UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Maiara Medeiros de Sousa

Modelo Random Forest aplicado a precificação de imóveis à venda em Aracaju, SE

São Cristóvão - SE

2023

Maiara Medeiros de Sousa

Modelo Random Forest aplicado a precificação de imóveis à venda em Aracaju, SE

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Cleber Martins Xavier

São Cristóvão - SE

2023

Maiara Medeiros de Sousa

Modelo Random Forest aplicado a precificação de imóveis à venda em Aracaju, SE

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Aprovado em DATA DE APROVAÇÃO.

Prof. Dr. Cleber Martins Xavier
Orientador

Prof. Dr. Luiz Henrique Gama Dore de Araujo

Prof. Dr. Sadraque Eneas de Figueiredo Lucena

São Cristóvão - SE
2023

Este trabalho é dedicado a todos os professores que me influenciaram na minha trajetória.

Agradecimentos

Gostaria de expressar minha profunda gratidão à minha família, especialmente aos meus pais, Joseneide e Edvaldo, meu irmão Alex e meu sobrinho Théo, pela compreensão e apoio incondicional ao longo da minha jornada. Vocês foram o alicerce que sustentou cada passo desta caminhada acadêmica.

Aos meus queridos amigos, Mayara, Tiago, Juliene, Dara e Davi, que estiveram sempre ao meu lado, compartilhando risos e me incentivando nos momentos mais desafiadores, tornando esta jornada ainda mais especial.

Aos meus respeitáveis professores, quero expressar minha profunda gratidão pela dedicação, orientação e sabedoria que compartilharam ao longo deste percurso acadêmico. Em especial, ao meu orientador Cleber Xavier, pela paciência, suporte e estímulo incansáveis durante todo esse período. Suas palavras e ensinamentos foram fundamentais para o meu crescimento intelectual e formação.

A todos vocês, minha sincera gratidão. Este trabalho é dedicado a cada um que contribuiu de alguma forma para a minha trajetória acadêmica e pessoal.

“Uno studioso al microscopio vede molto più di noi.
Ma c’è un momento, un punto, in cui anch’egli deve fermarsi.
Ebbene, è a quel punto che per me comincia la poesia.
(René Magritte)

Resumo

O mercado imobiliário brasileiro é considerado um dos mais promissores do mundo. Segundo a Associação Brasileira de Incorporadoras Imobiliárias (Abrainc), entre os anos de 2012 a 2022, a valorização média anual do preço dos imóveis foi de 12,2%. Essa valorização torna o mercado atrativo não apenas para pessoas que desejam adquirir uma propriedade para morar, mas também para aquelas que desejam transformar o bem em um investimento, seja alugando ou vendendo. No entanto, antes de efetuar a compra de uma habitação, é necessário realizar uma análise dos fatores que compõem o preço do imóvel, a fim de verificar se o mesmo está sendo vendido por um valor justo e se está localizado em áreas com potencial de valorização. Para auxiliar os consumidores, foram coletados dados de imóveis à venda na cidade de Aracaju – SE, utilizando a técnica de *web scraping*. Esses dados foram posteriormente analisados e utilizados para desenvolver um modelo de predição baseado no algoritmo de *machine learning*, *Random Forest*, capaz de precificar imóveis com base em suas características. Após as análises, verificou-se que um imóvel em Aracaju tem um preço médio de R$ 561.000,00. Os bairros da Zona Sul e de Expansão possuem habitações mais caras em comparação com os localizados no Centro, Zona Oeste e Zona Norte. Quanto ao modelo de predição, observou-se que o mesmo apresentou um bom poder preditivo para residências de até R$ 1.000.000,00. O fator de maior importância para a precificação na cidade de estudo foi o tamanho da área de construção.

Palavras-chave: Mercado imobiliário. Precificação de imóveis. *Machine learning*. *Random Forest*.

Abstract

The Brazilian real estate market is considered one of the most promising in the world. According to the Brazilian Association of Real Estate Developers (Abrainc), between the years 2012 to 2022, the Brazilian real estate market was considered one of the most promising in the world. According to the Brazilian Association of Real Estate Developers (Abrainc), between the years 2012 to 2022, the average annual appreciation of property prices was 12.2%. This appreciation makes the market attractive not only for people looking to acquire a property to live in, but also for those who want to turn the property into an investment, either by renting or selling. However, before purchasing a property, it is necessary to conduct an analysis of the factors that make up the property's price, in order to determine if it is being sold at a fair value and if it is located in areas with potential for appreciation. To assist consumers, data from properties for sale in the city of Aracaju – SE, Brazil, was collected using the web scraping technique. These data were later analyzed and used to develop a prediction model based on the Random Forest machine learning algorithm, capable of pricing properties based on their characteristics. After the analyses, it was found that a property in Aracaju has an average price of R$ 561,000.00. The neighborhoods in the South Zone and Expansion Zone have more expensive housing compared to those located in the Central, West, and North Zones. Regarding the prediction model, it was observed that it showed good predictive power for residences up to R$ 1,000,000.00. The most important factor for pricing in the city of study was the size of the construction area.

Keywords: Real estate market. Property pricing. Machine learning. Random forests.

Resumen

El mercado inmobiliario brasileño es considerado uno de los más prometedores del mundo. Según la Asociación Brasileña de Desarrolladores Inmobiliarios (Abrainc), entre los años 2012 y 2022, la apreciación media anual del precio de las propiedades fue del 12,2 %. Esta apreciación hace que el mercado sea atractivo no solo para personas que desean adquirir una propiedad para habitar, sino también para aquellas que desean convertir el bien en una inversión, ya sea alquilándolo o vendiéndolo. Sin embargo, antes de realizar la compra de una vivienda, es necesario llevar a cabo un análisis de los factores que componen el precio del inmueble, con el fin de verificar si se está vendiendo a un valor justo y si está ubicado en zonas con potencial de apreciación. Para ayudar a los consumidores, se recopilaron datos de propiedades en venta en la ciudad de Aracaju, SE, utilizando la técnica de *web scraping*. Estos datos fueron posteriormente analizados y utilizados para desarrollar un modelo de predicción basado en el algoritmo de aprendizaje automático, *Random Forest*, capaz de establecer precios para inmuebles según sus características. Tras los análisis, se determinó que una propiedad en Aracaju tiene un precio medio de R$ 561.000,00. Los barrios de la Zona Sur y de Expansión cuentan con viviendas más caras en comparación con las ubicadas en el Centro, Zona Oeste y Zona Norte. En cuanto al modelo de predicción, se observó que tuvo una buena capacidad predictiva para viviendas de hasta R$ 1.000.000,00. El factor de mayor importancia para la fijación de precios en la ciudad de estudio fue el tamaño del área de construcción.

Palabras clave: Mercado inmobiliario. Fijación de precios de propiedades. *Machine Learning*. *Random Forest*.

Lista de ilustrações

Figura 1 – Estrutura de um modelo de árvore de decisão 26
Figura 2 – Distribuição da quantidade de imóveis à venda diponíveis nos bairros da cidade de Aracaju - SE considerando a variável Cidade. 31
Figura 3 – Análise do preço dos imóveis à venda na cidade de Aracaju - Sergipe considerando as variáveis Tipo e Zonas. 31
Figura 4 – Análise da área dos imóveis à venda na cidade de Aracaju - SE considerando as variáveis Tipo e a Zonas. 32
Figura 5 – Resultado da correlação das variáveis coletadas de anúncios de venda de imóveis na cidade de Aracaju - SE. 33
Figura 6 – Comparação entre os valores Previstos e Observados do preço de casas à venda na cidade de Aracaju - SE a partir do melhor modelo de Random Forest. 35
Figura 7 – Classificação de importância das variáveis para o modelo Random Forest aplicado ao preço de venda de imóveis em Aracaju - SE. 35

Lista de quadros

Quadro 1 – Descrição das variáveis presentes no banco de dados para estudo do preço de imoveis na cidade de Aracaju-se. 21
Quadro 2 – Detalhamento da classificação dos tipos de imóveis à venda na cidade de Aracaju - SE. 22
Quadro 3 – Descrição do agrupamento de bairros para estudo do preço de imóveis na cidade de Aracaju - SE 23
Quadro 4 – Descrição do agrupamento de bairros em zonas para o estudo do preço de imóveis em Aracaju - SE. 23

Lista de tabelas

Tabela 1 – Frequência absoluta e percentual de imóveis na cidade de Aracaju - SE considerando as regiões da variável Zona. 30
Tabela 2 – Análise descritiva do preço de imóveis à venda na cidade de Aracaju - SE considerando a variável Tipo. 30
Tabela 3 – Análise descritiva da quantidade de garagens por imóvel à venda em Aracaju 33
Tabela 4 – Resultado dos 10 melhores modelos utilizando o algoritimo do modelo Random Forest aplicado ao preço de venda de imóveis em Aracaju - SE. . . 34

Sumário

1 INTRODUÇÃO 14

2 OBJETIVOS 16
2.1 Geral 16
2.2 Específicos 16

3 REVISÃO LITERÁRIA 17
3.1 Mercado imobiliário 17
3.1.1 Mercado imobiliário em Aracaju - SE 18
3.2 Precificação de imóveis 19
3.3 Aplicações de algoritmos de *Machine Learning* na precificação de imóveis 19

4 METODOLOGIA 21
4.1 Base de dados e Pré - Processamento 21
4.2 Métodos 23
4.2.1 *Web Scraping* 23
4.2.2 *Machine Learning* 24
4.2.2.1 Árvore de decisão 25
4.2.2.2 *Bagging* 27
4.2.2.3 *Random Forest* (Florestas Aleatórias) 27
4.2.2.4 Método de Avaliação de Modelos 28
4.3 Suporte Computacional 29
4.3.1 Python 29
4.3.2 *Software R* 29

5 RESULTADOS 30
5.1 Análise Descritiva 30
5.2 Modelo *Random Forest* 34

6 CONCLUSÃO 36
6.1 Pesquisas futuras 36

REFERÊNCIAS 38

APÊNDICE A – CÓDIGO DO PRÉ - PROCESSAMENTO, ANÁLISE DOS DADOS E CONSTRUÇÃO DO MODELO DE *RANDOM FOREST* 41

1 INTRODUÇÃO

O mercado imobiliário brasileiro é considerado um dos mais promissores do mundo. É composto por pessoas e empresas que atuam na comercialização de empreendimentos imobiliários e na administração destes, como as empresas de construção civil e imobiliárias. Isso se caracteriza pelo desenvolvimento urbano das cidades, visando gerar uma maior qualidade de vida para a população, consequentemente proporcionando o crescimento da economia por meio da geração de empregos diretos e indiretos relacionados aos serviços da construção civil (MATOS; BARTKIW, 2013).

Segundo a Associação Brasileira de Incorporadoras Imobiliárias (Abrainc), houve uma valorização média anual do preço dos imóveis de 12,2% entre os anos de 2012 e 2022. A Abrainc afirma que ocorreu uma alta de 9,2% na venda de imóveis novos em 2022 quando comparado com o ano anterior, totalizando 156.730 unidades vendidas (ABECIP, 2023; ABRAINC, 2022).

De acordo com Freire et al. (2011), Aracaju vivenciou no período de 2001 a 2011 a supervalorização de imóveis já construídos e terrenos, assim como um aumento na demanda por imóveis, o que levou a um aumento no custo médio de construção na cidade.

Com o notável crescimento e valorização do mercado imobiliário, torna-se ainda mais evidente a sua importância para a economia. Carvalho (2017) destaca que o estudo do comportamento dos preços de venda e aluguel de residências desempenha um papel crucial na economia, principalmente na estimativa da demanda do setor imobiliário.

Para Leeuw (1993), medir a mudança de preços de construções é um problema de longa data devido a estruturas não serem produzidas a partir de modelos de produção fixos. No entanto, nos últimos anos a precificação dos imóveis tem sido aprimorada através de modelos hedônicos (regressão), que utilizam características de residências para estimar o seu preço.

Com base no exposto anteriormente, o presente trabalho tem como objetivo coletar dados de imóveis à venda na cidade de Aracaju – SE utilizando a técnica de *web scraping*. Com esses dados, pretende-se propor um modelo de precificação utilizando algoritmos de *machine learning*, analisando as características das construções que compõem o seu valor.

O presente trabalho está dividido em 6 tópicos, começando pelo Capítulo 1, a introdução, onde é exposto de forma breve o conteúdo desenvolvido ao longo do estudo. No segundo capítulo, são abordados os objetivos a serem explanados. No Capítulo 3, é apresentada a revisão de literatura, que explana sobre a história do mercado imobiliário brasileiro, a precificação de imóveis e os métodos de *machine learning* utilizados nesse processo. O quarto capítulo expõe os métodos e materiais utilizados para a construção da análise dos dados até a obtenção de resultados. Inicia-se com a apresentação dos dados e o método pelo qual foram obtidos (*web scraping*), além das modificações feitas neles. Também são abordados o algoritmo de *machine*

$learning$, as árvores de decisão e o $random\ forest$. O penúltimo capítulo apresenta os resultados obtidos após análises e aplicações dos métodos expostos no capítulo anterior, juntamente com as discussões sobre eles. No Capítulo 6, o estudo é finalizado com a apresentação da conclusão obtida após análises e levando em consideração os objetivos do mesmo.

2 OBJETIVOS

2.1 Geral

O objetivo desse trabalho consiste em propor um modelo capaz de precificar o valor de venda de um imóvel de Aracaju - SE utilizando algoritmo de *machine learning*.

2.2 Específicos

- Extrair dados de anúncios de imóveis de venda de imóveis em Aracaju utilizando a técnica de *Web Scraping*.
- Descrever e analisar os dados.
- Construir e treinar modelo de *Random Forest*.
- Analisar as variáveis que possuem mais influência para o modelo e o seu poder preditivo.

3 REVISÃO LITERÁRIA

3.1 Mercado imobiliário

De acordo com Kremer (2008), o desenvolvimento do mercado imobiliário no Brasil está correlacionado com os ciclos econômicos ocorridos no país. Esses ciclos foram divididos por Lacerda et al. (2006) em dois períodos distintos: o período mercantil, no qual se destaca a fase colonial marcada pelos ciclos da cana-de-açúcar e do ouro, e a fase primário-exportadora, caracterizada pela expansão da cafeicultura, quando a burguesia local começou a acumular capital próprio. Posteriormente, ocorreu o período industrial.

Durante o período mercantil, as residências eram espaçosas e frequentemente constituíam propriedades rurais, uma vez que abrigavam famílias numerosas e também serviam como locais de produção. Entretanto, durante o período industrial, com o êxodo rural, as moradias assumiram uma nova configuração, estendendo-se ao longo das vias de acesso às cidades (KREMER, 2008).

Ao longo do êxodo rural, houve uma migração interna significativa da população brasileira em direção à região Sudeste. Isso resultou em um crescimento exponencial das principais cidades da região, levando ao desenvolvimento de núcleos urbanos de pequeno e médio porte (KREMER, 2008).

O período de urbanização no Brasil ocorreu de forma acelerada, resultando em um aumento na quantidade de imóveis urbanos que necessitavam de avaliação. Muitos desses imóveis eram construídos com materiais de pouca resistência, devido ao fato de que grande parte das pessoas que migravam do meio rural em busca de oportunidades possuía baixo poder aquisitivo (CARVALHO, 2019).

> Verifica-se um cenário de urbanização acelerada, estrutura desigual de distribuição de renda e precariedade dos processos de planejamento público, o que resultou em ocupação desordenada das cidades e na fragmentação do tecido urbano e social (KREMER, 2008).

Somente em 1964 o mercado imobiliário no Brasil começou a ser regulamentado com a criação do memorial de incorporação, instrumento que, por lei, exigia a existência de informações sobre o empreendimento e o seu futuro, bem como informações sobre empresa e sócios, com o objetivo de garantir o processo de compra e venda de um imóvel mais seguro (MATOS; BARTKIW, 2013). No mesmo ano, cria-se pelo governo com a intenção de desenvolver formas de financiamento de imóveis e facilitar a construção e obtenção da casa própria, o Banco Nacional de Habitação, gestor do Sistema Financeiro de Habitação (SFH) (KREMER, 2008).

O SFH teve um bom desempenho até o início da década de 1980. Nesse período, estima-se que tenham sido financiadas cerca de 6 milhões de unidades residenciais. No entanto, a economia

brasileira entrou em um período de estagnação, no qual a inflação começou a crescer, chegando a 80% ao mês. As tentativas de contê-la afetaram o Sistema Financeiro de Habitação e o BNH foi extinto (FARIAS, 2010; MATOS; BARTKIW, 2013).

Em 1997, como uma tentativa de aperfeiçoar o sistema de crédito imobiliário, que estava afetado pelo alto nível de inadimplência dos consumidores do setor e substituir o SFH, criou-se o Sistema Financeiro Imobiliário (SFI). Este tem como objetivo promover o financiamento imobiliário e novos empreendimentos com mais recursos, provenientes de investidores de diversas partes do mercado, em oposição ao modelo anterior que dependia de uma única fonte (poupança), e com menos burocracia (FARIAS, 2010).

Com a criação do SFI, foi introduzida a Alienação Fiduciária de Bens e Imóveis, utilizada até os dias atuais. Seu objetivo é garantir o pagamento de uma dívida por meio do bem adquirido pelo devedor.

Atualmente, há interesse por parte dos bancos em disponibilizar linhas de crédito para a construção de novos imóveis. No entanto, a demanda por crédito está diretamente ligada à renda e, no país nos dias atuais, existe uma deficiência habitacional maior para a população cuja faixa salarial é de até cinco salários mínimos, considerados fora do mercado financeiro (KREMER, 2008). Como alternativa para suprir o déficit imobiliário de pessoas de baixa renda, foram lançadas políticas governamentais, como o programa Minha Casa, Minha Vida, que promove o direito à moradia para núcleos familiares com renda per capita de até 8 mil reais.

3.1.1 Mercado imobiliário em Aracaju - SE

Aracaju, município e capital do estado de Sergipe, está situada no leste sergipano e abrange uma área de 182,163 km^2, com uma população de 602.757 habitantes, segundo dados do IBGE (2022).

Na década de 1960, com a criação da Companhia de Habitação de Sergipe (COHAB/SE), Aracaju teve seu território valorizado com a aquisição de inúmeras áreas pela empresa e a construção de conjuntos habitacionais em diferentes zonas do município, especialmente nas zonas norte e oeste (SANTOS, 2023). Entre os anos 1970 e 1980, ocorreu a disseminação da construção de condomínios fechados, inicialmente localizados em bairros da zona sul, como o Grageru, São José, Jardins, Salgado Filho e Treze de Julho (FRANçA; REZENDE, 2014)).

Na década de 1990, começou a surgir a região metropolitana de Aracaju, marcada pela integração das redes intraurbanas e dos municípios fronteiriços, bem como pelo crescimento vertical dos bairros. Nesse período, houve uma concentração de prédios em bairros localizados mais ao sul da capital, principalmente nas avenidas Hermes Fontes e Barão de Maruim (bairros Luzia, Grageru, Jardins e Treze de Julho) (FRANçA; REZENDE, 2014).

A partir dos anos 2000, houve um aumento no número de empreendimentos imobiliários e uma alta nos valores dos imóveis, decorrente do aquecimento da construção civil no país e do

incentivo ao crédito imobiliário. Nesse período, condomínios residenciais fechados e conjuntos habitacionais passaram a ser construídos em bairros como Farolândia, Atalaia, Jabotiana, Luzia, Coroa do Meio e Zona de Expansão, locais de grande especulação fundiária (FRANçA; REZENDE, 2014).

Segundo Santos (2023), atualmente o mercado imobiliário de Aracaju vem sendo impulsionado pelo lançamento de empreendimentos residenciais em bairros localizados mais próximos ao litoral do município, como Aruana, Atalaia e Coroa do Meio, devido à busca estimulada por qualidade de vida, conforto e a presença de recursos ambientais, aliados a fatores econômicos.

3.2 Precificação de imóveis

De acordo com Arraes e Sousa Filho (2008), a aquisição de imóveis pode ser dividida entre as pessoas que pretendem adquirir uma propriedade para habitação e aquelas que têm a intenção de comercializar o bem como investimento. Cada tipo de consumidor seleciona os atributos que desejam que a habitação possua, geralmente associados a aspectos físicos do imóvel, sua localização e fatores ambientais que não estão diretamente ligados a ele. Essas características são fundamentais para a precificação das habitações (ARRAES; FILHO, 2008).

Uma maneira de deduzir o valor de imóvel é utilizando o método de comparação, que consiste em avaliar habitações disponíveis no mercado com base nas suas características intrínsecas (quantidade de quartos, banheiros, garagens, área total, etc.) e extrínsecas (distância do imóvel de pontos de influência como hospitais, escolas e parques). Por meio dessa avaliação, identificam-se as variáveis que influenciam na determinação do preço do imóvel (ARRAES; FILHO, 2008).

O método de Preços Hediônicos é comumente aplicado à precificação de habitações. Por meio dele busca-se obter a importância de cada atributo da habitação para que o preço seja determinado (DANTAS; MAGALHãES; VERGOLINO, 2007). Esse método fundamenta-se na técnica estatística de análise de regressão múltipla, em que os preços são explicados através das características (aspectos físicos, econômicos e localização) dos imóveis.

Nos últimos anos, percebeu-se ainda mais o quão dinâmico é o mercado imobiliário, havendo a necessidade de utilização de técnicas que auxiliem na estratégia de negócio para obter métricas e resultados mais precisos diante do grande volume de dados disponível sobre esse mercado. Assim, o uso de algoritmos de aprendizado de máquina está se tornando cada vez mais comum para auxiliar na precificação de imóveis.

3.3 Aplicações de algoritmos de *Machine Learning* na precificação de imóveis

Segundo Araruna (2022), o mercado imobiliário cresceu substancialmente nos últimos anos e tal crescimento causou a escassez de mão de obra qualificada no setor responsável pela

avaliação e precificação dos imóveis (corretores e imobiliárias). O autor alega que o avanço do mercado causou aumento da concorrência no setor, atraindo profissionais mais focados na prospecção de clientes do que na estimativa de preços dentro dos parâmetros do mercado.

Para tentar corrigir as falhas relacionadas a precificação no mercado imobiliário, Araruna (2022) propôs a comparação de modelos utilizando algoritmos de *Machine learning* (Regressão Linear, Árvores de Regressão, *Random Forest* e Redes Neurais) aplicados a dados de apartamentos à venda em Braília - DF, para auxiliar na precificação de imóveis, comparando - os a partir do risco preditivo. Como resultado de sua pesquisa foi constatado que o modelo de Redes Neurais possui o menor erro preditivo, seguido do modelo *Random Forest*.

Com o intuito de precificar imóveis localizados nos subúrbios de Chicago utilizando modelos de predição, Xu e Nguyen (2022) utilizaram informações como: preço de venda, ano em que foi construído, quantidade de banheiros, quartos, vagas de garagem, entre outros, para construir os modelos de regressão, árvore de decisão, *Random Forest*, XGBoost e *Support Vector Regression*, e assim compará-los. Como resultado, foi observado que o modelo XGBoost obteve o melhor desempenho, tendo como variáveis mais importantes a área, quantidade de banheiros e quartos.

Zaghi (2023) utilizou três modelos de *machine learning*, *Lasso*, *Random Forest* e *XGBoost*, com o objetivo de encontrar o que obteria maior eficácia na predição do preço de imóveis em Florianópolis - SC. Para a construção dos modelos, o autor utilizou dados dos imóveis, como a área, quantidade de quartos, banheiros, vagas de garagem, preço e condomínio. O resultado mostrou que o modelo de *Random Forest* foi o que obteve melhor desempenho diante das quatro métricas avaliadas (RMSE, RSR, R^2, MAPE).

4 METODOLOGIA

4.1 Base de dados e Pré - Processamento

Os dados utilizados no presente estudo foram coletados por meio do método de *Web Scraping* (explicado na subseção 4.2.1) em cinco *websites* de imobiliárias localizadas em Aracaju (Barros e Nobres, Century, COHAB, Planeta Imóveis e Valor Imobiliária), utilizando a linguagem de programação Python. Por meio dessa técnica foram obtidos 1149 anúncios de imóveis nos quais foram disponibilizadas 10 variáveis que podem ser observadas no Quadro 1.

Quadro 1 – Descrição das variáveis presentes no banco de dados para estudo do preço de imoveis na cidade de Aracaju-se.

Variável	Descrição
Tipo	Casa ou Apartamento
Cidade	Bairro localizado o imóvel
Preço	Preço de venda do imóvel
Quarto	Quantidade de quartos do imóvel
Suíte	Quantidade de suítes do imóvel
Banheiro	Quantidade de banheiro do imóvel
Garagem	Quantidade de garagens do imóvel
Área	Área do imóvel em m^2
Código	Código do imóvel exibido nos sites
URL	Link do anúncio

Fonte: Elaborado pela autora

O pré-processamento de dados consiste em identificar problemas para aprimorar a qualidade dos dados, que por vezes apresentam valores desconhecidos, incorretos, ou atributos que não contribuem para um modelo preditivo, entre outros (BATISTA, 2003).

No seu trabalho, Batista (2003) divide as tarefas de processamento de dados em dois grupos: tarefas fortemente e fracamente dependentes do conhecimento de domínio. O primeiro grupo refere-se a atividades que dependem de conhecimentos específicos como a identificação de inconsistências, atributos duplicados e redundantes, presença de dados distorcidos, verificação da integridade dos dados, detectar e descrever dados extremos, entre outros. Já as tarefas que dependem fracamente são aquelas que as informações necessárias para resolver o problema de pré-processamento podem ser extraídas dos próprios dados.

Silva, Peres e Boscarioli (2016) conotam o pré-processamento como parte do processo de Descoberta de Conhecimento de em Bases de Dados (KDD – Knowledge Discovery in Databases) proposto por Piatetsky- -Shapiro (1989). O objetivo desse processo é buscar padrões

nos dados por meio de um procedimento analítico, sistemático e automatizado, tornando-os mais facilmente assimiláveis para a obtenção de conhecimento. Para Silva, Peres e Boscarioli (2016), o pré-processamento inclui procedimentos como a organização dos dados em um único repositório, remoção de elementos duplicados e discrepantes, seleção de variáveis relevantes para mineração de dados e a normalização delas.

Neste estudo, após a coleta de dados, foi realizada a limpeza e padronização dos dados utilizando a linguagem de programação R. Inicialmente, foram removidos da base de dados os anúncios que não continham informações de Preço e Área, bem como os valores discrepantes presentes nessas variáveis e as informações que não se adequavam ao objetivo do estudo, como bairros que não pertenciam à cidade de Aracaju. Logo após, verificou-se a necessidade de agrupar as informações presentes na variável "Tipo", já que devido à coleta ter sido realizada em diferentes *websites*, cada um possuía uma padronização para o tipo de imóvel anunciado (ver Quadro 2).

Quadro 2 – Detalhamento da classificação dos tipos de imóveis à venda na cidade de Aracaju - SE.

Variável	Agrupamento
Apartamento	Apartamento Cobertura
	Apartamento Padrão
	Cobertura
Casa	Casa de Condomínio
	Casa Padrão

Fonte: Elaborado pela autora

Outra variável que precisou ser reorganizada foi "Bairro". Havia certa confusão entre os habitantes de Aracaju quanto à distinção e delimitação dos bairros e conjuntos. Portanto, foi necessário agrupar alguns bairros utilizando como referência os limites delineados em um mapa disponibilizado pela Prefeitura Municipal de Aracaju[1] (ver Quadro 3).

Por último, foi criada uma nova variável na qual os bairros foram agrupados em cinco zonas apresentadas no Quadro 4.

[1]Disponível em: <https://www.aracaju.se.gov.br/index.php?act=leitura&codigo=32496>

Quadro 3 – Descrição do agrupamento de bairros para estudo do preço de imóveis na cidade de Aracaju - SE .

Bairro	Conjuntos
Jabutiana	Santa Lúcia, Jabotiana, Conjunto Sol Nascente
José Conrado de Araújo	José Conrado de Araújo, Dom Pedro I
São Conrado	São Conrado, Orlando Dantas
Zona de Expansão	Aruana, Areia Branca, Gameleira, Robalo, Mosqueiro, Matapoã, São José dos Náufragos, Loteamento Beira Mar, Parque Santo Antônio, Loteamento Santa Maria

Fonte: Elaborado pela autora

Quadro 4 – Descrição do agrupamento de bairros em zonas para o estudo do preço de imóveis em Aracaju - SE.

Zonas	Bairros
Centro	Ponto Novo, América, Siqueira Campos, Jabutiana, Olaria, Capucho, José Conrado de Araújo, Novo Paraíso.
Norte	Porto Dantas, Soledade, Lamarão, Bugio, Jardim Centenário, Santos Dumont, Santo Antônio, Industrial, Cidade Nova, Dezoito do Forte, Palestina, Japãozinho, Dom Luciano.
Oeste	Getúlio Vargas, Cirurgia, Centro.
Sul	Atalaia, Coroa do Meio, Aeroporto, São José, Treze de Julho, Jardins, Luzia, Grageru, Pereira Lobo, Suíssa, Salgado Filho, Inácio Barbosa, Farolândia, São Conrado, Santa Maria.
Expansão	Aruana, Areia Branca, Gameleira, Robalo, Mosqueiro, Matapoã, São José dos Náufragos, Loteamento Beira Mar, Parque Santo Antônio, Loteamento Santa Maria

Fonte: Elaborado pela autora

4.2 Métodos

4.2.1 *Web Scraping*

O *Web Scraping* é a prática de coleta de dados realizada por meio da criação de um programa automatizado capaz de consultar um servidor *web*, solicitar os dados e, posteriormente, analisá-los para extrair as informações necessárias (MITCHELL, 2018).

Essa prática é comum quando há a necessidade de coletar grandes quantidades de dados, os quais podem estar distribuídos em várias páginas *web*. Fazer essa coleta manualmente seria inviável devido ao grande esforço e tempo necessários. Através do *Web Scraping*, é possível acessar dados que, de forma manual, seriam muito difíceis de obter, devido ao enorme volume

de informações disponíveis na internet.

Zhao (2017) divide o processo de *web scraping* em duas etapas: primeiro, obter os dados de páginas da internet, e em seguida, extrair as informações desejadas. Para isso, é necessário começar solicitando a HTTP da página através do *link* correspondente. Após a solicitação ser atendida, as informações serão enviadas para que possam ser trabalhadas. Ele ressalta ainda que os recursos obtidos podem estar em diversos formatos, dependendo de como os *websites* foram construídos, como HTML, XML, JSON, ou mesmo arquivos de dados multimídia.

No estudo de Khder (2021), são mencionadas algumas dificuldades que devem ser consideradas antes de aplicar o *Web Scraping*, como verificar se os dados estão apresentados em uma tabela no HTML e se é fácil utilizar os seletores CSS para extraí-los.

Zhao (2017) destaca o Beautiful Soup como uma ferramenta para extrair dados de documentos nos formatos HTML e XML. Além disso, ele ressalta dois módulos importantes para a coleta de dados: Urllib2 e Selenium. O primeiro define funções para requisitar informações de HTTP, tratar redirecionamentos, autenticações, cookies, entre outros. Já o segundo permite que o processo de navegação no site seja automatizado, pois o Selenium é capaz de simular cliques em páginas da *web*, bem como inserir conteúdos e percorrer todo o *site*.

Neste trabalho, o método de *Web Scraping* foi utilizado para a coleta de dados de cinco *websites*. Cada um desses *sites* foi construído utilizando diferentes formatos e linguagens, sendo necessárias diferentes abordagens para a coleta. Portanto, o uso das bibliotecas Beautiful Soup, Urllib2 e Selenium foi essencial para o processo de extração dos dados.

4.2.2 *Machine Learning*

Zhou (2017) define *Machine Learning* como uma técnica que melhora o desempenho de um sistema por meio da experiência adquirida através de métodos computacionais. A principal tarefa é desenvolver algoritmos de aprendizagem capazes de construir modelos a partir de dados.

O objetivo fundamental do *Machine Learning* é estudar e aprimorar modelos que podem ser treinados por meio de dados para fazer inferências sobre o futuro e tomar decisões sem possuir conhecimento completo sobre os fatores externos. Em outras palavras, um *software* que recebe informações sobre um certo assunto utiliza conhecimentos estatísticos para tentar determinar distribuições de probabilidade e usá-las para encontrar o valor ou decisão com maior acurácia (BONACCORSO, 2017).

Janiesch, Zschech e Heinrich (2021) destinguem *machine learning* em três tipos: Aprendizagem supervisionada, não supervisionada e por reforço.

Nos modelos de aprendizagem supervisionada, são separados um grupo de dados (treinamento) com informações de entrada e de saída (variáveis preditoras e resposta). Esses dados são utilizados para que o modelo aprenda e, em seguida, é fornecida uma medida precisa de seu erro para que posteriormente os parâmetros possam ser corrigidos e a função de perda global

seja reduzida. Esses modelos são suscetíveis a desenvolver *overfitting*, ou seja, a capacidade de generalização onde o modelo se adapta bem aos dados de treinamento mas falha em prever as demais (BONACCORSO, 2017).

Por outro lado, os modelos de aprendizagem não supervisionada são capazes de detectar padrões sem a necessidade de rótulos pré-existentes. Eles são bastante utilizados quando há a necessidade de agrupar elementos de acordo com a similaridade entre eles (BONACCORSO, 2017; JANIESCH; ZSCHECH; HEINRICH, 2021).

A aprendizagem por reforço tem como base o aprendizado através das próprias ações, as quais geram recompensas ou penalidades. Utiliza o princípio de tentativa e erro para adquirir informações e ser capaz de mapear estados e ações que possam aumentar a recompensa cumulativa, sendo este um parâmetro que indica se a ação foi boa ou ruim (CORRêA; MIRANDA, 2023).

O principal objetivo do *machine learning* é encontrar um modelo que preveja com precisão valores futuros a partir de dados existentes, ou seja, que seja capaz não somente de se ajustar bem aos dados existentes previamente, mas também aos futuros (BOEHMKE; GREENWELL, 2019).

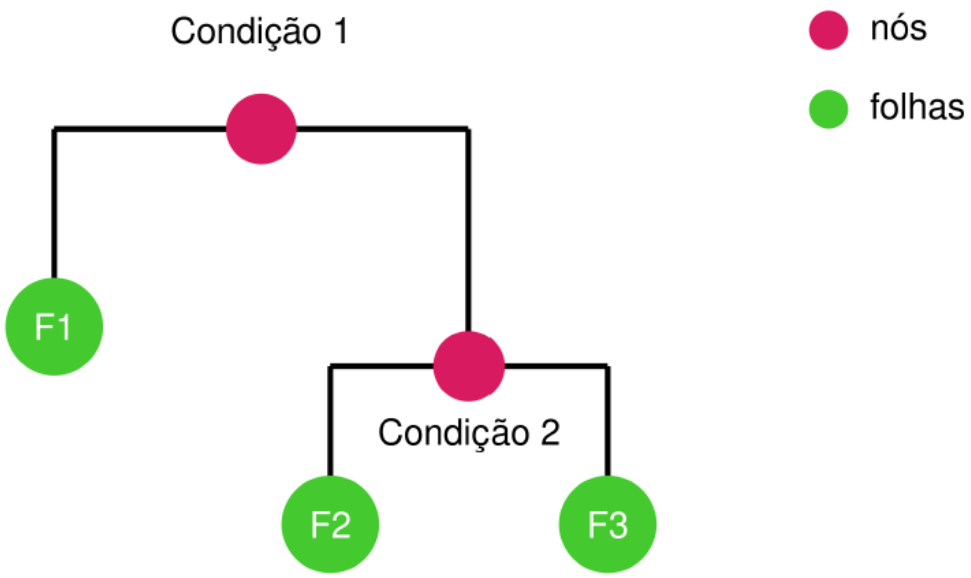
Para controlar a generalização dos modelos, as bases de dados são divididas em dois conjuntos: o de treinamento e o de teste. A separação dos grupos de dados geralmente ocorre em proporções de 60% - 40%, 70% - 30% ou 80% - 20% (BOEHMKE; GREENWELL, 2019). Essa separação deve ser analisada cautelosamente, pois a alta porcentagem de dados de treinamento podem levar o modelo ao *overfitting*, ou seja, ele se adaptará demais aos dados de treinamento e terá baixa capacidade de generalização. Por outro lado, quando há mais dados de teste do que de treinamento, o modelo não terá uma boa capacidade preditiva e não será suficiente para explicar bem os dados (*underfitting*) (IZBICKI; SANTOS, 2020).

4.2.2.1 Árvore de decisão

Modelos de árvores de decisão consistem na estratificação ou segmentação do espaço preditor em regiões menores e mais simples (JAMES et al., 2013). Nele é utilizada a estrutura de uma árvore para representar as possíveis relações entre as variáveis preditoras e os potenciais resultados, podendo eles serem discretos, como categorias ou rótulos, (árvores de classificação) ou contínuos (árvores de regressão) (NWANGANGA; CHAPPLE, 2020). Neste trabalho a variável resposta é quantitativa, sendo assim será utilizado o modelo de árvore de regressão, pertencente a classe de aprendizados supervisionados.

As árvores de regressão são modelos não paramétricos no qual a estrutura é construída por particionamentos que recebem o nome de nó e cada resultado deles são denominados folhas, como está ilustrada na Figura 1 (IZBICKI; SANTOS, 2020). Sua estrutura se assemelha a uma árvore invertida na qual se inicia com um nó que será particionado a cada decisão tomada baseada nos valores do preditor até que a última decisão seja tomada e assim gerando as folhas (NWANGANGA; CHAPPLE, 2020).

Figura 1 – Estrutura de um modelo de árvore de decisão



Fonte: (IZBICKI; SANTOS, 2020)

Para James et al. (2013), o processo de construção da estrutura de uma árvore se dá por dois passos:

1. Divisão do espaço preditor a partir dos possíveis valores $X_1, X_2, ..., X_p$ em J distintas e não sobrepostas regiões $Z_1, Z_2, ..., Z_j$.

2. Para cada observação na região Z_j, é realizada a predição, que será a média dos valores de resposta (Y) para as observações de treinamento em Z_j.

A predição da resposta é dada por (IZBICKI; SANTOS, 2020)

$$g(x) = \frac{1}{\left|\{i : x_p \in Z_j\}\right|} \sum_{i:x_p \in Z_j} y_i \tag{4.1}$$

Izbicki e Santos (2020) afirmam que a criação da estrutura da árvore de regressão é feita através de duas etapas: a criação de uma árvore completa e complexa e a poda da mesma, evitando assim o super ajuste. Na primeira etapa criam-se partições nas quais os valores de y sejam homogêneos em cada uma das folhas.

Logo após a criação, esses modelos são avaliados através do Erro Quadrático Médio (MSE),onde procura-se minimizá-lo. No entanto, a repetição do processo em busca do menor MSE muitas vezes é computacionalmente inviável, por isso são construídas divisões binárias particionando o espaço em duas regiões diferentes em busca da combinação cuja partição possua

predições com menor erro quadrático. Esse processo é repetido até que se contrua uma árvore com poucas observações em suas folhas (IZBICKI; SANTOS, 2020).

4.2.2.2 *Bagging*

Bagging é um método utilizado para aumentar o baixo poder preditivo de modelos afetados pela alta variância como a árvore de decisão, que podem retornar valores um pouco diferentes ao dividirmos os dados de treino aleatoriamente em duas partes e ajustar um modelo para elas (JAMES et al., 2013).

O método de *bagging* acontece a partir da construção de modelos ($\widehat{f}(x)$, $\widehat{f^2}(x)$, ..., $\widehat{f^B}(x)$) utilizando diferentes dados de treino (B) de uma população, dados dos quais são extraídas amostras utilizando o *bootstrap* e gerados B diferentes grupos de treinamento para que o método seja treinado b-ésimas vezes ($\widehat{f^{*b}}(x)$), logo após calculando a média deles para a obtenção um único modelo que possua baixa variância, dada por (JAMES et al., 2013)

$$\widehat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f^{*b}}(x). \tag{4.2}$$

4.2.2.3 *Random Forest* (Florestas Aleatórias)

Em seu livro, Kelleher, Mac Namee e D'Arcy (2015) definem os modelos de *random forest* como a combinação do *bagging*, sub amostras e árvores de decisão.

Para Pardalos et al. (2015), *Random Forest* é um método desenvolvido por Breiman que combina um conjunto de árvores de decisão no qual cada árvore é construída por um método determinístico que seleciona um grupo de variáveis e amostras aleatórias de um conjunto de treinamento. Para a construção do modelo é necessário a otimização de três parâmetros:

- *ntree*: é caracterizado pela quantidade de árvores a serem criadas. Não há um método para definição do número de árvores a serem construídas, usualmente base de dados pequenas e mais simples requerem poucas árvores, já as bases maiores e complexas demandam mais. Entende-se que quanto maior o número de árvores maior será a performance do modelo, no entanto aumenta também a tendência ao *overfiting* e o custo computacional (PROBST; BOULESTEIX; WRIGHT, 2018).

- *mtry*: trata-se da divisão aleatória do número de variáveis que serão consideradas para a geração dos nós. Modelos com alto valor de *mtry* tendem a produzir árvores mais diversificadas que podem sofrer *overfiting*. Já para aqueles com baixos valores de *mtry*, tendem a explorar com mais afinco variáveis com efeito moderado na variável resposta, no entanto induzem a ávores com desempenho baixo, uma vez que são construídas com base em variáveis que foram selecionadas de um conjunto menor, fazendo com que variáveis não importantes sejam escolhidas (PROBST; BOULESTEIX; WRIGHT, 2018). Para

problemas de regressão, o cáculo do *mtry* é dado por $mtry = \frac{p}{3}$, sendo p o número de variáveis preditoras.

- *nodesize*: consite no menor tamanho que o nó de uma árvore deverá ter, controlando o tamanho da mesma. Estimar baixos valores para o *nodesize* faz com que o custo computacional diminua e as árvores sejam mais profundas, ou seja, existam mais divisões até o nó final, já quando são estimadas altas quantidades, as árvores passam a ser mais curtas e tendem ao *overfiting*.

O algoritmo de um modelo *Random Forest* é dado por (IZBICKI; SANTOS, 2020):

1. Considere $z = 1, 2, 3, ..., Y$:
2. Cria-se Y amostra *bootstrap*, dos dados originais, de tamanho n;
3. Para cada amostra obtida, cria-se uma árvore $(g_y(x))$;
4. A função de predição é dada por: $g(x) = \frac{1}{Y} \sum_{y=1}^{Y} g_y(x)$.

4.2.2.4 Método de Avaliação de Modelos

Após a construção de um modelo *Random Forest*, é estimada a média de erro de predição de cada árvore através do erro quadrático médio (MSE - Mean Squared Error), conforme definido por Pardalos et.al. 2015:

$$MSE = n^{-1} \sum_{n=1}^{n} [\widehat{Y}(X_i) - Y_i]^2. \tag{4.3}$$

Onde, $\widehat{Y}(X_i)$ é o valor previsto correspondente a amostra de entrada, enquanto Y_i é o valor observado, sendo n o número total de amostras.

Kelleher, Mac Namee e D'Arcy (2015) apresentam uma crítica ao MSE, argumentando que em alguns cenários essa medida pode não ser suficientemente significativa. Por isso, sugerem o uso da raiz quadrada do erro quadrático médio (RMSE - Root Mean Squared Error), que é calculada como:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{n} [\widehat{Y}(X_i) - Y_i]^2}{n}}. \tag{4.4}$$

Essas medidas possuem como objetivo avaliar o desempenho do modelo proposto em determinado banco de dados. Por meio delas é possível medir o quão bem os valores previstos correspondem aos observados (JAMES et al., 2013). Tanto o MSE quanto o RMSE serão grandes se os valores previstos e os valores reais forem muito diferentes. Portanto, quanto menor o valor

do MSE e do RMSE, melhor será o desempenho preditivo do modelo. Neste trabalho, o RMSE será utilizado como medida de avaliação de modelos preditivos.

4.3 Suporte Computacional

Neste trabalho, foram utilizadas duas linguagens de programação, ambas de código aberto: Python e R.

4.3.1 Python

A linguagem de programação Phyton (versão 3.10.11) e o ambiente de desenvolvimento Pycharm (edição 2022.3.1) foram utilizados na etapa de *Web Scraping*.

No processo de *web scraping* foram utilizados os pacotes *requests* para a requisição de acesso ao conteúdo dos *sites*, *Beautifulsoup* para análise e extração de informações de arquivos $HTML$ e XML, e $Selenium$ utilizada na coleta e automação da mesma em *sites* renderizados via JavaScript. Também foi utilizado nessa etapa a biblioteca $pandas$ para organização dos dados extraídos em data frame.

4.3.2 *Software R*

O *software R* (versão 4.2.2) e o ambiente de desenvolvimento $RStudio$ (edição 2023.06.0) foram utilizados nas etapas de pré-processamento, análise dos dados e construção de modelos.

Foram utilizadas as bibliotecas $tidyverse$, para a limpeza e análise dos dados, $osmdata$, sf e $leaflet$ na elaboração do mapa e para a construção dos modelos de *Random Forest* utilizou-se os pacotes $caret$, $ranger$, $rsample$ e vip.

5 RESULTADOS

5.1 Análise Descritiva

Após coleta e tratamento dos dados, foram identificadas 1103 anúncios de imóveis disponíveis no dia 29 de janeiro. Desse total, 611 são de apartamentos e 492 casas. Estas propriedades estão distribuídas entre 40 bairros de Aracaju agrupados em 5 regiões, sendo as regiões com maior quantidade de imóveis à venda a Zona de Expansão, Zona Oeste e a Zona Sul como é possivel observar na Figura 2 e na Tabela 1 .

Tabela 1 – Frequência absoluta e percentual de imóveis na cidade de Aracaju - SE considerando as regiões da variável Zona.

Zona	**Freq.**	**(%)**
Centro	56	5,08
Zona de Expansão	117	10,61
Zona Norte	58	5,26
Zona Oeste	154	13,29
Zona Sul	718	65,09
Total	**1103**	**100,00**

Fonte: Elaborado pela autora.

A análise dos preços dos imóveis revelou que o valor médio de uma propriedade em Aracaju é de aproximadamente R$ 561.000,00 com uma variação entre R$ 80.000,00 e R$ 2.000.000,00. Os apartamentos têm uma média de preço de R$ 457.887,00 , enquanto as casas têm uma média de R$689.922,00, conforme demonstrado na Tabela 2.

Tabela 2 – Análise descritiva do preço de imóveis à venda na cidade de Aracaju - SE considerando a variável Tipo.

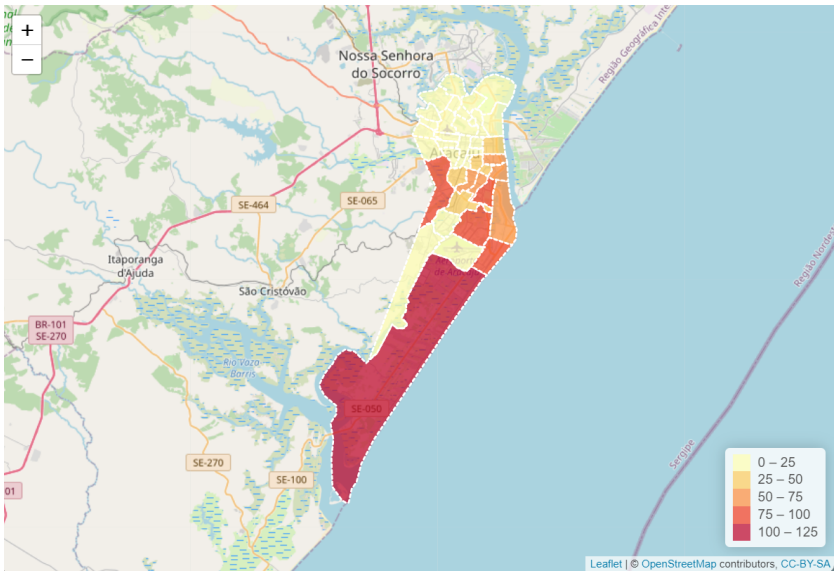
Tipo	Mínimo	Média	Mediana	Máximo	Amplitude
Apartamento	80.000,00	457.887,00	360.000,00	2.000.000,00	1.920.000,00
Casa	110.000,00	689.922,00	550.000,00	2.000.000,00	1.890.000,00

Fonte: Elaborado pela autora.

Na Figura 3, é possível observar que os preços de venda das casas são mais elevados do que os dos apartamentos em todas as zonas. Além disso, constata-se que o custo habitacional na Zona de Expansão e na Zona Sul é mais alto, com uma média de aproximadamente R$ 900.000,00 e R$ 600.000,00, respectivamente. Observa-se também que as zonas Norte e Oeste têm construções com valores mais baixos, com uma média estimada de R$ 257.000,00 e R$ 310.000,00, nessa ordem.

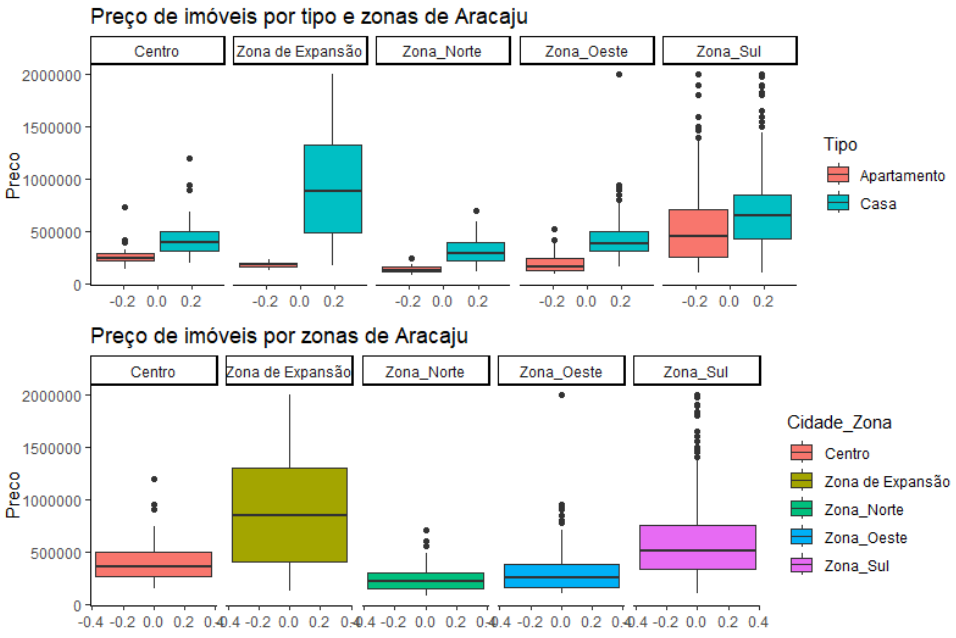
Em um estudo realizado por Santos (2023), que visava identificar a influência da distância da praia no valor das residências em Aracaju, foi observado que quanto mais próximos os imóveis

Figura 2 – Distribuição da quantidade de imóveis à venda diponíveis nos bairros da cidade de Aracaju - SE considerando a variável Cidade.



Fonte: Elaborado pela autora.

Figura 3 – Análise do preço dos imóveis à venda na cidade de Aracaju - Sergipe considerando as variáveis Tipo e Zonas.

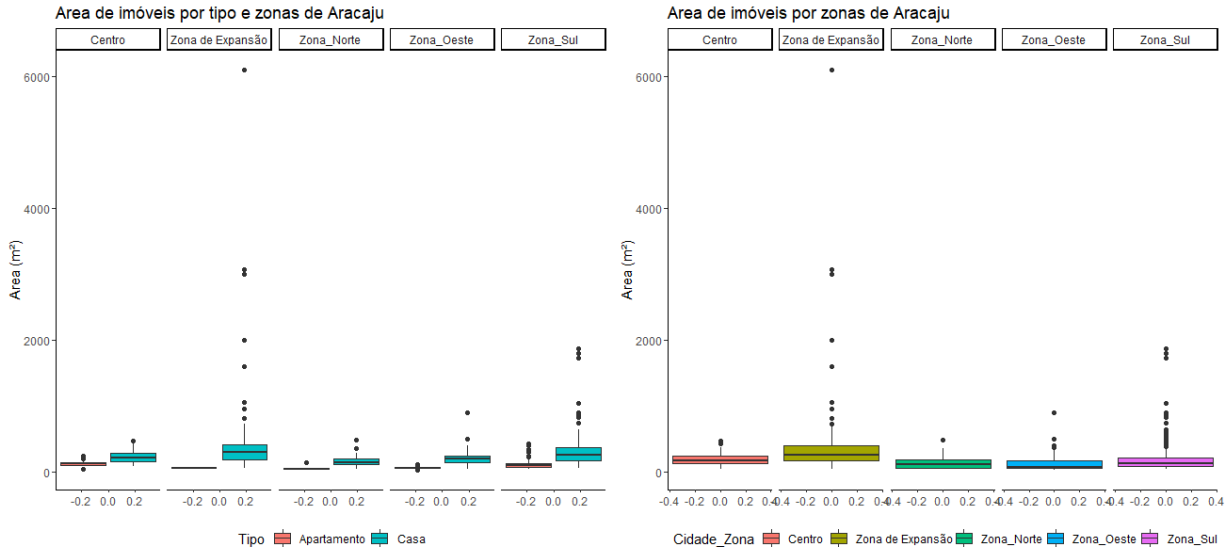


Fonte: Elaborado pela autora.

estão localizados da área litorânea da cidade, especialmente nos bairros Atalaia e Coroa do Meio, localizados na zona sul, maior é a valorização imobiliária, resultando em valores unitários mais elevados para os imóveis.

Explorando a variável Área, observa-se que as habitações em Aracaju possuem, em média, 200 m^2 e variam de 40 a 6100 m^2.Foi notado que as casas tendem a ter uma área maior do que os apartamentos, com a Zona de Expansão apresentando as maiores habitações, com imóveis variando de 44 m^2 a 6100 m^2. Enquanto isso, a Zona Oeste, Norte e Centro são caracterizadas por residências de menor tamanho, variando de 40 a 900 m^2, como detalhado na Figura 4.

Figura 4 – Análise da área dos imóveis à venda na cidade de Aracaju - SE considerando as variáveis Tipo e a Zonas.



Fonte: Elaborado pela autora.

Quanto a quantidade de quartos, a média por habitação no municipio é de 3 quartos podendo variar de 1 a 11. Na Zona de Expansão e na Zona Sul podem ser encontrados imóveis com 10 e 11 quartos, enquanto nas demais zonas os imóveis podem ter até 6 quartos.

Outras variáveis a serem analisadas são a quantidade de banheiros e suítes em um domicílio. Em Aracaju, observou-se que os imóveis à venda apresentam, em média, 2 banheiros. Somente na Zona de Expansão essa média é elevada para 3, enquanto nas demais zonas segue a média do município. Quanto à quantidade de suítes, em média há 1 suíte por imóvel, podendo ser encontrados imóveis sem suítes ou com até 7 suítes. Nas zonas norte e oeste, a média de suítes por habitação é inferior a 1, indicando que muitas casas não possuem suítes.

Analisando a variável "Garagem", é possível observar que a quantidade de vagas de garagem por imóvel variam de 0 a 15. No Centro, Zonas Norte e Oeste a média é de 1 garagem por habitação, já na Zona Sul e de Expansão é de 2 e e 3 respectivamente. No entanto, é possivel encontrar nessas regiões imóveis com 14 e 15 vagas de garagem, como podemos observar na

Tabela 3.

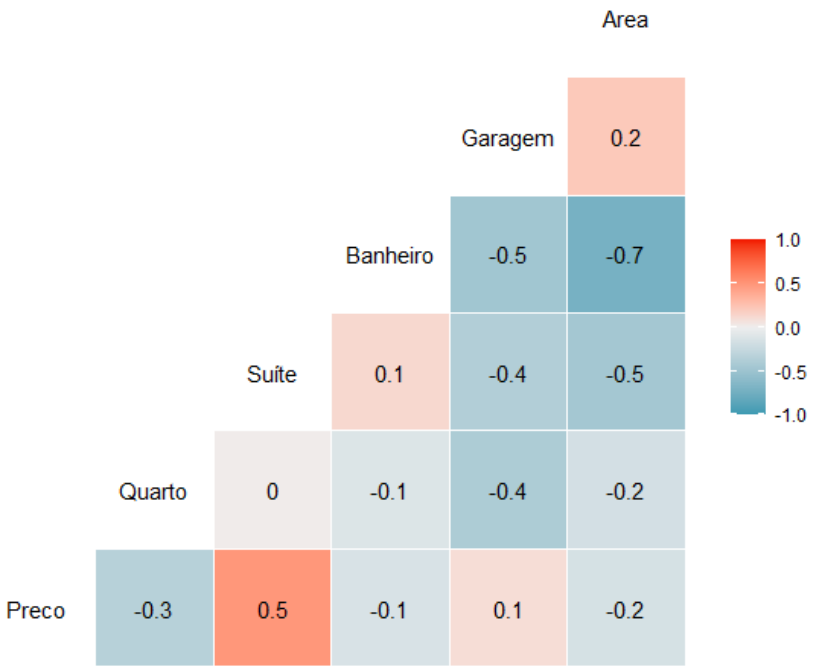
Tabela 3 – Análise descritiva da quantidade de garagens por imóvel à venda em Aracaju

Zona	Mínimo	Média	Máximo	Desvio Padrão
Centro	0	1,62	8	1,73
Expansão	0	3,18	15	2,47
Norte	0	1,28	5	1,06
Oeste	0	1,51	10	1,17
Sul	0	2,12	14	1,66

Fonte: Elaborado pela autora.

Na Figura 5, pode-se observar a correlação das variáveis quantitativas coletadas. Destaca-se a presença de uma correlação positiva entre a variável "Preço"e "Suíte", indicando que a quantidade de suítes em um imóvel influencia no preço do mesmo. Nesse caso, a existência de suítes aumenta o preço da residência.

Figura 5 – Resultado da correlação das variáveis coletadas de anúncios de venda de imóveis na cidade de Aracaju - SE.



Fonte: Elaborado pela autora.

5.2 Modelo *Random Forest*

Para a construção do modelo de *Random Forest*, foram utilizadas as variáveis: preço, tipo, area, quartos, banheiro, suíte, garagem e cidade_zona.

Utilizando as váriáveis disponíveis, foi construído um modelo de *random forest* no qual foram ajustados os hiperparametros mencionados no Capítulo 4 a fim de se obter o modelo que melhor previsse o preço de um imóvel. O resultado pode ser observado na Tabela 4, na qual estão listados os 10 melhores modelos, ordenados pelo menor erro quadrático médio (RMSE) e melhor percentual de ganho.

Tabela 4 – Resultado dos 10 melhores modelos utilizando o algoritimo do modelo Random Forest aplicado ao preço de venda de imóveis em Aracaju - SE.

	mtry	min.node.size	replace	sample.fraction	RMSE	perc_gain
1	5	1	TRUE	0,5	210918,82	1,4931
2	5	5	TRUE	0,5	211711,19	1,1230
3	5	5	TRUE	0,63	211771,70	1,0947
4	5	3	TRUE	0,5	211830,36	1,0674
5	5	3	TRUE	0,63	212013,82	0,9817
6	5	1	TRUE	0,63	212088,82	0,9467
7	5	10	TRUE	0,63	212169,71	0,9089
8	4	3	TRUE	0,5	212306,95	0,8448
9	3	1	TRUE	0,5	212828,36	0,6013
10	4	5	FALSE	0,5	213209,52	0,4233

Fonte: Elaborado pela autora.

O modelo 1, que apresenta $mtry = 5$, número mínimo de nó igual a 1 e utiliza 50% da amostra para o treinamento do modelo utilizando a amostragem por reposição, foi o escolhido por apresentar melhores métricas. Verificou-se também para a avaliação do modelo o RMSE nesse caso igual a 210.918,82.

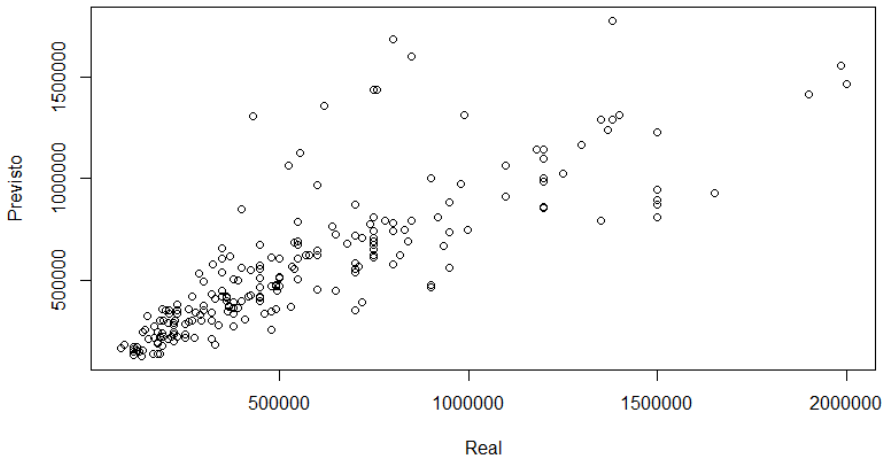
Após a construção e escolha do melhor modelo, foram inseridos novos dados presentes na base de teste com o intuito de observar como o modelo se comporta na presença de novas informações, a comparação entre os valores observados (base de treino) e previstos (base de teste). Através da Figura 6, é possível observar que os valores previstos e os observados apresentam bons desempenhos nos imóveis com preços mais baixos, prevendo bem imóveis de até aproximadamente R$ 1.000.000,00.

Na Figura 7, observa-se a importância de cada variável para o modelo ajustado escolhido. As variáveis Área, Suíte e Garagem se destacam como as principais preditoras do modelo, enquanto as demais possuem menor impacto no mesmo.

Amaral (2018) e Zaghi (2023) também identificaram em seus trabalhos o tamanho do imóvel como a variável mais importante para o modelo de predição utilizado. No entanto, Amaral (2018) obteve o número de banheiros como a segunda variável mais importante, enquanto o

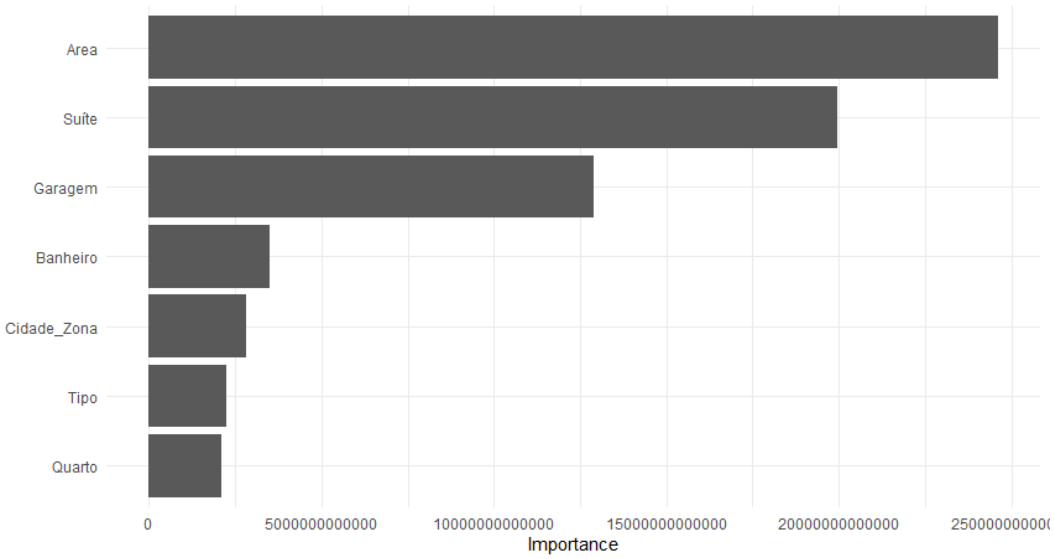
modelo de predição de Zaghi (2023) teve a variável "Bairro Jurerê Internacional"como a segunda mais relevante. Ele atribuiu esse fato a fatores relacionados à segurança, os quais não puderam ser mapeados na variável bairro, mas estão diretamente ligados ao preço do imóvel. Neste estudo, como os bairros foram agrupados em Zonas devido à escassez de informações e/ou à falta de detalhes mais específicos sobre as regiões dos imóveis, as variáveis de área, presença de suíte, quantidade de garagens e banheiros demonstram uma relevância maior na construção do modelo.

Figura 6 – Comparação entre os valores Previstos e Observados do preço de casas à venda na cidade de Aracaju - SE a partir do melhor modelo de Random Forest.



Fonte: Elaborado pela autora.

Figura 7 – Classificação de importância das variáveis para o modelo Random Forest aplicado ao preço de venda de imóveis em Aracaju - SE.



Fonte: Elaborado pela autora.

6 CONCLUSÃO

Através deste trabalho buscou-se construir um modelo utilizando o algoritmo de *machine learning Random Forest* capaz de precificar o valor de venda de um imóvel na cidade de Aracaju - SE, utilizando dados obtidos por meio da técnica de *web scraping*.

Por meio da análise dos dados obtidos, foi possível identificar anúncios de venda de imóveis em 40 bairros de Aracaju, posteriormente agrupados em Zonas. Durante a análise descritiva, observou-se que o preço de venda de um imóvel varia de R$ 80.000,00 a R$ 2.000.000,00, sendo o preço médio de venda de apartamentos menor do que o de casas. Além disso, o preço médio de venda de imóveis nas zonas de Expansão e Sul é maior do que nas outras. Também foi percebido que a área dos imóveis no município varia de 40 a 6100 m^2, sendo as casas maiores do que os apartamentos.

Notou-se que a quantidade média de quartos é de 3 por habitação, no entanto, na Zona Sul e de Expansão, é possível encontrar imóveis com 10 e 11 quartos. Ao observar a quantidade de banheiros por residência, percebeu-se que a média na Zona de Expansão (2) é maior do que a do município (3). A quantidade de suítes, um fator com bastante influência no valor de um imóvel nesse estudo, oscila entre nenhuma e 7 por imóvel, estando pouco presente em imóveis da zona norte e oeste. O número de garagens por habitação em Aracaju pode chegar a 15 vagas na Zona de Expansão e Sul, enquanto nas demais zonas a média é de 1 por construção.

Quanto a aplicação do algoritmo de *machine learning*, foi observado que o modelo *Random Forest* construído é capaz de prever bem valores de venda de imóveis a partir de novas informações até R$ 1.000.000,00 e a variável que possui maior influência para o modelo é a "Área".

6.1 Pesquisas futuras

Ao longo da pesquisa, percebeu-se que as características das habitações e o ambiente em que estão localizadas influenciam no seu valor de mercado. Sendo assim, a ausência de informações a respeito dos imóveis diminui a performance do modelo de predição. Observou-se também que existem outros algoritmos de *machine learning* que podem obter melhores desempenhos na predição e precificação de imóveis. Por fim, notou-se que é necessário um modo de apresentação dos resultados obtidos mais eficiente, de forma que as pessoas possam utilizá-lo no processo de compra ou venda de um imóvel. Neste contexto, em trabalhos futuros, surge a necessidade de:

- Obter mais informações sobre os imóveis, como a localização exata (latitude e longitude), a posição do imóvel em relação ao sol, o andar do apartamento e o valor da taxa de

condomínio, bem como a presença de elevadores;

- Obter informações acerca da vizinhança, como a existência de escolas, supermercados, hospitais e parques nas proximidades do imóvel;
- Propor novos modelos de *machine learning* para compará-los;
- Construir um aplicativo *web* para apresentação dos resultados de forma dinâmica.

REFERÊNCIAS

ABECIP, A. B. d. E. d. C. I. e. P. *Antes de queda de juros, preço de imóveis já se valoriza e supera inflação (InfoMoney)*. 2023. Acessado: 2023-09-24T16:15:33Z. Disponível em: <https://www.abecip.org.br/imprensa/noticias/antes-de-queda-de-juros-preco-de-imoveis-ja-se-valoriza-e-supera-inflacao-infomoney>. Citado na página 14.

ABRAINC, A. B. d. I. I. *Em coletiva de imprensa, ABRAINC apresenta resultados consolidados do mercado imobiliário em 2022*. 2022. Acessado: 2023-09-24T16:05:53Z. Disponível em: <https://www.abrainc.org.br/abrainc-news/2023/03/29/em-coletiva-de-imprensa-abrainc-apresenta-resultados%C2%A0consolidados-do-mercado-imobiliario-em-2022>. Citado na página 14.

AMARAL, A. S. d. *Uma Metodologia Orientada a Dados para Precificação de Imóveis*. 2018. Disponível em: <https://repositorio.ufrn.br/handle/123456789/43618>. Citado na página 34.

ARARUNA, R. S. *Análise imobiliária : qual o melhor método para prever o valor de um imóvel?* 2022. 87 p. Disponível em: <https://bdm.unb.br/handle/10483/34271>. Citado 2 vezes nas páginas 19 e 20.

ARRAES, R.; FILHO, E. Externalidades e formação de preços no mercado imobiliário urbano brasileiro: Um estudo de caso. *Economia Aplicada*, v. 12, 01 2008. Citado na página 19.

BATISTA, G. *Pré-processamento de dados em aprendizado de máquina supervisionado.* Tese (Doutorado), 01 2003. Citado na página 21.

BOEHMKE, B.; GREENWELL, B. *Hands-On Machine Learning with R*. [S.l.: s.n.], 2019. ISBN 9780367816377. Citado na página 25.

BONACCORSO, G. *Machine learning algorithms: a reference guide to popular algorithms for data science and machine learning*. [S.l.]: Packt, 2017. 343 p. ISBN 9781785889622. Citado 2 vezes nas páginas 24 e 25.

CARVALHO, L. F. B. d. *Identificação do percentual ideal a ser utilizado como fator na avaliação de imóveis na grande Aracaju-SE: Tipo, idade, área e padrão construtivo dos imóveis*. 2019. 78 p. Disponível em: <https://ri.ufs.br/jspui/handle/riufs/15668>. Citado na página 17.

CARVALHO, S.; MEDEIROS, R. Modelagem econométrica do preço de alugueis de apartamentos na cidade de petrópolis-rj utilizando regressão linear múltipla. 08 2017. Citado na página 14.

CORRêA, F. S.; MIRANDA, P. H. C. *Aprendizado por reforço aplicado em jogo de FPS.* 2023. Disponível em: <http://hdl.handle.net/11449/239535>. Citado na página 25.

DANTAS, R.; MAGALHãES, A.; VERGOLINO, J. Avaliação de imóveis: A importância dos vizinhos no caso de recife. *Economia Aplicada*, v. 11, 04 2007. Citado na página 19.

FARIAS, B. M. d. C. *A evolução do mercado imobiliário brasileiro e o conceito de Home Equity*. 2010. Citado na página 18.

FRANçA, S. L. A.; REZENDE, V. L. F. Produção do espaço urbano e novos eixos imobiliários em aracaju-se, brasil: mercado e estado. In: *XI Simposio de la Asociación Internacional de Planificación Urbana y Ambiente (UPE 11) (La Plata, 2014)*. [s.n.], 2014. p. 890–900. Disponível em: <http://sedici.unlp.edu.ar/handle/10915/55466>. Citado 2 vezes nas páginas 18 e 19.

IBGE, I. B. d. G. e. E. *Panorama de Aracaju - SE*. 2022. Acessado: 26 set. 2023. Disponível em: <https://cidades.ibge.gov.br/brasil/se/aracaju/panorama>. Citado na página 18.

IZBICKI, R.; SANTOS, T. M. d. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. 254 p. ISBN 978-65-00-02410-4. Citado 4 vezes nas páginas 25, 26, 27 e 28.

JAMES, G. et al. *An Introduction to Statistical Learning*. [S.l.: s.n.], 2013. 426 p. ISBN 1461471389. Citado 4 vezes nas páginas 25, 26, 27 e 28.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. *Electronic Markets*, 04 2021. Citado 2 vezes nas páginas 24 e 25.

KELLEHER, J.; NAMEE, B. M.; D'ARCY, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. [S.l.: s.n.], 2015. ISBN 9780262029445. Citado 2 vezes nas páginas 27 e 28.

KHDER, M. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and its Applications*, v. 13, p. 145–168, 12 2021. Citado na página 24.

KREMER, J. *Mercado Imobiliário*. [S.l.]: Grupo UNIASSELVI, 2008. 45 p. ISBN 978-85-7830-090-6. Citado 2 vezes nas páginas 17 e 18.

LACERDA, A. C. d. et al. *Economia brasileira*. [S.l.]: Editora Saraiva, 2006. Citado na página 17.

LEEUW, F. d. A price index for new multifamily housin. *Bureau of Economic Analysis*, 1993. Disponível em: <apps.bea.gov/scb/pdf/NATIONAL/NIPA/1993/0293dlw.pdf>. Citado na página 14.

MATOS, D.; BARTKIW, P. I. N. *Introdução ao Mercado Imobiliário*. [S.l.]: INSTITUTO FEDERAL DO PARANÁ – EDUCAÇÃO A DISTÂNCIA e-Tec Brasil, 2013. Citado 3 vezes nas páginas 14, 17 e 18.

MITCHELL, R. *Web Scraping with Python: Collecting More Data from the Modern Web*. [S.l.]: OReilly Media, 2018. 306 p. ISBN 978-1-491-98557-1. Citado na página 23.

NUNES, D. et al. O poder da nova classe média na estratégia de brand das construtoras de sergipe. estudo de caso da empresa celi. 09 2011. Citado na página 14.

NWANGANGA, F.; CHAPPLE, M. *Practical Machine Learning in R*. [S.l.: s.n.], 2020. ISBN 9781119591542. Citado na página 25.

PARDALOS, P. et al. *Machine Learning, Optimization, and Big Data*. 1. ed. Springer International Publishing, 2015. 387 p. ISBN 9783319279251. Disponível em: <http://link.springer.com/10.1007/978-3-319-27926-8>. Citado 2 vezes nas páginas 27 e 28.

PROBST, P.; BOULESTEIX, A.-L.; WRIGHT, M. Hyperparameters and tuning strategies for random forest. 04 2018. Citado na página 27.

SANTOS, A. A. S. d. *Aplicação da teoria de preços hedônicos na análise da influência da distância à praia no valor dos imóveis residenciais em Aracaju/SE*. 2023. 76 p. Disponível em: <https://ri.ufs.br/jspui/handle/riufs/17786>. Citado 3 vezes nas páginas 18, 19 e 31.

SILVA, L.; PERES, S.; BOSCARIOLI, C. *Introdução a Mineração de Dados com aplicações em R*. [S.l.: s.n.], 2016. ISBN 9788535284478. Citado 2 vezes nas páginas 21 e 22.

XU, K.; NGUYEN, H. Predicting housing prices and analyzing real estate markets in the chicago suburbs using machine learning. *Journal of Student Research*, v. 11, 03 2022. Citado na página 20.

ZAGHI, L. M. *Modelo de previsão de preços de imóveis na cidade de Florianópolis/SC a partir de técnicas de machine learning*. 2023. Disponível em: <https://repositorio.ufsc.br/handle/123456789/248689>. Citado 3 vezes nas páginas 20, 34 e 35.

ZHAO, B. Web scraping. In: ____. [S.l.: s.n.], 2017. p. 1–3. ISBN 978-3-319-32001-4. Citado na página 24.

ZHOU, Z.-H. Machine learning research: some recent progress in china and beyond. *National Science Review*, v. 5, 12 2017. Citado na página 24.

APÊNDICE A – Código do pré - processamento, análise dos dados e construção do modelo de *Random Forest*

```
### Banco de dados web scraping
df <- readxl::read_xlsx('DADOS_IMOB_.xlsx')
#### Data Cleaning ####

df$Tipo <- as.factor(df$Tipo)
df$Cidade <- as.factor(df$Cidade)
df$Preco <- as.numeric(df$Preco)

df <- df %>% filter(Area!= 0)
df <- df %>% filter(Preco <= 2000000)

# Classificações de Tipo

df$Tipo[df$Tipo %in% c('Apartamento Cobertura','Apartamento
   Padrão','Cobertura')] = 'Apartamento'
df$Tipo[df$Tipo %in% c('Casa de Condomínio','Casa Condomínio','
   Casa Padrão')] = 'Casa'
df$Tipo = df$Tipo[df$Tipo != 0]
df = df %>% filter(Tipo %in% c('Casa','Apartamento'))

# Classificação de Cidade

df$Cidade[df$Cidade %in% c('Areia Branca','Gameleira-Robalo','
   Mosqueiro','Mosqueiro
(Loteamento Santa Maria)','Mosqueiro (Parque Santo Antonio)', '
   Robalo','Zona de Expansão (Areia Branca)',
'Zona de Expansão(Mosqueiro)','Zona de Expansão (Robalo)', '
   Matapoã',
'São Josédos Náufragos','Zona de Expansão (Aruana)', 'Aruana','
   Loteamento
Beira Mar')] = 'Zona de Expansão'

df$Cidade[df$Cidade == '13 de Julho'] = 'Treze de Julho'
df$Cidade[df$Cidade == 'Suissa'] = 'Suíssa'
```

```
df$Cidade[df$Cidade %in% c('Santa Lúcia','Jabotiana', 'Conjunto
    Sol Nascente')] = 'Jabutiana'
df$Cidade[df$Cidade == '18 do Forte'] = 'Dezoito do Forte'
df$Cidade[df$Cidade == 'Dom Pedro I'] = 'José Conrado de Araújo
   '
df$Cidade[df$Cidade == 'Marivan'] = 'Santa Maria'
df$Cidade[df$Cidade =='Orlando Dantas'] = 'São Conrado'
df = df[!df$Cidade == 'Conjunto Marcos Freire 2',]

df = df[!df$Cidade %in% c('Fernando Collor','Barra dos
   Coqueiros','Parque dos Farois','São Cristovão'),]

df['Cidade_Zona'] = df$Cidade

df$Cidade_Zona[df$Cidade_Zona %in% c('Atalaia','Coroa do Meio',
   'Aeroporto','São José', 'Treze de Julho', 'Farolandia','
   Jardins','Luzia','Grageru','Pereira Lobo','Suíssa','Salgado
   Filho', 'Inácio Barbosa','São Conrado', 'Santa Maria')] = '
   Zona_Sul'

df$Cidade_Zona[df$Cidade_Zona %in% c('Porto Dantas','Soledade',
   'Lamarão','Bugio','Jardim Centenario','Santos Dumont','Santo
    AntÃ´nio','Industrial','Cidade Nova','Dezoito do Forte', '
   Palestina', 'Japãozinho','Dom Luciano')] = 'Zona_Norte'

df$Cidade_Zona[df$Cidade_Zona %in% c('Ponto Novo','América','
   Siqueira Campos','Jabutiana','Olaria','Capucho','José
   Conrado de Araújo','Novo Paraíso')] = 'Zona_Oeste'

df$Cidade_Zona[df$Cidade_Zona %in%  c('Getúlio Vargas','
   Cirurgia','Centro')] = 'Centro'

write.csv(df,'df_mod.csv',row.names = FALSE)

#### Analise Descritiva ####

df <- read.csv('df_mod_alt.CSV')

summary(df)
```

```
55
56 ### Correlação
57
58 df1 <- df %>% select(.,-Tipo,-Cidade_Zona) %>% na.omit()
59 knitr::kable(cor(df1))
60 ggcorr(cor(df1), label = TRUE)
61
62 df2 <- df %>% select(.,-Cidade,-Codigo,-Url) %>%
63   na.omit()
64
65 ggplot(df2,aes(y=Preco,fill=Cidade_Zona))+
66   geom_boxplot()+
67   facet_grid(~Cidade_Zona)+
68   theme_classic()
69 ## Zona sul e Expansao variacao alta, outliers zona sul
70
71 ggplot(df2,aes(y=Preco,fill=Tipo))+
72   geom_boxplot()+
73   facet_grid(~Cidade_Zona)+
74   theme_classic()
75
76 ggplot(df2,aes(y=Area,fill=Tipo))+
77   geom_boxplot()+
78   facet_grid(~Cidade_Zona)+
79   theme_classic()
80
81 ggplot(df2,aes(y=Area,fill=Cidade_Zona))+
82   geom_boxplot()+
83   facet_grid(~Cidade_Zona)+
84   theme_classic()
85 ## Outliers Zonal sul e Expansao
86
87 ggplot(df2,aes(y=Quarto,fill=Cidade_Zona))+
88   geom_boxplot()+
89   facet_grid(~Cidade_Zona)+
90   theme_classic()
91
92 ggplot(df2,aes(y=Suíte,fill=Cidade_Zona))+
93   geom_boxplot()+
```

```
94    facet_grid(~Cidade_Zona)+
95    theme_classic()
96
97  ggplot(df2,aes(y=Banheiro,fill=Cidade_Zona))+
98    geom_boxplot()+
99    facet_grid(~Cidade_Zona)+
100   theme_classic()
101
102 ggplot(df2,aes(y=Garagem,fill=Cidade_Zona))+
103   geom_boxplot()+
104   facet_grid(~Cidade_Zona)+
105   theme_classic()
106
107 library(psych)
108 describe(df2)
109
110 ### Modelos Random Forest
111
112 # Divisao dos dados
113
114 set.seed(123)
115 split <- initial_split(df2, prop = 0.8,
116                        strata = "Preco")
117
118 house_train <- training(split)
119 house_test <- testing(split)
120
121 n_features <- length(setdiff(names(df2), "Preco"))
122
123 # Modelagem
124 mod1 <- ranger(
125   Preco ~ .,
126   data = house_train,
127   mtry = floor(n_features / 3),
128   respect.unordered.factors = "order",
129   seed = 123
130 )
131
132 # get OOB RMSE
```

```r
(default_rmse_mod1 <- sqrt(mod1$prediction.error))
# [1] 214115.8

### Hiperparametro
# Ajustando hiperparametos
hyper_grid_mod1 <- expand.grid(
  mtry = floor(n_features * c(.15, .4, .5, .6,.8)),
  min.node.size = c(1, 3, 5, 10),
  replace = c(TRUE, FALSE),
  sample.fraction = c(.5, .63, .8),
  rmse = NA
)

for(i in seq_len(nrow(hyper_grid_mod1))) {
  # fit model for ith hyperparameter combination
  fit <- ranger(
    formula         = Preco ~ .,
    data            = house_train,
    num.trees       = n_features * 10,
    mtry            = hyper_grid_mod1$mtry[i],
    min.node.size   = hyper_grid_mod1$min.node.size[i],
    replace         = hyper_grid_mod1$replace[i],
    sample.fraction = hyper_grid_mod1$sample.fraction[i],
    verbose         = FALSE,
    seed            = 123,
    respect.unordered.factors = 'order',
  )
  # export OOB error
  hyper_grid_mod1$rmse[i] <- sqrt(fit$prediction.error)
}

hyper_grid_mod1 %>%
  arrange(rmse) %>%
  mutate(perc_gain = (default_rmse_mod1 - rmse) / default_rmse_
     mod1 * 100) %>%
  head(10)

## Ajustando melhor modelo

```

```
mod_final <- ranger(
  Preco ~ .,
  data = house_train,
  mtry = 5,
  num.trees = n_features * 10,
  min.node.size = 1,
  replace = TRUE,
  sample.fraction = 0.50,
  respect.unordered.factors = "order",
  seed = 123)

(sqrt(mod_final$prediction.error))

### Avaliando na base de teste

pred <- predict(mod_final,
                data = house_test)

pred$predictions


# Calcular RMSE
(rmse <- sqrt(mean((pred$predictions - house_test$Preco)^2)))

# Calcular MAE na escala original
(mae <- mean(abs(pred$predictions - house_test$Preco)))

comp<-data.frame(Pred=pred$predictions,
                 real=house_test$Preco)

plot(comp$real,comp$Pred, xlab = 'Real', ylab = "Previsto")

## Variaveis de importancia

mod_final_imp <- ranger(
  Preco ~ .,
  data = house_train,
  mtry = 5,
  num.trees = n_features * 10,
```

```
210   min.node.size = 1,
211   replace = TRUE,
212   importance = "impurity",
213   sample.fraction = 0.50,
214   respect.unordered.factors = "order",
215   seed = 123)
216
217 vip::vip(mod_final_imp, num_features = 25, bar = FALSE)+
218   theme_minimal()
```