso em casos de overfitting.

2.9 Implementação como Análise e Disseminação

A fase de implementação do CRISP-DM será adaptada como um processo de análise, interpretação e disseminação dos resultados, ao invés de uma implantação operacional em ambiente de produção.

A implementação como análise é composta pela interpretação dos modelos preditivos desenvolvidos e da melhor estratégia de remoção de outliers, com ênfase na extração de insights sobre os determinantes de preços no mercado imobiliário de São José do Rio Preto. Quais as principais influências na valoração e como a proximidade a diferentes pontos de interesse impacta o preço.

3. Metodologia e Desenvolvimento

Neste capítulo, será descrita a estruturação do processo para o desenvolvimento deste experimento, etapa por etapa, seguindo o algoritmo implementado. Nele será explicado brevemente a etapa de engenharia de dados, por não ser o foco deste trabalho. O foco principal será no fluxo de ciência de dados, no qual terá início pela análise exploratória de dados, seguidos por pré-processamento, feature engineering e modelagem, deixando a etapa de validação para o capítulo subsequente.

Diagrama 3: Resumo do Algoritmo Implementado

Fonte: Próprio Autor (2025).

Neste trabalho, além dos algoritmos de machine learning, serão avaliados também algumas estratégias de remoção de outliers com base em clusters. Os clusters e a remoção ocorrerão apenas no conjunto de treino, evitando assim vazamento de dados e deixando a validação mais realista, com testes cegos. As estratégias são :

1. Todo o conjunto de treino X;
2. Latitude - Longitude;
3. Características estruturais;
4. Sem remoção de outliers;

O programa principal foi desenvolvido dentro da IDE online do Kaggle, plataforma competitiva de ciência de dados que permite várias horas de acesso a GPU, foi utilizada a configuração 2 × T4 GPUs.

3.1 Engenharia de Dados

O principal desafio deste experimento foi a disponibilidade dos dados. Não havia nenhum dataset pronto ao público de São José do Rio Preto, então foi necessário obtê-los por meio de Web Scraping.

As bibliotecas utilizadas para realização das extrações foram Selenium e BeautifulSoup e cada uma foi atribuída para a seguinte tarefa:

Selenium: abre o navegador, navega até a URL, realiza ações (como rolar a página, clicar em botões, aguardar o carregamento do conteúdo) e obtém o código-fonte.

Beautiful Soup: converte o código-fonte (em HTML ou XML) para uma árvore estruturada e navegável, extrai os elementos para construção do dataset.

Os dados raspados foram extraídos de 2 websites:

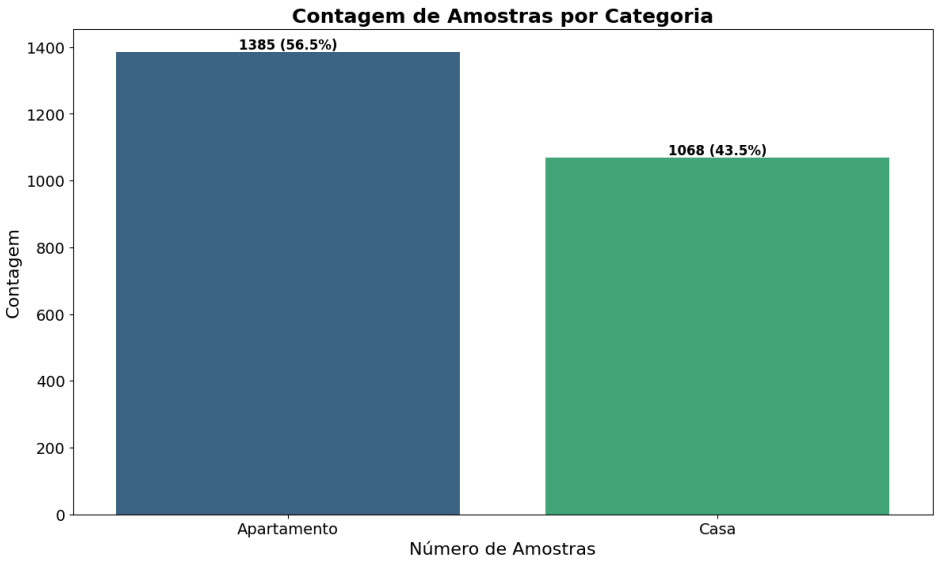
1. Compacto Imóveis para o dataset de imóveis: dados hedônicos sobre o imóvel (Quantidade de quartos, banheiros, garagem, área útil, piscina, condomínio, localização aproximada e preço.)
2. Google para o dataset dos amenities: dados de geolocalização de shoppings, supermercados, hospitais, farmácias, escolas públicas e particulares da cidade de São José do Rio Preto.(Categoria, Nome, Endereço, Latitude e Longitude)

As informações extraídas são públicas e foi configurado um delay de 3 segundos a cada iteração da raspagem, com o intuito de não sobrecarregar os servidores.

3.2 Análise exploratória de dados

3.2.1 Amostras

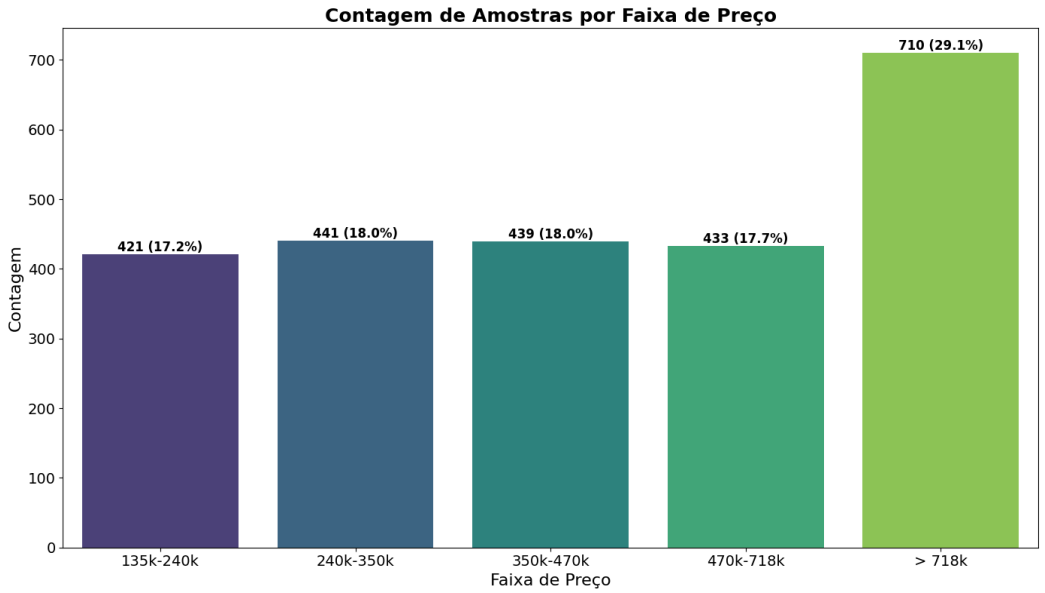
Figura 5: Contagem de Amostras por Categoria



Fonte: Próprio Autor (2025).

A primeira análise a ser feita é sobre a quantidade de amostras por categoria de imóveis, contendo um total de 2453 imóveis, 1385 (56,5%) do tipo apartamento e 1068 (43,5%) do tipo casa.

Figura 6: Contagem de Amostras por Faixa de Preço



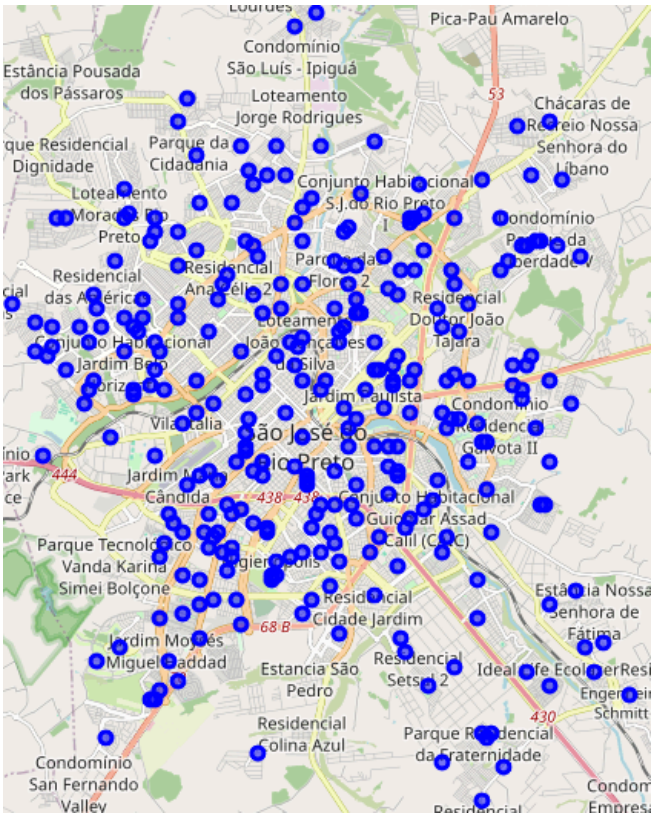
Fonte: Próprio Autor (2025).

Na contagem das amostras por faixa de preços, é possível observar uma quantidade razoável de amostras, 421 na faixa mais baixa de 135 a 240 mil reais, 441 na faixa dos 240 a 350 mil reais, 439 na faixa de 350 a 470 mil reais, 433 na faixa dos 470 a 718 mil reais, a partir de 718 mil reais há 710 amostras.

3.2.2 Geolocalização:

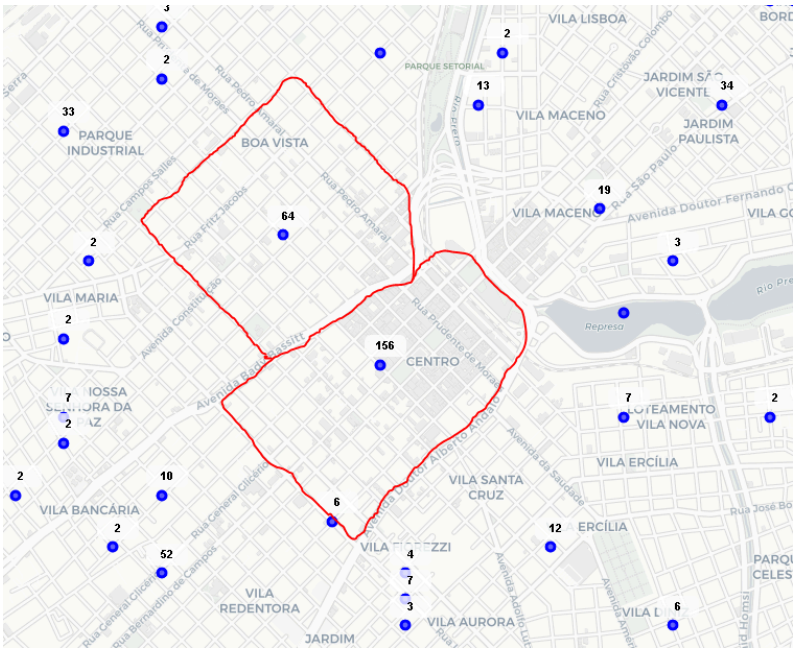
Um ponto de atenção importante das amostras é o baixo nível de granularidade na localização, como os dados obtidos vieram de fontes públicas, não havia disponível a localização exata da maioria dos imóveis, somente a localização aproximada, ao invés do endereço completo, foi obtido o bairro. Em alguns casos de condomínios, residenciais ou apartamentos foram obtidas as localizações exatas,contudo essa granularidade pode variar de local para local.

Figura 7: Distribuição das Amostras em São José do Rio Preto



Fonte: Próprio Autor (2025)

Figura 8: Contagem de Amostras por Localização



Fonte: Próprio Autor (2025)

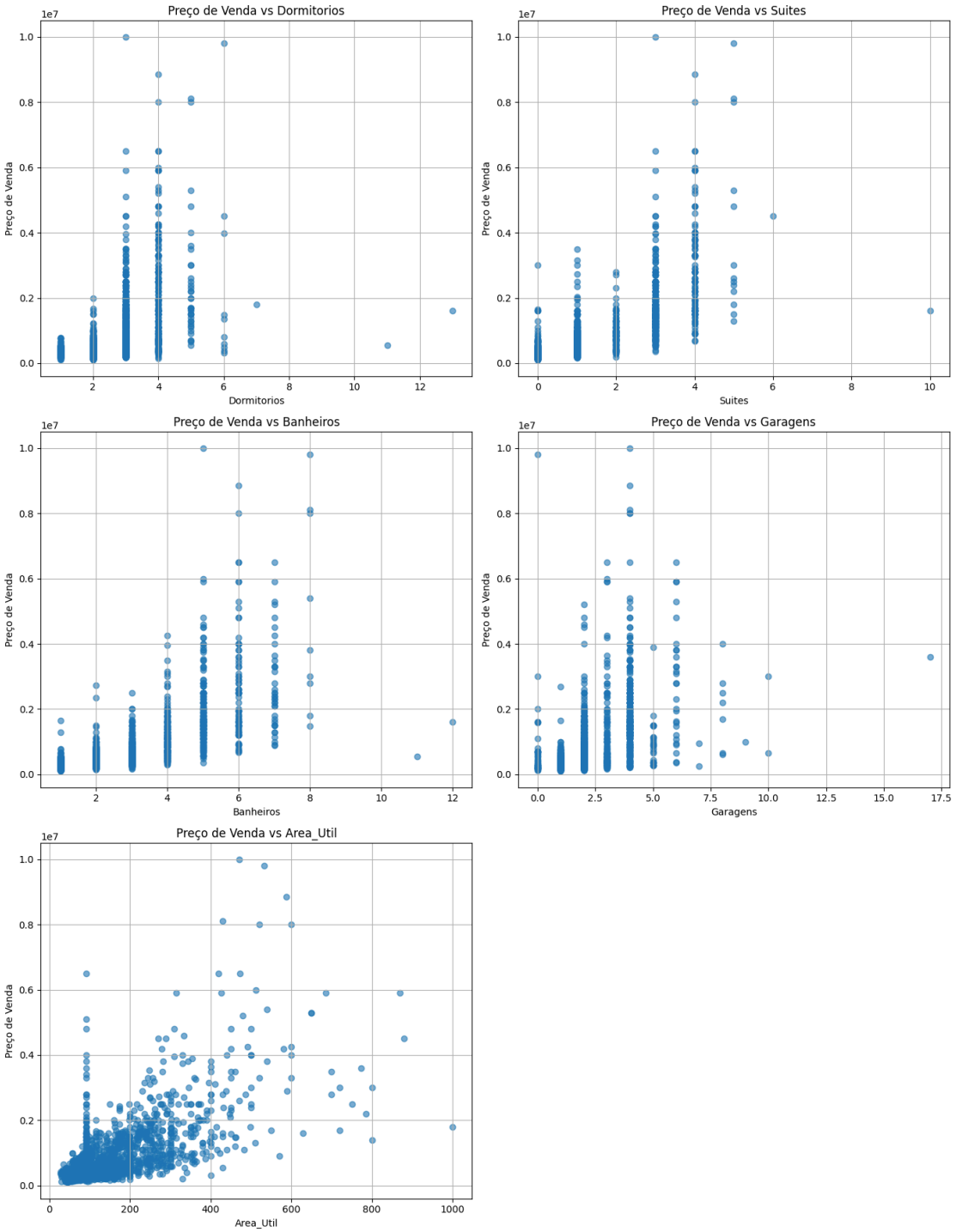
Veja nas imagens acima a forma em que os pontos dos imóveis estão distribuídos no espaço geográfico de São José do Rio Preto, se aproximarmos o mapa, é possível analisar a densidade de amostras pela localização, com a contagem de amostras por local. Os bairros marcados acima, centro e boa vista são a localização com maior densidade de amostras, ambas contendo respectivamente 156 e 64 imóveis, 220 se juntas (aproximadamente 9% de todas as amostras).

3.2.3 Correlação Variáveis Estruturais

Para entender a relação entre as variáveis estruturais e o preço de venda, foram analisados gráficos de dispersão e calculados os coeficientes de correlação de Pearson e Spearman. Esta análise comparativa é fundamental para identificar tanto relações lineares quanto tendências monotônicas entre as características dos imóveis e seu valor de mercado.

3.2.3.1 Gráficos de Dispersão

Figura 9: Gráfico de Dispersão com Características Estruturais.



Fonte:Próprio Autor (2025)

A variável área útil apresentou a associação mais visível com o preço de venda, caracterizada por uma tendência positiva clara. Contudo, observa-se a presença de heterocedasticidade, evidenciada pelo aumento da dispersão dos

preços à medida que a área do imóvel cresce. Esse comportamento indica que, embora exista uma correlação forte, ela não é exclusivamente linear em toda a faixa de valores.

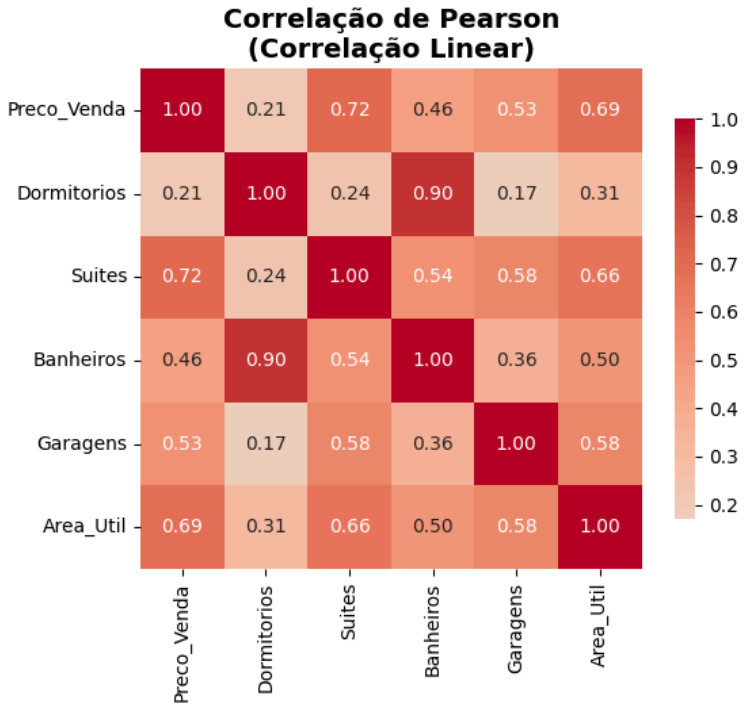
O número de dormitórios e garagens apresentou maior dispersão dos dados, sem alinhamento claro a uma relação linear. Ainda assim, observa-se uma tendência geral de elevação dos preços conforme essas variáveis aumentam, embora com baixo poder explicativo quando analisadas isoladamente. Esse padrão sugere a existência de uma relação monotônica, porém não linear, em que o efeito marginal dessas variáveis varia ao longo de seus intervalos.

Em relação à quantidade de suítes, banheiros e vagas de garagem, verifica-se uma associação positiva mais correlacionada com o preço de venda. O acréscimo dessas características eleva o preço do imóvel, mas com efeito reduzido em níveis mais elevados. Tal comportamento é típico de relações não lineares, nas quais o acréscimo de unidades adicionais contribui para o valor do imóvel principalmente dentro de faixas específicas.

De modo geral, a análise gráfica indica que as relações entre o preço de venda e as variáveis estruturais são predominantemente não lineares. No próximo tópico iremos complementar a análise de correlação com a análise de Pearson e Spearman.

3.2.3.2 Correlação de Pearson

Figura 10: Correlação de Pearson Características Estruturais

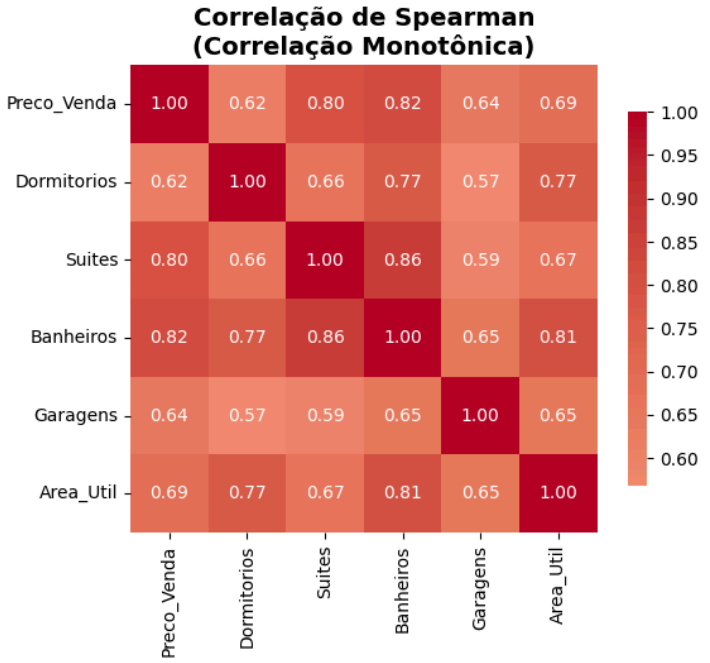


Fonte: Próprio Autor (2025).

Na correlação linear de Pearson, observa-se que a quantidade de suítes (0,72) e área útil (0,69) apresentam as correlações mais fortes com o preço, indicando que aumentos lineares nestas variáveis estão associados a aumentos significativos no valor do imóvel. O número de garagens (0,53) mostra correlação moderada, enquanto o de banheiros (0,46) apresenta relação linear fraca a moderada. A quantidade de dormitórios (0,21) demonstra a menor correlação linear entre todas as variáveis estruturais.

3.2.3.3 Correlação de Spearman

Figura 11: Correlação de Spearman Características Estruturais



Fonte: Próprio Autor (2025).

A correlação monotônica de Spearman revela um cenário distinto: quantidade de banheiros (0,82) e suites (0,80) são as variáveis com maior correlação com o preço. Número de dormitórios apresentando um salto significativo (0,62 contra 0,21 em Pearson).

As discrepâncias entre os dois métodos revelam padrões importantes, o baixo valor de Pearson para quantidade de dormitórios (0,21) comparados ao médio valor da correlação de Spearman (0,62) sugere que, embora a relação não seja estritamente linear, existe uma tendência monotônica clara: imóveis com mais dormitórios tendem a ter preços mais elevados. Número de banheiros apresentam a maior diferença (0,46 × 0,82), indicando que a relação com o preço é fortemente monotônica. Suítes e área útil mantêm correlações elevadas em ambas as métricas, confirmando que sua relação com o preço é consistente tanto linear quanto monotonicamente.

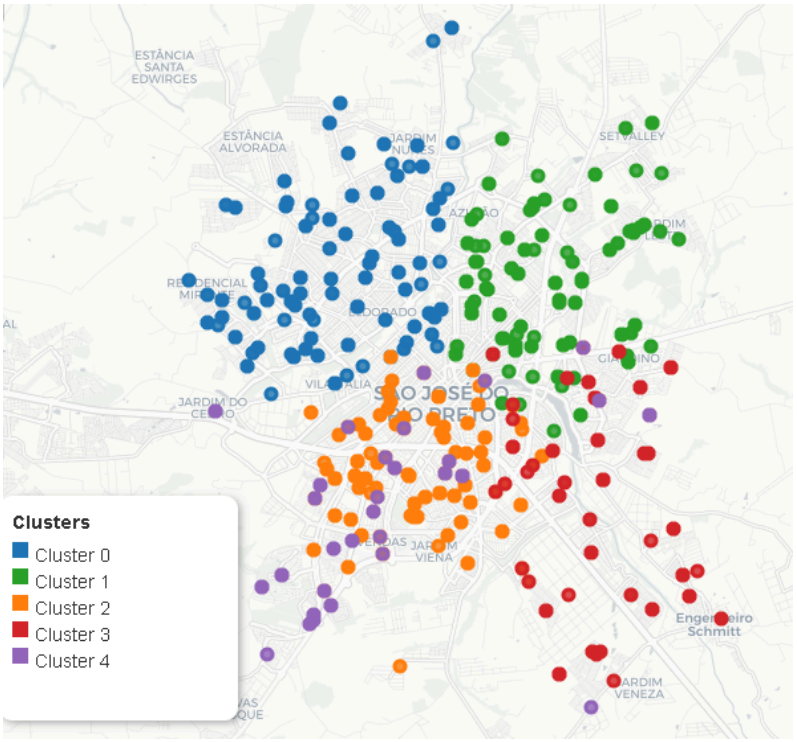
Portanto, a análise de correlação indica que a área útil, o número de banheiros e a quantidade de suítes apresentam as associações mais relevantes com o preço de venda dos imóveis. Observa-se que essas relações não seguem

estritamente um padrão linear, o que levanta a hipótese de que modelos capazes de capturar não linearidades, como aqueles baseados em árvores de decisão, possam apresentar melhor desempenho na etapa de modelagem.

3.2.4 Preços

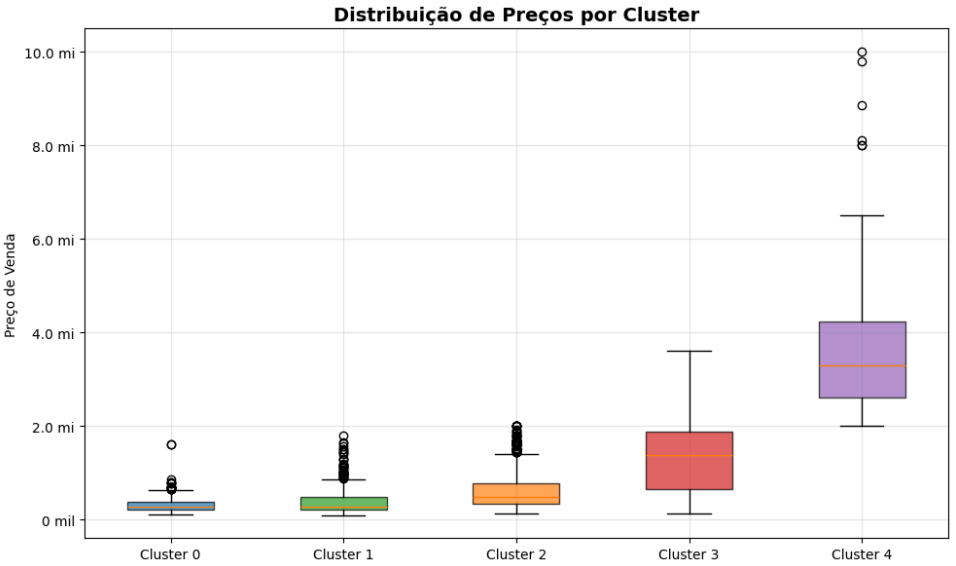
Exclusivamente para análise, a clusterização foi realizada com base nas variáveis latitude, longitude e preço de venda, empregando o algoritmo K-Means com K=5, foi obtido uma pontuação de Silhouette = 0,36 e Calisnki-Harabasz = 1200 indicando que ainda há uma heterogeneidade razoável dentro de cada cluster. O intuito foi agrupar imóveis geograficamente próximos e com faixas de preço semelhantes, de forma a identificar zonas de valorização e desvalorização no município de São José do Rio Preto.

Figura 12: Clusterização por Latitude-Longitude e Preço.



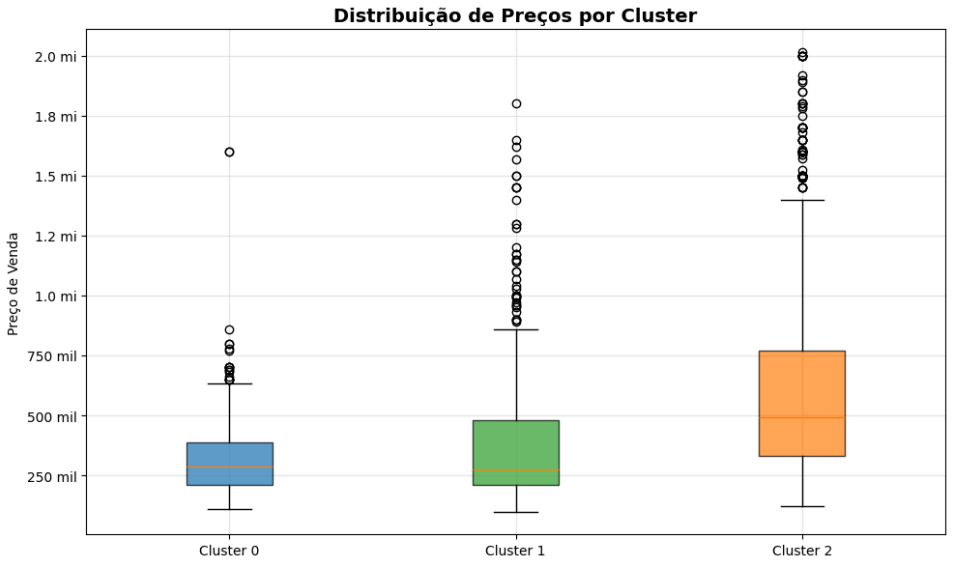
Fonte: Próprio Autor (2025).

Figura 13: Box Plot do Preço por Cluster.



Fonte: Próprio Autor (2025).

Figura 14: Box Plot do Preço por Cluster Ampliado.



Fonte: Próprio Autor (2025).

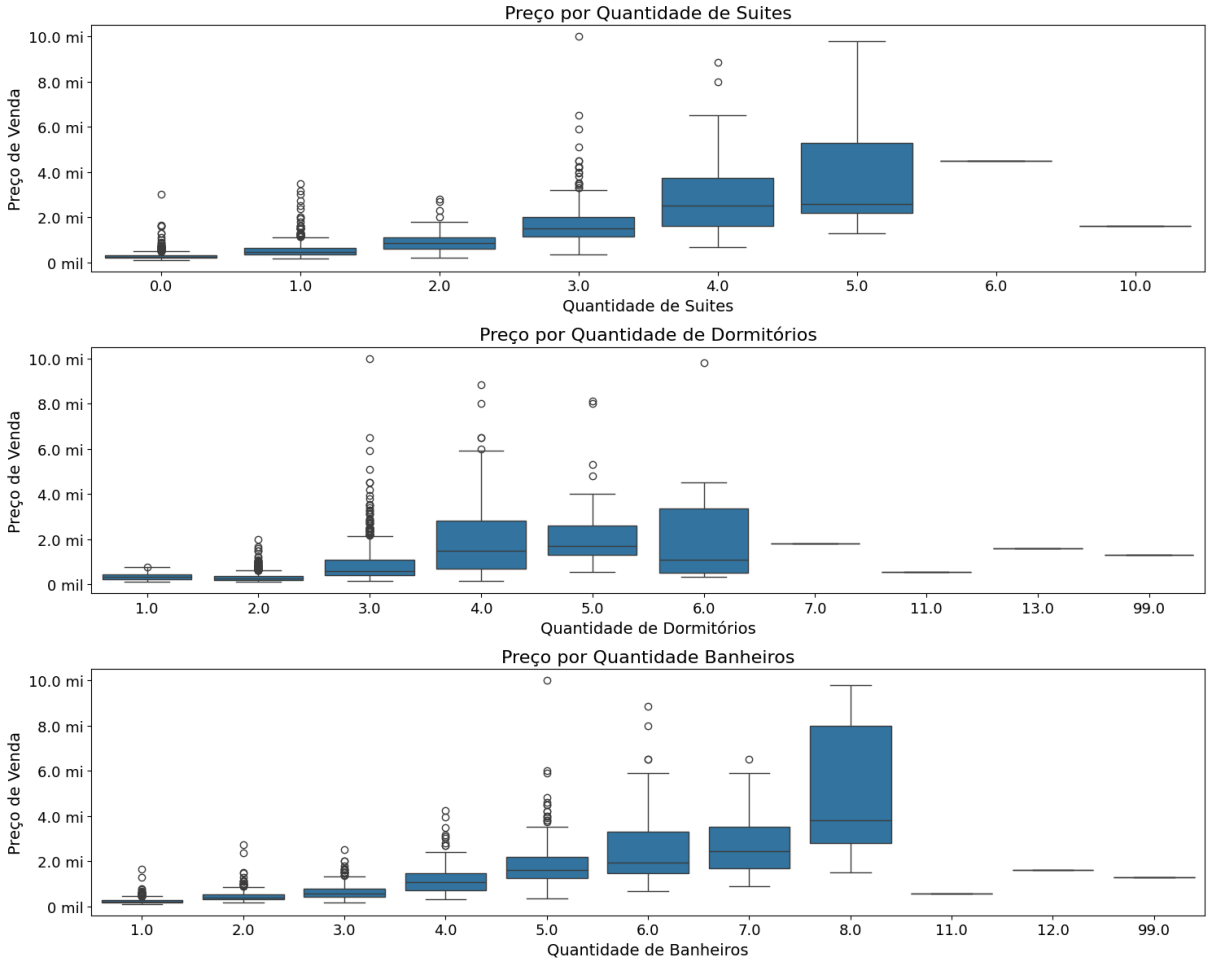
O mapa mostra uma segmentação espacial do mercado imobiliário local. O cluster 0 (em azul) representa as áreas com menores valores médios de venda, esse cluster representa a maior parte da Zona Norte na cidade, região com menor infraestrutura urbana, onde predominam imóveis populares, apesar de haver

exceções, pois os pontos azuis mais próximos ao centro representam áreas de transição que ainda há uma boa infraestrutura.

O cluster 1 (em verde) e o cluster 2 (em laranja) apresentam valores intermediários, o cluster 1 que mistura um pequeno pedaço da Zona Norte com a Zona Leste, região de melhor infraestrutura, contendo alguns condomínios de classe Alta (Damha), o que resultou em um grande número de outliers. Há um efeito semelhante na região do cluster 2, pois é uma região de boa localização e infraestrutura na cidade, contudo há uma mistura de regiões de classe Alta e classe Média.

O cluster 3 (em vermelho) marca o início das áreas de maior valorização, onde o preço médio supera R$1 milhão e há maior presença de imóveis de alto padrão, com acesso mais rápido a regiões centrais pela rodovia. Por fim, o cluster 4 (em roxo) reúne os imóveis mais caros da cidade, frequentemente ultrapassando R$3 milhões, chegando a casos isolados acima de R$8 milhões. Essas áreas correspondem às regiões mais nobres, que refletem uma característica da cidade de São José do rio Preto, condomínios fechados em regiões mais distantes do centro, contudo com menor tempo de chegada, devido a Avenida Juscelino Kubitschek e Bady Bassit, que integram uma via de rápido acesso, sendo as margens da Avenida Juscelino Kubitschek e da represa as áreas mais valorizadas da cidade.

Figura 15: Box Plot do Preço por Características Estruturais.



Fonte: Próprio Autor (2025).

A imagem acima apresenta os boxplots da variável preço em função da quantidade de suítes, dormitórios e banheiros. Observa-se uma tendência clara de crescimento do preço de venda conforme aumenta o número de cômodos, evidenciando uma relação monotônica positiva entre essas variáveis.

De acordo com a matriz de correlação de Spearman, as quantidades de banheiros e suítes apresentam as correlações mais fortes com o preço, seguidas por área útil e dormitórios. Esse resultado é coerente com o comportamento visual observado nos boxplots: quanto maior o número de suítes e banheiros, maior tende a ser o valor mediano do imóvel.

Nos gráficos, nota-se que imóveis com 3 a 5 suítes, 3 a 6 dormitórios e 3 a 7 banheiros concentram a maior parte das observações, com medianas crescentes de preço à medida que o número de cômodos aumenta. Além disso, a dispersão dos

preços também se amplia, indicando que imóveis maiores tendem a apresentar maior variabilidade de valor, possivelmente associada a diferenças de localização, padrão construtivo e metragem.

Entretanto, observa-se a presença de outliers nas categorias extremas, especialmente imóveis com mais de 6 suítes, mais de 7 dormitórios e 8 ou mais banheiros cujos preços se distanciam significativamente da tendência central. Esses registros serão removidos da base de dados por serem considerados outliers, garantindo maior confiabilidade às análises estatísticas e modelagens preditivas subsequentes.

3.3 Pré-processamento dos Dados

Neste estudo, as etapas de pré-processamento concentraram-se principalmente em tratamento de valores nulos, normalização, remoção de outliers por cluster usando o método do IQR e One Hot Encoder para codificar as variáveis categóricas.

3.4 Feature Engineering

3.4.1 Metodologia

Partindo de um dataset contendo coordenadas geográficas de diversos pontos de interesse, o processo consistiu em criar duas categorias principais de features para cada tipo de localização:

Features de Contagem por Faixa de Distância: Para cada tipo de ponto de interesse (como escolas, hospitais, shoppings, farmácias, faculdades e supermercados), foram criadas cinco variáveis que quantificam a quantidade desses estabelecimentos em raio ao redor de cada imóvel:

1. {Tipo}_ate_0.5km
2. {Tipo}_ate_1km
3. {Tipo}_ate_2km
4. {Tipo}_ate_3km
5. {Tipo}_mais_que_3km

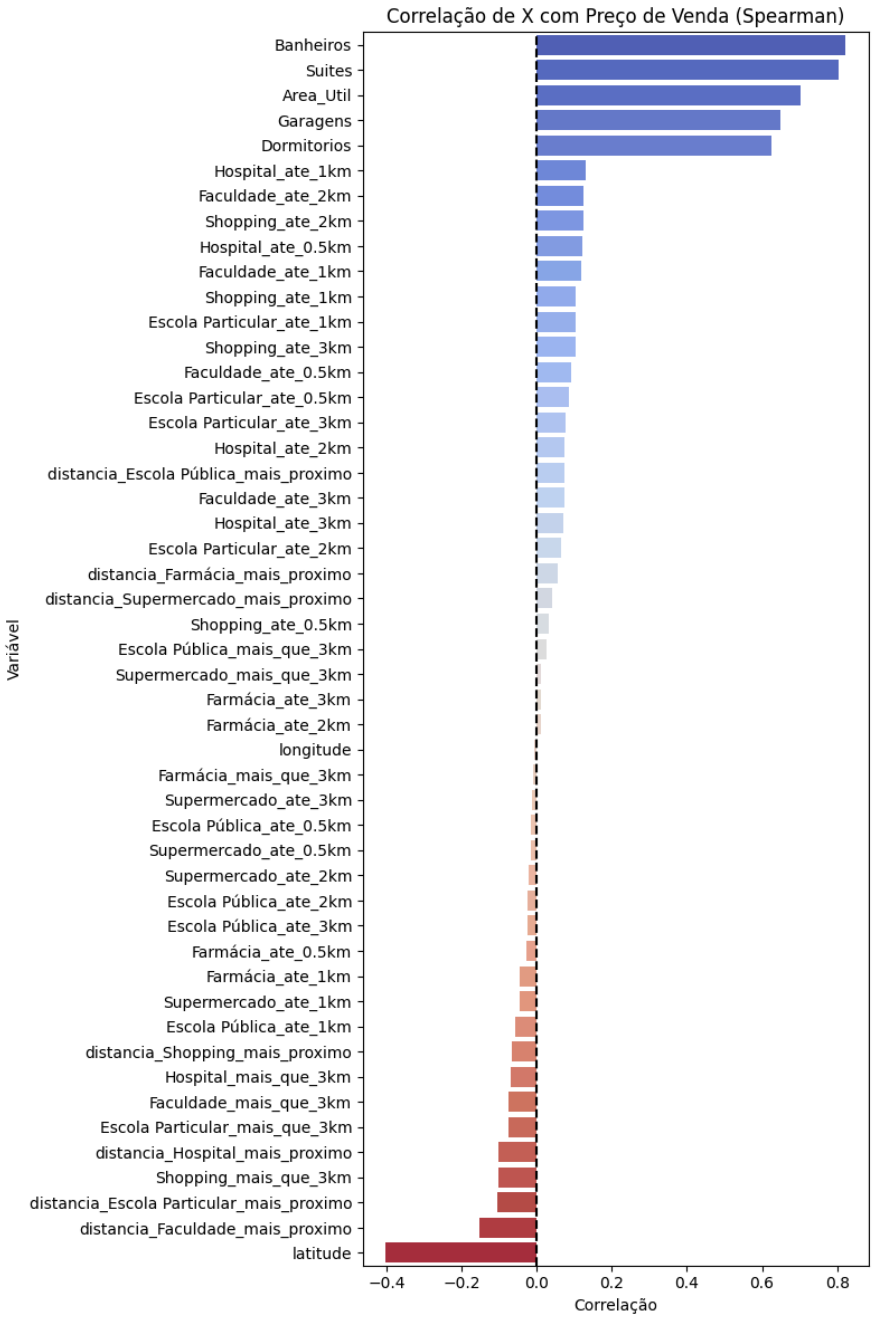
Esta abordagem permite que o modelo de aprendizado de máquina compreenda não apenas a presença, mas também a densidade e a distribuição desses pontos no entorno, capturando efeitos de proximidade imediata (até 0,5 km) e de alcance moderado (até 3 km).

Feature de Distância Mínima: Paralelamente, para cada tipo de ponto de interesse, foi calculada a distância exata até o estabelecimento mais próximo (distancia_{Tipo}_mais_proximo). Esta feature contínua é crucial para mensurar o impacto direto da acessibilidade a um serviço essencial, onde, em muitos casos, a conveniência do "mais próximo" tem um peso significativo.

A implementação utilizou a Fórmula de Haversine para calcular a distância em linha reta entre as coordenadas de cada imóvel e todos os pontos de interesse. O processo iterou por cada imóvel no dataset principal, calculou as distâncias para todos os pontos válidos, as juntou por tipo e, finalmente, preencheu as features criadas com os valores de contagem e distância mínima.

3.4.2 Análise novas features

Figura 16: Correlação de Spearman Pós-Feature Engineering.



Fonte: Próprio Autor (2025).

A análise das correlações após o feature engineering mostra que as variáveis de distância até pontos de interesse apresentam, em geral, baixa correlação monotônica com o preço de venda, o que já era esperado, dado que seus efeitos costumam ser não lineares e dependentes do contexto urbano. Observa-se que distâncias menores até hospitais, faculdades, escolas e shoppings tendem a

apresentar correlações ligeiramente negativas, indicando que a proximidade costuma estar associada a preços mais elevados. Já as variáveis que representam distâncias maiores ou presença de pontos de interesse em raios ampliados exibem correlações muito próximas de zero, sugerindo impacto quase nulo. No conjunto, os resultados reforçam que, embora importantes, essas variáveis não influenciam o preço de forma isolada.

3.5 Modelagem Preditiva

Na etapa de modelagem preditiva, foram selecionados e comparados 3 algoritmos de aprendizado de máquina:, Regressão Linear, Random Forest e XGBoost.

A validação dos modelos foi realizada mediante validação cruzada, técnica que permite uma estimativa mais eficiente do desempenho generalizável de cada algoritmo, reduzindo o risco de overfitting e garantindo que a avaliação seja representativa de diferentes subconjuntos dos dados.

4. Validação

Este capítulo apresenta a validação dos modelos preditivos desenvolvidos para estimativa de valores imobiliários, com base nas estratégias de remoção de outliers por cluster e algoritmos de machine learning implementados. A validação tem como objetivo principal identificar qual combinação de estratégia de clustering e algoritmo apresenta o melhor desempenho preditivo, validando a eficácia da abordagem proposta.

4.1 Desempenho Estratégias e Modelos

Figura 17: Top 10 Melhores Modelos e Estratégias.

Rank	Estratégia	Nº de Clusters	Modelo	MAPE	MAE	R2	Percentual Removido
1	Latitude-Longitude	3	RandomForest	17,35%	R$ 74.340,69	0,8315	8,10%
2	Latitude-Longitude	3	XGBoost	17,62%	R$ 74.384,84	0,8347	8,10%
3	Características Estruturais	2	RandomForest	17,72%	R$ 77.315,94	0,8318	8,19%
4	Características Estruturais	2	XGBoost	17,84%	R$ 78.978,80	0,8276	8,19%
5	Latitude-Longitude	9	RandomForest	17,91%	R$ 77.940,01	0,8246	7,97%
6	Latitude-Longitude	6	XGBoost	17,94%	R$ 75.969,92	0,8362	7,97%
7	Latitude-Longitude	9	XGBoost	18,02%	R$ 78.564,30	0,8168	7,97%
8	Latitude-Longitude	5	XGBoost	18,15%	R$ 75.653,22	0,8470	9,12%
9	Latitude-Longitude	8	XGBoost	18,15%	R$ 76.486,39	0,8494	8,02%
10	Características Estruturais	3	XGBoost	18,27%	R$ 83.298,92	0,8224	5,93%
-	Sem Remoção de Outliers	-	RandomForest	22,44%	R$ 165.452,82	0,8315	0,00%

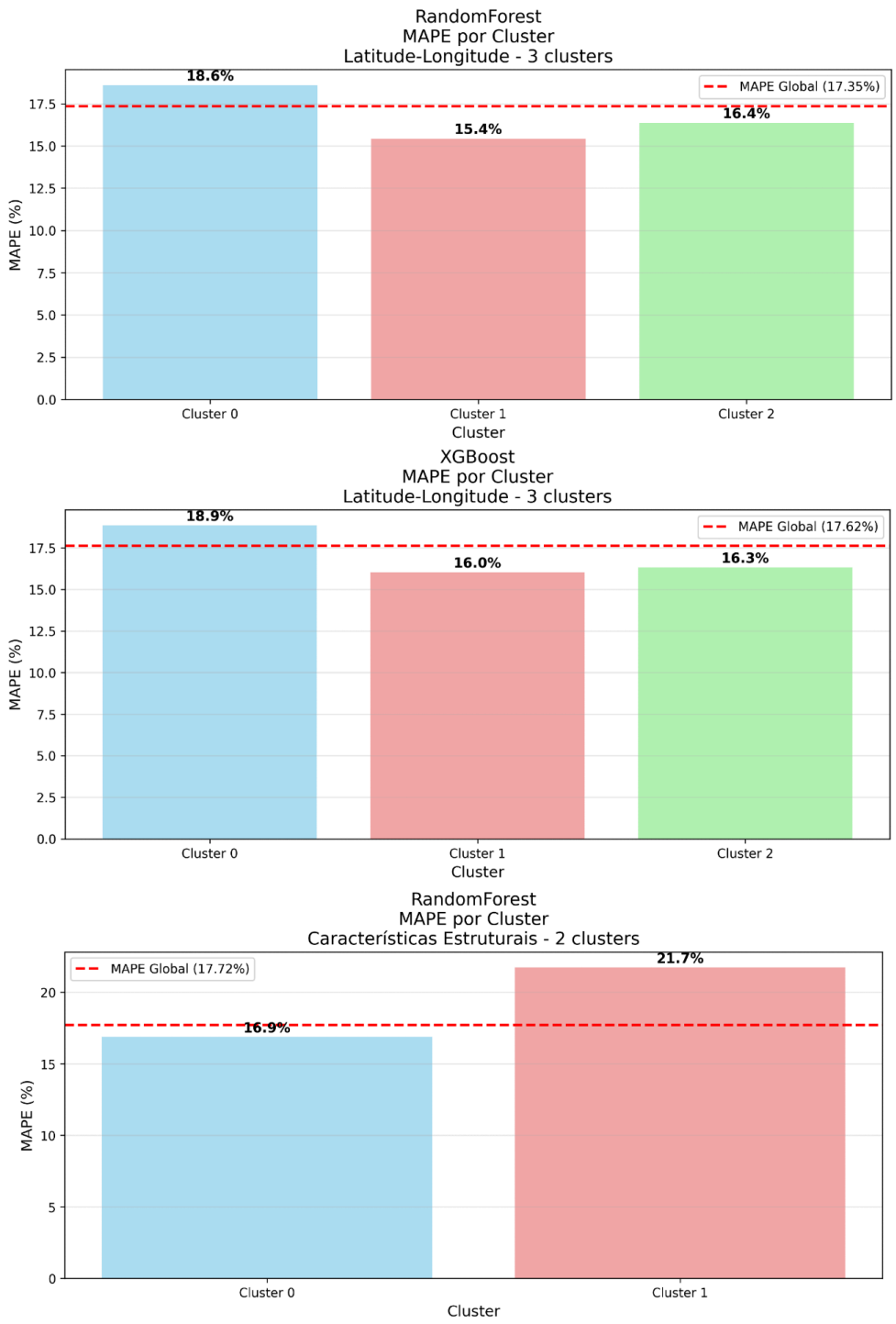
Fonte: Próprio Autor (2025).

Ao analisar os 10 melhores resultados, é possível observar que os testes com modelos de Regressão Linear não atingiram um resultado competitivo, com performance variando entre 25% a 20% de MAPE. A estratégia de remoção de outliers por clusterização com base em Geral X, que utilizava todo o conjunto de variáveis independentes também não alcançou resultado suficiente para entrar no ranking, variando entre 22% a 19% de MAPE nas melhores performances.

As estratégias de remoção de outliers por clusterização que obtiveram melhores performances, incluem a remoção por latitude e longitude e por características estruturais, com destaque para a estratégia latitude-longitude, que teve 7 dos 10 melhores resultados. Os modelos de Aprendizado de máquina Random Forest e XGBoost tiveram os melhores resultados, enquanto o XGBoost obteve 7 dos 10 melhores resultados, o Random Forest obteve 3 dos 5 melhores

resultados, tendo como destaque principal o modelo com menor MAPE e MAE. De forma geral, o modelo vencedor foi o Random Forest com a estratégia de remoção de outliers por clusters baseado em Latitude e Longitude com MAPE de 17,35% e MAE de R$ 74.340,69. Os resultados comprovam que os modelos baseados em árvores performam melhor neste conjunto de dados, devido à natureza não-linear das principais correlações, como mostrado na etapa de análise com a correlação de Spearman.

Figura 18: Erro Detalhado por Cluster Comparados ao Erro Global (Top 3).



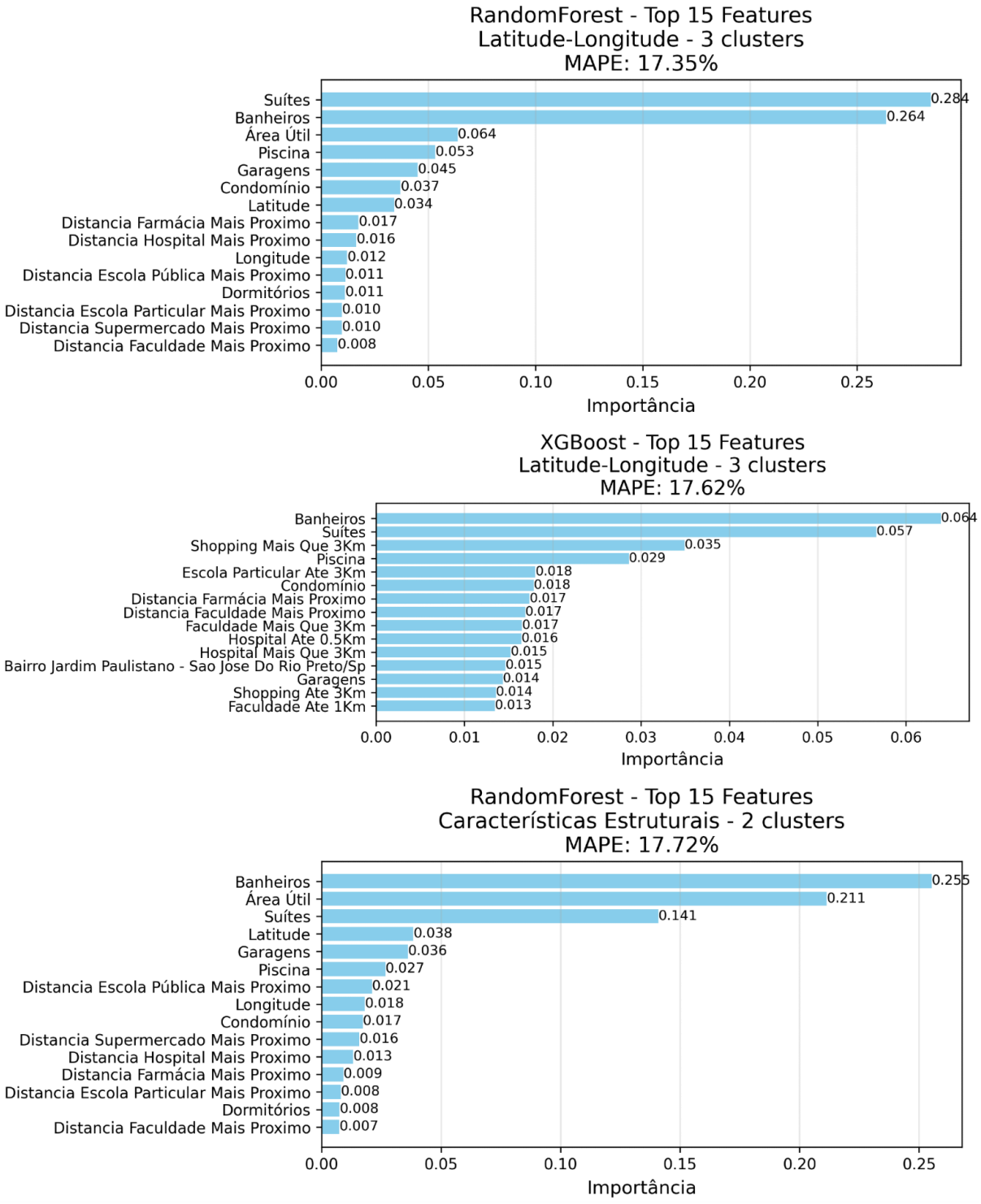
Fonte: Próprio Autor (2025).

Ao executar uma análise mais detalhada do erro por cluster, é possível observar que apesar das estratégia de remoção de outliers por cluster com base em características estruturais ter ficado em 3º lugar, há um desequilíbrio no MAPE por

cluster, enquanto a estratégia Latitude e Longitude obteve um maior equilíbrio no MAPE entre os clusters. A recomendação principal é usar a estratégia Latitude e Longitude para remoção de outliers por cluster, com 3 clusters combinados com o Random Forest.

4.2 Importância das Features

Figura 19: Feature Importance (Top 3).



Fonte: Próprio Autor (2025).

Para os 3 melhores resultados, é possível concluir que as características estruturais do imovel são as principais variáveis na estimativa do preço, a quantidade de suítes, banheiro e a área útil, assim como na correlação de spearman ficaram no topo das principais características associadas ao preço.

As características extraídas pela etapa de feature engineering, com as informações de distância e frequência de pontos de interesses em diferentes faixas de raio, tiveram, se somados, um impacto considerável na estimativa do preço 7,2%, 16,2% e 7,4% para modelos 1, 2 e 3 respectivamente, apesar da menor correlação. A latitude, a contagem de shopping a mais de 3 Km, a contagem de escolas particulares até 3 km, a distância da escola pública mais próxima e a distância da farmácia mais próxima foram as características espaciais mais relevantes, segundo os melhores modelos.

5. Conclusão e Trabalhos Futuros

5.1 Conclusão

Neste estudo foram comparados diferentes algoritmos de machine learning, bem como estratégias de remoção de outliers baseadas em agrupamentos (clusters). Os resultados indicam que os melhores modelos alcançaram erro percentual (MAPE) entre 17% e 18%, e erro médio absoluto (MAE) entre R$ 74.340,69 e R$ 79.000,00. Dentre os algoritmos avaliados, destacaram-se os modelos baseados em árvores Random Forest e XGBoost refletindo a predominância de relações não lineares no conjunto de dados.

A estratégia mais eficaz de tratamento de outliers foi a segmentação por latitude e longitude com 3 clusters, que proporcionou melhor equilíbrio no erro percentual entre os grupos analisados. Os resultados podem ser considerados satisfatórios para dados públicos, porém ainda há espaço para aprimoramentos. A ausência de variáveis estruturais que capturem o nível de luxo, estado de conservação, andar de apartamentos e localização exata limita parcialmente a capacidade preditiva dos modelos; a inclusão desses atributos tende a fortalecer substancialmente as previsões.

5.2 Trabalhos Futuros

Classificação de imagens de imóveis:

O desenvolvimento de um modelo de visão computacional capaz de classificar o imóvel com base em sua aparência, capturando padrões de conservação, luxo, acabamentos e outros atributos estruturais, tem grande potencial. Tais informações, ausentes no dataset utilizado, possuem alta probabilidade de reduzir o erro dos modelos. Trabalhos como Poursaeed et al. (2018) exploram abordagens similares ao estimar preços de imóveis diretamente a partir de fotos.

Usar dados de sensoriamento remoto:

A utilização de imagens de satélite pode adicionar informações atualmente não representadas, como densidade de áreas verdes, áreas pavimentadas,

sombreamento urbano e padrões espaciais não lineares. Esses fatores têm demonstrado correlação relevante com o valor imobiliário em estudos recentes.

Incorporação de dados socioeconômicos e de segurança pública:

Variáveis como índices de criminalidade, IDH médio e indicadores de desenvolvimento urbano por região podem complementar o modelo e reduzir vieses na precificação. Esses atributos capturam nuances contextuais importantes que não estavam presentes no dataset utilizado.

Referências

ANSELIN, Luc. Spatial Econometrics: Methods and Models. Dordrecht: Kluwer Academic Publishers, 1988.

BREIMAN, Leo. Random Forests. Machine Learning, 2001.

BREIMAN, Leo; FRIEDMAN, Jerome; OLSHEN, Richard; STONE, Charles. Classification and Regression Trees. Belmont: Wadsworth International Group, 1984.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. AI Magazine, 1996.

FIRJAN. IFDM – Índice FIRJAN de Desenvolvimento Municipal. Disponível em: https://www.firjan.com.br.

Hyndman, Rob J.; ATHANASOPOULOS, George. Forecasting: Principles and Practice. 2. ed. Melbourne: OTexts, 2018.

IBGE – Instituto Brasileiro de Geografia e Estatística. Disponível em: https://www.ibge.gov.br.

WRI BRASIL. Construções em cidades brasileiras crescem mais que a população, aponta estudo. Disponível em: https://noticias.uol.com.br/ultimas-noticias/agencia-brasil/2025/03/27/construcoes-em-cidades-brasileiras-crescem-mais-que-a-populacao.htm.

MACQUEEN, J. B. Some Methods for Classification and Analysis of Multivariate Observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967.

MITCHELL, Ryan. Web Scraping with Python: Collecting More Data from the Modern Web. Sebastopol: O'Reilly Media, 2018.

POURSAEED, Omid; et al. Vision-Based Real Estate Price Estimation. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018.

ROSEN, Sherwin. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 1974.

SCIKIT-LEARN DEVELOPERS. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, 2011.

SHEARER, Colin. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 2000.

Neves, F. T., Aparicio, M., & de Castro Neto, M. (2024). The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities. Applied Sciences, 14(5), 2209

Zhao, Y., Ravi, R., Shi, S., Wang, Z., Lam, E. Y., & Zhao, J. (2022). PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction. arXiv:2209.05471.

CELLMER, Radosław. *Points of Interest and Housing Prices*. Real Estate Management and Valuation, v. 31, n. 1, p. 69-77, 2023. DOI: 10.2478/remav-2023-0007.

CELLMER, Radosław; BEŁEJ, Mirosław; TROJANEK, Radosław. *Housing prices and points of interest in three Polish cities*. Journal of Housing and the Built Environment, v. 39, p. 1509–1540, 2024. DOI: 10.1007/s10901-024-10124-7.