UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
ENGENHARIA DE PRODUÇÃO CIVIL

Mateus Mota Gonçalves

PRECIFICAÇÃO DE IMÓVEIS UTILIZANDO REGRESSÃO LINEAR MÚLTIPLA E ÁRVORES DE DECISÃO

Florianópolis
2022

Mateus Mota Gonçalves

PRECIFICAÇÃO DE IMÓVEIS UTILIZANDO REGRESSÃO LINEAR MÚLTIPLA E ÁRVORES DE DECISÃO

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia de Produção e Sistemas da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Engenheiro Civil, do curso de Engenharia, área Civil, habilitação Engenharia de Produção Civil.
Orientador: Prof. Carlos Ernani Fries, Dr.

Florianópolis
2022

Ficha de identificação da obra

Gonçalves, Mateus Mota
PRECIFICAÇÃO DE IMÓVEIS UTILIZANDO REGRESSÃO LINEAR MÚLTIPLA E ÁRVORES DE DECISÃO / Mateus Mota Gonçalves ; orientador, Carlos Ernani Fries, 2022.
77 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Civil, Florianópolis, 2022.

Inclui referências.

1. Engenharia de Produção Civil. 2. Predição. 3. Imóveis. 4. Regressão linear múltipla. 5. Árvores de decisão. I. Fries, Carlos Ernani . II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Mateus Mota Gonçalves

Precificação de imóveis utilizando regressão linear múltipla e árvores de decisão

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título em Engenharia, área Civil, habilitação em Engenharia de Produção Civil e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Produção Civil

Florianópolis, 08 de dezembro de 2022.

_____

Prof. Mônica Maria Mendes Luna, Dra.
Coordenador do Curso

Banca Examinadora:

_____

Prof. Carlos Ernani Fries, Dr.
Orientador
Universidade Federal de Santa Catarina

_____

Prof. Diego de Castro Fettermann, Dr.
Avaliador
Universidade Federal de Santa Catarina

_____

Prof. Daniel Christian Henrique, Dr.
Avaliador
Universidade Federal de Santa Catarina

Este trabalho é dedicado à minha família, por sempre incentivarem meus estudos, à minha namorada, pelo companheirismo e apoio nessa jornada, e ao meu orientador Carlos, por toda dedicação e apoio no desenvolvimento dele.

AGRADECIMENTOS

Agradeço ao meu pai, Quintino Sebastião Gonçalves, por me incentivar, me ensinar lições da vida e ser meu exemplo. Agradeço à minha mãe, Gilséia Aparecida Mota Gonçalves, por todo o amor, apoio e ajuda que recebi. Agradeço aos meus irmãos, Maria Aparecida Gonçalves e Gabriel Mota Gonçalves, que sempre me apoiaram. Agradeço ao meu primo, Guilherme Gonçalves, que sempre me apoiou, me ensinou e trouxe perspectivas diferentes.

Agradeço à minha namorada, Bruna Gobbi, por sempre estar ao meu lado, me apoiando e me incentivando, e por ser minha companheira sempre acreditando e tirando o melhor de mim.

Agradeço ao meu orientador, Carlos Ernani Fries, por me guiar, participar ativamente de toda a pesquisa e por todo o aprendizado recebido.

Agradeço à A7 por todos os campeonatos, eventos, amigos e aprendizados durante o período da graduação.

Agradeço a todos que trabalharam comigo no estágio, principalmente ao Douglas e Ronal Balena por todo o conhecimento, aprendizado e responsabilidades transmitidas.

Agradeço à Universidade Federal de Santa Catarina por tudo o que vivi e aprendi nesse período de graduação, além de todas as oportunidades oferecidas. Por fim, agradeço a todos que de alguma forma contribuíram direta ou indiretamente para a minha pesquisa e para a minha formação. Muito obrigado!

“O valor de uma formação universitária não reside no aprendizado de muitos fatos, mas no treinamento da mente para conceber coisas novas.” (Albert Einstein, 1921)

RESUMO

A presente monografia pretende propor um modelo de precificação de imóveis em uma cidade do Centro-Oeste. Inicialmente, são apresentados periódicos que abordam sobre o tema deste trabalho para imersão de pesquisas relacionadas. Após, são trazidas as noções teóricas acerca dos temas relevantes para a compreensão do estudo, a partir de uma pesquisa bibliográfica. Nessa parte, são destacados os principais tópicos sobre regressão linear e árvores de decisão. A metodologia utilizada foi com coleta e tratamento de dados seguida de uma análise exploratória dos dados. A coleta de dados considerou dados disponíveis na *internet* levando em conta tanto imóveis novos quanto usados. Modelos de regressão linear foram utilizados no intuito de predizer a precificação de imóveis. Um modelo de árvore de decisão para classificação dos imóveis foi construído considerando variáveis contínuas e categóricas. Os resultados do trabalho mostram que a combinação de modelos de regressão com a técnica de árvores de decisão permite considerar simultaneamente variáveis contínuas e categóricas, obtendo-se desta forma, resultados bastante promissores com relação aos erros de predição na precificação de imóveis.

Palavras-chave: Predição, Regressão linear múltipla, Árvores de decisão, CHAID, imóveis.

ABSTRACT

This undergraduate thesis intends to propose a real estate pricing model in a city in the Midwest of Brazil. Initially, journals are presented, addressing the theme of the work for the immersion of related research. Afterwards, the theoretical notions about the relevant themes for the understanding of the study are brought, from bibliographical research. In this part, the main topics on linear regression and decision trees are highlighted. The methodology used relied on data collection and treatment followed by an exploratory data analysis. Data collection considered data available on the internet, considering both new and used properties. Linear regression models were used to predict property prices. A decision tree model for classifying properties was built considering continuous and categorical variables. The results of the work show that the combination of regression models with the decision tree technique allows the simultaneous consideration of continuous and categorical variables, thus obtaining very promising results regarding prediction errors in property pricing.

Keywords: Prediction. Multiple linear regression. Decision trees. CHAID. Real estate.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CART - *Classification And Regression Tree*

CHAID - Detecção Automática de Interação Qui-Quadrado

EQM - Erro Quadrático Médio

EPAM - Erro Percentual Absoluto Médio

ID3 - *Iterative Dichotomiser 3*

MPH - Metodologia de Preços Hedônicos

PRISMA - *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

R^2 - Coeficiente de determinação

$\bar{R}^2$ ou R^2 ajustado - Coeficiente de determinação ajustado

REQM - Raiz do Erro Quadrático Médio

SQT - Soma dos Quadrados Totais

SQE - Soma dos Quadrados dos erros

SQR – Soma dos Quadrados da Regressão

SUMÁRIO

1 INTRODUÇÃO

1.1 PROBLEMA

O imóvel é um ativo e um investimento e, devido ao seu alto custo, compreende a exigência de maior análise sobre o ativo adquirido. Isto porque muitas vezes esta é a maior compra na vida de uma pessoa. Lopes (2014) mostrou que os compradores de imóveis em São Paulo adquiriam em média 1,8 imóveis ao longo da vida.

A precificação de imóveis é um tema que conta com o interesse de todas as partes envolvidas no processo de compra e venda do imóvel. No Brasil, a precificação de imóveis urbanos é normatizada pela ABNT 14.653-2, onde consta a metodologia a ser seguida no processo de precificação de um imóvel.

A avaliação do imóvel depende de diversos fatores, dentre eles podem-se citar localização, idade, infraestrutura do condomínio, orientação solar, vista, varanda, garagem, andar, disposição dos ambientes, etc. Além destes, há o critério subjetivo, que não pode ser ignorado, tornando a precificação um processo complexo. Afinal, além de ser um mercado dinâmico, conta com itens que dependem de uma opinião para se tornar quantificado. Como esse processo necessita de um parecer humano, a situação pode gerar um desconforto no avaliador e, consequentemente, a algum conflito de interesse ou sentimentalismo no resultado da avaliação.

Por um lado, o proprietário do bem tem a questão sentimental, querendo valorizar o que é seu. Por outro, os avaliadores podem ser representados por um corretor de imóveis, que pode controlar o preço com a intenção de direcionar o processo de venda ao seu melhor interesse, visto que as escolhas ao se analisar investimentos são complexas (DOROW, 2009). Ainda, o corretor pode estar representando algum banco, avaliando a possibilidade do financiamento, carregando consigo os pensamentos de investimento, como a liquidez do ativo, os riscos envolvidos na transação, inclinando-se a precificar abaixo do valor de mercado.

O preço do imóvel é um dos critérios que o banco utiliza para fornecer o financiamento imobiliário, empregado na grande maioria das compras de imóveis do país. Ao se considerar que o máximo da renda mensal a ser comprometido pelo financiamento é de 30%, a aquisição de um imóvel é um investimento de longo prazo, podendo chegar a 35 anos para quitar o saldo.

Sendo assim, é de extrema importância que se realize a melhor escolha dentro das condições individuais de cada família.

Segundo Ecconit (2020), a demanda habitacional do Brasil até 2030 será de aproximadamente 11,9 milhões de habitações. Esse é um ponto que interfere de forma extensa no setor de construção, na oferta e demanda de imóveis, podendo gerar o aquecimento do mercado e uma sobrevalorização dos produtos.

Como se trata de um mercado cheio de nuances e todos precisam de moradia, torna-se um tema com muitos interessados, sendo ponto importante da economia e política do país. Com o fim de auxiliar o setor existem alguns canais de informações e dados disponíveis sobre o mercado de construção e serviços imobiliários. E, com base nesses dados, como criar um modelo de precificação de imóveis que torne o processo de precificação menos subjetivo?

1.2 OBJETIVOS

Nesta seção serão abordados o objetivo geral e os objetivos específicos do trabalho.

1.2.1 Objetivo Geral

O objetivo deste trabalho consiste em propor um modelo de precificação de imóveis em uma cidade do Centro-Oeste utilizando dados de imóveis disponíveis *online* e disponibilizados por uma empresa do ramo.

.

1.2.2 Objetivos Específicos

- Determinar variáveis que são importantes na precificação de imóveis;
- Combinar variáveis contínuas e categóricas com árvores de decisão.

1.3 JUSTIFICATIVA

A avaliação imobiliária é, em grande parte, utilizada em operações de arrendamento, venda, operações de garantia, seguro, decisões judiciais, autos de infração, indenizações, expropriações, servidões, permutas, entre outras.

Diante disso, torna-se imperativo que sejam realizados mais estudos com o objetivo de aprimorar o sistema vigente de precificação imobiliária. Assim, entre os diferentes métodos de avaliação imobiliária propostos pela NBR 14.653-2 (2010), o método mais utilizado e recomendado é o método comparativo de dados do mercado, levando em consideração as diferentes características e atributos que influem na composição do preço dos imóveis avaliados, cujas características e atributos são analisadas por inferência estatística. A regressão linear é geralmente a análise mais utilizada para a inferência estatística (NBR 14.653-2, 2010).

Este trabalho pretende utilizar de regressão linear múltipla e árvore de decisão como métodos estatísticos para modelar a predição de precificação de imóveis.

1.4 LIMITAÇÕES

O trabalho apresenta limitações com relação à regionalidade da amostra, permitindo considerar os resultados apenas para a região de estudo e para o período em que foram coletados os dados da amostra. Outra limitação diz respeito à dificuldade de conseguir dados que poderiam ser relevantes ao modelo, tais como orientação solar do imóvel, andar do imóvel, tipo de garagem, ventilação dos banheiros e outras facilidades que o imóvel possui. Isso se deve à indisponibilidade por parte dos anúncios e pela não computação na base de dados que a empresa do ramo forneceu.

O preço do imóvel é outra limitação do estudo, apesar da NBR 14.653-2 (2010) considerar o método comparativo de dados do mercado. Visto que os preços praticados em anúncios nos portais e pelas construtoras em imóveis novos são relacionados à oferta e não a valor final de aquisição do imóvel.

1.5 ESTRUTURA DO TRABALHO

O trabalho é estruturado em quatro capítulos. Primeiramente, a introdução contextualiza e descreve o tema estudado e a importância do tema. Nela, também são listados os objetivos propostos, justificativa e limitações do trabalho.

No capítulo 2 é apresentada a fundamentação teórica da avaliação imobiliária, apresentando-se estudos que abordam o tema de precificação de imóveis e os métodos de regressão linear e árvore de decisão.

Na sequência, no capítulo 3 são abordadas as metodologias utilizadas para o desenvolvimento do projeto de forma linear. Também são detalhadas as análises referentes à construção do modelo preditor de precificação de imóveis para o conjunto de dados de agosto de 2022. Para as análises são apresentados tabelas e gráficos para melhor entendimento das inferências estatísticas, onde foram definidas as variáveis importantes para o modelo preditor de precificação de imóveis.

Por fim, no último capítulo são apresentadas as conclusões finais do que foi executado no processo de modelar o preditor de precificação de imóveis. Também são comparados os resultados obtidos com os objetivos propostos para mostrar se foi atingido o objetivo inicial do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Com base nos periódicos encontrados, foi possível determinar os métodos, conceitos e áreas de estudo importantes para desenvolver este trabalho. Por isso, nessa seção será apresentada a fundamentação teórica sobre a avaliação imobiliária e os principais métodos utilizados para precificação de imóveis. Com a finalidade de contextualizar os tópicos e suas teorias em meio à literatura.

2.1 ESTUDOS PERTINENTES AO TEMA

Neste tópico, portanto, será abordada uma perspectiva sobre estudos realizados que trataram de abordar o tema de avaliação de imóveis, trazendo, assim, um contexto do assunto no cenário científico.

Sardá (2018) utilizou a regressão linear clássica associada a análise multivariada para dispor das variáveis que influenciam o valor de imóveis em Florianópolis - SC. O trabalho teve como objeto de estudo avaliação de apartamentos. Nesta pesquisa foram avaliadas oito variáveis, sendo (1) área total, (2) quantidade de quartos, (3) suítes, (4) vagas de garagem, (5) idade do apartamento, (6) padrão, (7) distância à Beira-Mar e (8) andar. A análise multivariada possibilitou o estudo das variáveis e a definição das classes homogêneas das amostras para, após, serem calculados modelos de regressão em cada classe e comparados a um modelo único. Nesse estudo foi concluído que a análise multivariada contribuiu para a melhora em determinadas características do modelo de regressão, como a utilização de todos os dados da amostra, significativa diminuição do coeficiente de variação, aumento de alguns coeficientes de determinação, diminuição dos valores absolutos dos resíduos e além desta melhora os agrupamentos possibilitaram a melhor compreensão do mercado imobiliário abordado.

Em um estudo realizado na Grécia, Doumpos (2020) alega que modelos automatizados de avaliação são amplamente utilizados para estimar preço de imóveis e que são usualmente desenvolvidos por abordagens de regressão. Nesse trabalho apresentou-se uma comparação sobre o desempenho de técnicas de regressão paramétrica e não-paramétrica ao desenvolver modelos confiáveis de avaliação automatizada de imóveis residenciais. Para isso, utilizou-se amostras de propriedades gregas no período de 2012 a 2016. O autor concluiu que modelos de regressão linear que recorrem a esquema espacial ponderado fornecem melhores resultados, superando abordagens de machine learning e modelos que não consideram os efeitos locais.

Uma análise feita no Japão, Ishijima e Maeda (2013) propuseram uma teoria de precificação de imóveis aplicando análise empírica, em que demonstraram que, sob condições definidas, o preço de equilíbrio teórico pode ser descrito como a combinação linear de atributos comuns a todos os imóveis. No entanto, em circunstâncias mais realistas, os preços podem divergir do preço de equilíbrio. Sendo assim, aplicaram um modelo misto do modelo hedônico com o Box-Cox, a qual é uma técnica de transformação de dados útil usada para estabilizar a variância, tornar os dados mais semelhantes à distribuição normal, melhorar a validade das medidas de associação e para outros procedimentos de estabilização de dados. Ao fim do estudo, aplicaram o modelo para analisar os dados de registros de imóveis do Japão, este apresentou resultados precisos, sendo disponibilizada uma ferramenta online de avaliação imobiliária.

Na Coreia do Sul, Choi (2021) critica o Método dos Preços Hedônicos (MPH) por não considerar a flutuação do preço do imóvel avaliado e a volatilidade do mercado temporalmente. Sendo assim, propôs o método da pseudo auto comparação que visa encontrar um pseudo imóvel que mais se assemelhe às características do imóvel avaliado. Com ajustes no seu preço de transação anterior faz com que esteja em sincronia com as alterações do mercado imobiliário. Portanto, foi proposto e testado um modelo em dois cenários de volatilidade, nas cidades de Seul e Gyeonggi na Coreia do Sul, utilizando dados de transações imobiliárias. Como resultado do estudo, o método proposto apresentou erros de estimativa aproximadamente cinco vezes menor ao prever os preços de transação em comparação ao método hedônico, provando-se uma boa ferramenta para avaliação em massa de apartamentos.

Com base nos estudos apresentados, conclui-se que a precificação de imóveis é uma questão recorrente no cenário mundial, principalmente pelos métodos serem, por vezes, subjetivos, trazendo um viés errôneo nos resultados avaliados. Além disso, identifica-se que o avanço da tecnologia e crescente acesso à informação, colaboram para a elaboração de novos modelos para realizar a precificação de imóveis, os quais buscam remover da equação a subjetividade.

2.2 AVALIAÇÃO IMOBILIÁRIA

Conforme explicado anteriormente, a avaliação imobiliária é elemento de extrema importância para a precificação dos imóveis, e tem sua regulamentação dada pela Associação Brasileira de Normas Técnicas.

Nesse sentido, de acordo com a NBR 14653-1 (2019), que estabelece os procedimentos gerais para avaliação de bens: “Valor de mercado é a quantia mais provável pela qual se negociaria voluntariamente e conscientemente um bem, em uma data de referência, dentro das condições do mercado vigente”. Ou seja, o valor de mercado é uma expectativa do valor a ser negociado pelo bem. De outro lado, está definição do preço, que é o valor efetivo que foi pago pelo bem.

Sendo assim, avaliação de bens é definida como uma análise técnica utilizada para identificar o valor, custo e viabilidade econômica de um determinado propósito, circunstância e época (NBR 14653-1/2019).

Em vista disso, no Brasil, a NBR 14653-2 (2010) normatiza o processo de avaliação de bens para imóveis urbanos, descrevendo algumas metodologias de avaliação já reconhecidas, bem como formas de aplicar métodos não detalhados na norma.

Posto isso, um método eficaz para predição de valores pode permitir que as transações dos imóveis sejam justas para o comprador e o vendedor, além de reduzir a possibilidade de manipulação pelos corretores de imóveis. Sendo o mercado de imóveis um setor reflexo da economia do país, a área de previsão de valores dos imóveis sempre esteve presente nos estudos (VERAS, 2019).

2.3 CORRELAÇÃO E REGRESSÃO LINEAR

A análise de correlação linear busca responder se duas variáveis têm algum tipo de associação entre si. De certa forma, estima o grau de relacionamento linear entre as variáveis para inferir se um valor baixo ou alto de uma das variáveis influencia em um valor baixo, ou alto da outra variável (HOFFMANN, 2016).

A análise de regressão linear, por outro lado, preocupa-se com a dependência estatística entre as variáveis analisadas (CHEIN, 2019). O modelo de regressão linear consiste em analisar e compreender a relação de uma variável dependente com uma ou mais variáveis independentes, supondo que esta dependência é linear (FARIAS et al, 2020).

Segundo a NBR 14653-2 (2010), o método mais utilizado quando a intenção é estudar como se comporta uma variável dependente em relação a outras variáveis responsáveis pela variabilidade percebida nos preços é a análise de regressão.

2.3.1 Regressão Linear Simples

De acordo com Pereira e Pereira (2016), o modelo de regressão linear simples é definido como a relação linear entre a variável dependente (Y), uma variável independente (X) e um erro. Como o erro não pode ser explicado com a relação linear entre X e Y, supõe-se na regressão linear simples que o erro é zero. Dessa forma, obtêm-se a equação da reta estimada:

$$\hat{Y} = b_0 + b_1 X \tag{1}$$

Se há "n" pares de dados $(x_1, y_1), \ldots, (x_n, y_n)$, pode-se estimar os valores de b_0 e b_1 utilizando do método dos mínimos quadrados, como mostrado nas Equações 2 e 3 (PEREIRA e PEREIRA, 2016):

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{2}$$

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} \tag{3}$$

onde, $\bar{X}$ é a média dos valores de X_i, em outros termos: $\bar{X} = \frac{\sum X_i}{n}$. A mesma analogia aplica-se a $\bar{Y}$.

A significância estatística do modelo obtido deve ser verificada por meio de testes de hipóteses. Para a validação de hipóteses utiliza-se das somas dos quadrados totais (SQT - Equação 4), dos erros (SQE - Equação 5), da regressão (SQR - Equação 6) e do coeficiente de determinação (R^2 – Equação 7) (PEREIRA e PEREIRA, 2016).

$$SQT = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \tag{4}$$

$$SQE = \sum (Y_i - \hat{Y}_i)^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum X_i Y_i \tag{5}$$

$$: SQR = SQT - SQE = \sum (\hat{Y}_i - \bar{Y})^2 \tag{6}$$

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \tag{7}$$

O coeficiente de determinação (R²) equivale à proporção da variância dos valores de Y que pode ser atribuída a regressão com a variável X (BRAULE, 2001).

Utilizando do teste F com a Equação 8, obtêm-se o valor de F que deve ser comparado com o valor F da tabela de Tabela F-Snedocor (Anexo A). Se $H_0: \beta = 0$ for a hipótese nula e se o valor de F calculado for maior que o valor de F da tabela, rejeita-se H_0 (PEREIRA e PEREIRA, 2016).

$$F = \frac{SQR}{1} \times \frac{n-2}{SQE} \tag{8}$$

2.3.2 Regressão Linear Múltipla

O Modelo de Regressão Linear Múltipla é definido como a relação linear entre a variável dependente (Y) e várias variáveis independentes ($X_1, \dots, X_k$). Sendo k o número de variáveis independentes e "n" o tamanho da amostra. Pode-se simplificar através de uma abordagem matricial da seguinte forma (MAIA, 2017):

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{9}$$

Como a regressão linear múltipla envolve várias variáveis e está simplificada na forma matricial, as equações de β (Eq. 10), da soma dos quadrados totais (Eq. 11), dos erros (Eq. 12) e de regressão (Eq. 13) sofrem ajustes para a forma matricial. A equação do coeficiente de determinação (R²) mantém-se especificado como na Equação 7 (MAIA, 2017).

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{10}$$

$$SQT = \sum (Y_i - \bar{Y})^2 = Y^T Y - n\bar{Y}^2 \tag{11}$$

$$SQE = \sum (\hat{Y}_i - Y_i)^2 = Y^T Y - \hat{\beta}^T X^T Y \tag{12}$$

$$SQR = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^T X^T Y - n\bar{Y}^2 \tag{13}$$

Na regressão linear múltipla também é utilizada a tabela Tabela F-Snedocor (Anexo A) para comparar com o valor F calculado da Equação 14. Se $H_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0$ for a hipótese nula e se o valor F calculado for maior que o valor F da tabela, rejeita-se H_0 (PEREIRA e PEREIRA, 2016).

$$F = \frac{SQR}{k} \times \frac{n-k-1}{SQE} \tag{14}$$

No método de regressão linear múltipla é interessante destacar que o R^2 nunca diminui quando uma nova variável independente é adicionada ao modelo, ainda que essa nova variável não seja significativa para ele. Com isso tem-se o coeficiente de determinação ajustado, ou $\bar{R}^2$ (Eq. 15), que penaliza o uso excessivo de variáveis independentes sem importância, tendo um valor menor do que o R^2. (BERTI et al, 2019)

$$\bar{R}^2 = 1 - \frac{\frac{SQE}{(n-k-1)}}{\frac{SQT}{(n-1)}} \tag{15}$$

2.3.3 Método de Precificação Hedônico

As referências teóricas citam Kelvin Lancaster (1966) e Sherwin Rose (1974) como os autores dos trabalhos mais importantes na consolidação da teoria dos preços hedônicos. Lancaster (1966) apresentou que a utilidade não é atingida somente pelo bem, mas também por suas características.

Segundo Rosen (1974), preços hedônicos são definidos como preços implícitos de atributos e resultam dos preços notados em produtos diferenciados e adicionados às suas respectivas propriedades. Em seu trabalho, abordou o comportamento do consumidor e o equilíbrio do mercado, trazendo a questões importantes, como os consumidores valorizarem as utilidades dos produtos e a tendência dos produtores a buscar satisfazer a procura dos consumidores com o menor custo possível.

Na formulação de Rosen (1974), o MPH determina o equilíbrio em um plano onde o vetor de coordenadas $z = (z_1, z_2, \cdots, z_n)$ corresponde a cada ponto do plano. Rosen (1974) definiu que cada preço $p(z) = p(z_1, z_2, \cdots, z_n)$ é determinado em cada ponto do plano, sendo cada "z" uma característica do produto. Malpezzi (2002), em seu trabalho, apresentou a função

p(z) considerando características estruturais do imóvel, características da vizinhança, localização no mercado estudado, condições contratuais e o momento que ocorre a venda como as variáveis do modelo.

Owusu-Ansah (2011) realizou uma revisão do MPH em habitação, na qual explica que a curva de regressão hedônica, de forma simplificada, demonstra a relação entre a variável dependente (preço do imóvel no caso deste trabalho) e a variável independente ou exploratória (uma característica do imóvel como tamanho). Transformando em uma equação, seria na seguinte forma:

$$y_i = f(x_i) + \varepsilon_i \tag{16}$$

onde, "y" representa o preço do imóvel a ser estimado, $f(x_i)$ é a média do valor de "y" dado o número de quartos e ε_i é 0 assumindo que é uma distribuição normal. Sendo a Equação 16 análoga a Equação 1.

Härdle (1990 apud OWUSU-ANSAH, 2011) define o modelo paramétrico de precificação hedônica como um tipo do modelo hedônico no qual supõe que a curva de regressão tem uma forma pré-definida onde é totalmente descrita por um conjunto finito de parâmetros. Os parâmetros geralmente são os coeficientes das variáveis independentes, traduzindo esta definição para uma forma matemática seria similar a Equação 9 apresentada anteriormente.

Owusu-Ansah (2011) apresenta modelos paramétricos, não paramétricos e semi-paramétricos, concluindo com os benefícios e malefícios de cada tipo de modelo. Os modelos paramétricos são os modelos mais utilizados na análise de regressão hedônica, devido ao fato de serem totalmente determinados por parâmetros, simplicidade na estimação e interpretação dos coeficientes. Se as suposições feitas para os modelos paramétricos estiverem corretas, as estimativas serão precisas e os modelos ajustados podem ser interpretados facilmente.

Por outro lado, a desvantagem é que as suposições podem tornar o modelo muito restrito para se ajustar a variáveis inesperadas. Outra crítica ao modelo é que a premissa da linearidade das variáveis pode tornar o modelo inconsistente e fornecer uma regressão enganosa se alguma das suposições não forem seguidas.

No modelo não paramétrico Owusu-Ansah (2011), conclui como positivo o fato deste modelo ter uma abordagem flexível que funciona mesmo com conjuntos pequenos de dados e

até com valores ausentes. Outro fator é que o modelo não paramétrico fornece uma estimativa objetiva da curva de regressão, tornando-a adaptável de explorar as relações entre as variáveis. Como críticas ao modelo, o principal ponto está na complexidade da sua natureza técnica, visto que não há parâmetros para descrever torna complexo a dimensionalidade da equação. Referida complexidade faz com que seja difícil atribuir pesos às variáveis, comparar duas populações e estimar os valores.

No modelo semi-paramétrico Owusu-Ansah (2011), complementa que, como é a combinação dos dois modelos acima, apresenta os prós e contras deles. Com essa combinação é possível realizar estimativas precisas e robustas para o modelo de regressão. Neste modelo, o problema da dimensionalidade pode ser reduzido, pois a maioria das variáveis entra na parte paramétrica. Contudo, como no modelo paramétrico, se as premissas não se realizam, as estimativas podem ser inconsistentes.

Desta forma, os preços implícitos da econometria são estimados com base em análise de regressão linear múltipla, onde o preço do produto é regredido de acordo com suas características (ANGELO; FAVERO, 2003). É importante destacar que, em linhas gerais, o MPH pressupõe que o mercado estudado esteja em equilíbrio. No MPH o principal objetivo é encontrar o valor que cada característica agrega ou não no produto dado a sua significância para o modelo.

2.3.4 Erro Quadrático Médio e Raiz do Erro Quadrático Médio

O Erro Quadrático Médio (EQM) considera a média do quadrado dos erros das previsões. Em outras palavras, utiliza-se a diferença entre o valor predito pelo modelo e o valor real da amostra, eleva-se essa diferença ao quadrado e, assim, obtém-se o erro de um valor da amostra. Repete-se esse passo para cada valor da amostra a ser testado, soma-se os resultados e divide-se pelo número de elementos preditos (AZANK, 2020).

O EQM apresenta para a métrica o valor mínimo como zero, não possui valor máximo. Para a avaliação da métrica do EQM quanto maior o valor da soma, pior é o modelo. E pode ser descrito pela Equação 17 (AZANK, 2020).

$$EQM = \frac{1}{n} \quad \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{17}$$

Como o EQM eleva o erro ao quadrado, o valor predito que está mais distante do valor real aumenta o valor da soma de forma acelerada, tornando, assim, o EQM uma métrica de avaliação útil para problemas em que grandes erros não são tolerados. O EQM é análogo ao desvio padrão (NAKAMURA, 2022).

Um problema do EQM é a sua interpretação, visto que a predição, por exemplo, é na unidade R$, e a unidade do EQM seria R$², que não tem significado físico. Para solucionar o problema de diferença das unidades físicas, a Raiz do Erro Quadrático Médio (REQM) reduz os valores do EQM para a mesma unidade física dos valores a serem preditos. A Equação 18 mostra o cálculo da REQM (AZANK, 2022).

$$REQM = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} = \sqrt{EQM} \tag{18}$$

2.3.5 Erro Percentual Absoluto Médio

O Erro Percentual Absoluto Médio (EPAM), dado pela Equação 19, é uma medida resultante em uma porcentagem, obtida através da divisão da diferença entre o valor predito e o valor real da amostra (LOPES, 2002).

$$EPAM = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{19}$$

Por se tratar de uma porcentagem, o EPAM é uma métrica intuitiva tanto para o analista de previsões quanto para a comunicação dos resultados para os contratantes. Assim como o EQM e o REQM, quanto menor o valor obtido, mais preciso é o modelo de regressão (AZANK, 2022).

2.4 ÁRVORE DE DECISÃO

Na análise de decisão, uma árvore pode ser utilizada para representar visual e explicitamente as decisões. Nesta abordagem, é utilizado um modelo de decisão que é representado em forma de árvore. Para Kumari (2021), dentre outros modelos de aprendizagem supervisionado, as árvores são modelos preditivos com maior precisão e de simples compreensão. Alguns dos algoritmos mais utilizados são, *Classification And Regression Tree*

(CART), *Iterative Dichotomiser 3* (ID3), C4.5 e *Chi-square Automatic Interaction Detector* (Detecção Automática de Interação Qui-Quadrado) (CHAID) (KUMARI, 2021).

Uma árvore de decisão é desenhada de cabeça para baixo, isto é, com a sua raiz no topo, onde se encontram os dados do problema. A cada etapa do processo ele se divide em um nó interno que representa uma condição da variável, sendo a árvore dividida em ramos. O final do ramo que não mais se divide é a decisão ou folha.

Entre as diversas técnicas e algoritmos que são utilizados no processo de descoberta de conhecimento em bases de dados, foi selecionado o método CHAID, visto que aceita variáveis categóricas nominais ou ordinais como variáveis dependentes e um nó pode ter múltiplas divisões. Quando os preditores são contínuos, são transformados em preditores ordinais. Kass (1980) foi quem desenvolveu a técnica. Este autor afirma que se trata de um método estatístico bastante eficiente para classificar dados.

Segundo Kumari (2021), qui-quadrado é uma métrica estatística utilizada para encontrar a diferença entre os nós filho e pai. Calcula-se o qui-quadrado encontrando-se a diferença entre as contagens observadas e esperadas da variável de destino, para cada nó e a soma quadrada padronizada dessas diferenças, resultando no valor do qui-quadrado, que é demonstrado na Equação 20, sendo y o valor observado e y’ o valor esperado.

$$Qui - Quadrado(X^2) = \sqrt{\frac{(y - y')^2}{y'}} \tag{20}$$

O método CHAID é embasado nos testes de associação qui-quadrado e divide o conjunto de dados em subconjuntos, os quais não podem ocorrer simultaneamente, que melhor descrevem a variável resposta de forma exaustiva (TURE et al., 2009). Kumari (2021) define que o CHAID utiliza da métrica do qui-quadrado para descobrir o recurso (ou variável) mais importante e aplica-o sucessivamente até que os subconjuntos tenham uma única decisão.

Para exemplificar o método, na sequência será exemplificado um passo do método. Os demais passos da classificação desse exemplo estão incluídos no Anexo B. O exemplo apresentado no trabalho de Kumari (2021) é apresentado na Figura 1.

Figura 1 - Base de dados Kumari (2021) para árvore CHAID.

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Fonte: Kumari (2021).

Os cálculos do qui-quadrado para a variável *Outlook* são apresentados na Figura 2. Para o cálculo do qui-quadrado, soma-se os valores para “sim” e “não”. Obtém-se assim: 0,316 + 0,316 + 1,414 + 1,414 + 0,316 + 0,316 = 4,092. Este é o valor do qui-quadrado para a variável *Outlook*.

Figura 2 - Cálculo do qui-quadrado para a variável *Outlook*

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Sunny	2	3	5	2.5	0.316	0.316
Overcast	4	0	4	2	1.414	1.414
Rain	3	2	5	2.5	0.316	0.316

Fonte: Kumari (2021).

Os resultados dos cálculos do valor de qui-quadrado das demais variáveis do exemplo são apresentados na Figura 3. O maior valor do qui-quadrado (4.092) foi obtido para a variável *Outlook.*

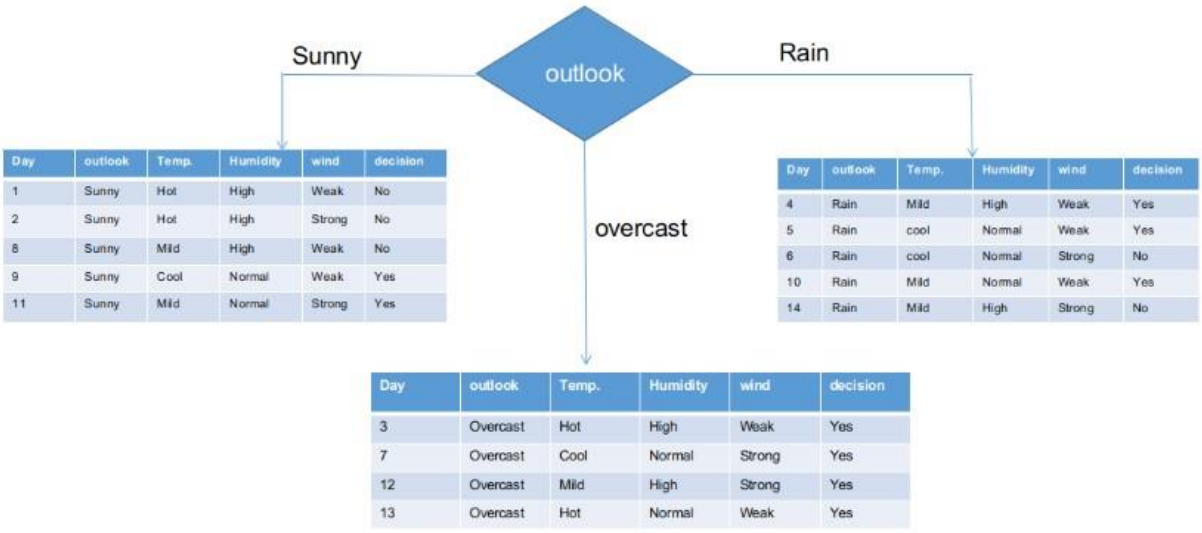
Figura 3 - Qui-quadrado das variáveis das amostras

Feature	Chi-square value
Outlook	4.092
Temperature	2.569
Humidity	3.207
Wind	1.604

Fonte: Kumari (2021).

Para a variável *Outlook,* três diferentes valores estão contidos na amostra: *Sunny, Rain* e *Overcast.* A amostra original é então dividida considerando estes três atributos, conforme apresentado na Figura 4. Esta é a primeira divisão de nós da árvore. Os passos intermediários para as demais variáveis são apresentados no Anexo B. O resultado final da classificação do exemplo de Kumari (2021) é apresentado na Figura 5.

Figura 4 - Primeira divisão da Árvore CHAID



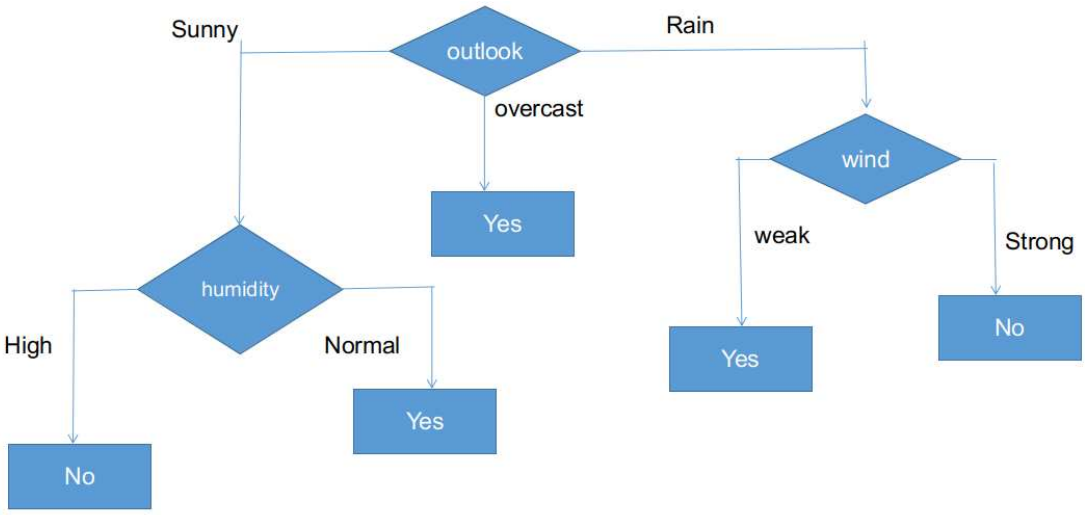
Fonte: Kumari (2021).

O resultado final conduz a cinco conclusões referentes a sair de casa para comprar leite:

i. **SE** *Outlook* **FOR** *Sunny* **E** *Humidity* **FOR** *High* **ENTÃO** Não (não sair de casa).
ii. **SE** *Outlook* **FOR** *Sunny* **E** *Humidity* **FOR** *Normal* **ENTÃO** Sim (sair de casa).
iii. **SE** *Outlook* **FOR** *Overcast* **ENTÃO** Sim (sair de casa).
iv. **SE** *Outlook* **FOR** *Rain* **E** *Wind* **FOR** *Weak* **ENTÃO** Sim (sair de casa).

v. **SE** *Outlook* **FOR** *Rain* **E** *Wind* **FOR** *Strong* **ENTÃO** Não (não sair de casa).

Figura 5 - Resultado final da classificação do exemplo de Kumari (2021)



Fonte: Kumari (2021).

3 METODOLOGIA E ANÁLISES

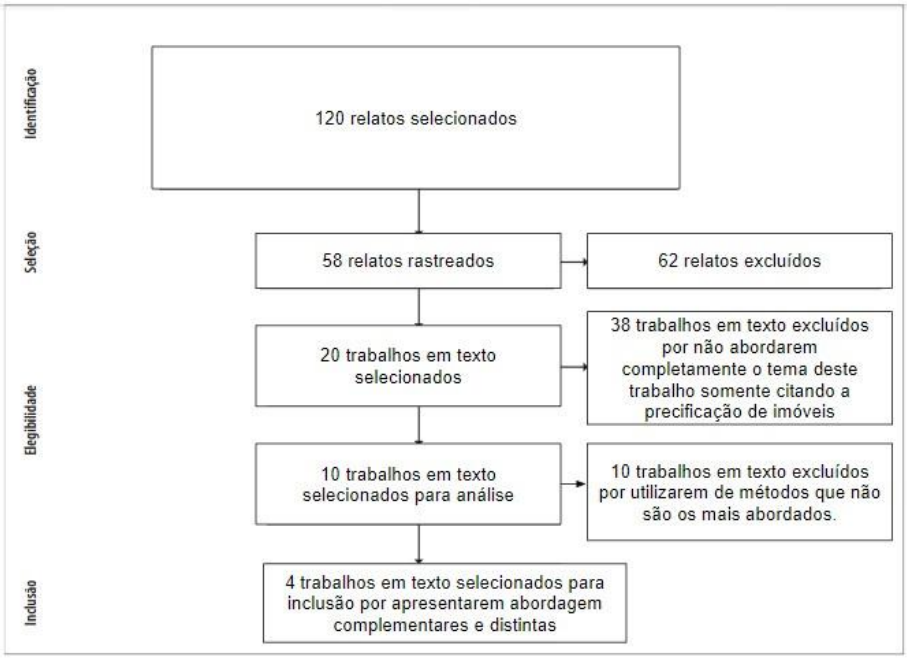
Nesse capítulo, são apresentados os métodos empregados para a proposição do modelo de precificação de imóveis usados em uma cidade do Centro-Oeste. Para alcançar os objetivos propostos neste trabalho, foi definido o roteiro metodológico com etapas descritas a seguir.

3.1 LEVANTAMENTO BIBLIOMÉTRICO

As pesquisas de periódicos retornaram um resultado muito amplo de amostras. Para reduzir essa amostragem utilizou-se dos princípios do método *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) de revisão sistemática, reduzindo-se o número de trabalhos a serem apresentados para quatro fontes principais.

Como os resultados obtidos apresentaram uma amostra ampla de periódicos realizou-se uma leitura dinâmica do resumo e com base no critério de estar abordando diretamente o tema deste trabalho. O total de trabalhos selecionados nesta etapa foi de 58 periódicos que abordavam o tema de precificação de imóveis de 120 resumos lidos. O método PRISMA segue fluxograma de indicação das etapas até definir os artigos incluídos no trabalho. Este fluxograma foi utilizado para identificar a quantidade de artigos nas etapas e o motivo da exclusão de artigos. O fluxograma é apresentado na Figura 6.

Figura 6 - Fluxograma de etapas para seleção dos periódicos incluídos no trabalho



Fonte: Autoria própria.

Os quatro trabalhos selecionados foram de Sardá (2018), Doumpos (2020), Ishijima e Maeda (2013) e Choi (2021). Estes trabalhos foram apresentados na seção 2.1 deste trabalho.

A verificação de conteúdo concentrou-se no Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), visto que integra o resultado a diversas bases de dados. Além do portal CAPES, foi também pesquisado em sites de busca na internet para encontrar fontes complementares.

A pesquisa no portal da CAPES considerou todas as bases de dados disponíveis para acesso e filtrado para ter resultado apenas os periódicos revisados por pares. As palavras-chaves utilizadas foram os termos “Real Estate”, “Residencial Real Estate”, “Pricing Models”, “Prediction Model” e “Residential Real Estate” + “Prediction Model”.

O termo “Real Estate” retornou 686.559 periódicos revisados por pares, sendo história, habitação e economia os principais assuntos mencionados. Identificou-se que os temas Home Prices, Regression Analysis e Valuation são exibidos em 8.880 resultados. Em função deste termo ter retornado muitos resultados, foi utilizado o termo “Residencial Real Estate” como palavra-chave. Esse retornou 144.245 resultados, dos quais, os principais campos de estudos são habitação e planejamento urbano.

Testou-se ainda “Pricing Models” e “Prediction Model” com o intuito de encontrar o melhor termo para precificação, no qual “Pricing Models” resultou em 579.123 periódicos. O termo “Prediction Model” foi utilizado em uma segunda busca, retornando 7.412.047 periódicos, em ambas as palavras-chaves, os principais tópicos abordados são sobre os modelos e métodos utilizados.

Os termos “Residential Real Estate” e “Prediction Model” foram pesquisados de forma conjunta para refinar os resultados, sendo 18.778 o número de publicações pertinentes a estudos de imóveis residenciais e modelos de previsão com “Home Prices” sendo o segundo assunto mais abordado nos periódicos. Foi percebido que temas de métodos e modelos para previsão, tendo a “técnica de análise de regressão” como aquela com o maior número de aplicações com 333 amostras, seguida por Machine Learning com 276 resultados. Em seguida selecionando somente Machine Learning constatou-se que os primeiros estudos surgiram em 2008. Restringindo o período da busca para 2000, 2008, 2015 e 2020, detectou-se que os estudos envolvendo Machine Learning vêm recebendo maior atenção de pesquisadores.

Durante o processo de investigação de estudos realizados sobre o tema proposto neste trabalho inferiu-se que o setor imobiliário é de interesse nacional e internacional, visto que

foram encontrados trabalhos com referências em mercados locais de norte a sul do país e ao redor do mundo.

A Metodologia de Preços Hedônicos (MPH) não foi encontrada na lista dos assuntos com mais periódicos realizados com base nas palavras-chave, mas foi citada nos trabalhos encontrados. Esta metodologia considera os preços implícitos dos atributos do produto.

3.2 COLETA E TRATAMENTO DE DADOS

Os dados utilizados nesse trabalho foram fornecidos por uma empresa do setor imobiliário que tem como área de atuação uma cidade da região Centro-Oeste e divididas em duas bases. A primeira base de dados é referente ao preço dos imóveis novos e na planta das construtoras, não considerando os imóveis do Programa Casa Verde e Amarela. Os dados são atualizados mensalmente contendo informações desde maio de 2020 até setembro de 2022, contando atualmente com 215 empreendimentos, 755 tipologias de imóveis e 7.785 unidades disponíveis.

A segunda base de dados é referente a anúncios de imóveis em portais online que foram tratados e alocados em empreendimentos e tipologias de imóveis com base nas informações dos anúncios e no cadastro imobiliário. A base original tem disponível informação de 12.500 anúncios de imóveis já entregues na região estudada. Como os anúncios foram tratados por empreendimento e tipologia de imóvel, sendo obrigatório identificar a idade do condomínio, conta-se com uma redução nessa base em que foram identificados 5.600 anúncios em que se teve a confirmação da idade do empreendimento, agrupados em 626 tipologias de imóveis. Essas bases de dados são divididas em série histórica e não histórica. Portanto foram selecionados os dados referentes ao mês de agosto de 2022, 2021 e 2020, os quais serão analisados neste trabalho. Como os dados são de bases diferentes verificou-se que não havia dados duplicados entre os portais que representariam o mesmo imóvel. A Tabela 1 mostra as variáveis da base de dados.

Tabela 1 - Variáveis iniciais dos dados

Nome da variável	Descrição da variável
ID Empreendimento	Número identificador para o empreendimento
Nome	Nome do empreendimento
Construtora	Construtora do empreendimento
Bairro	Bairro em que está localizado
Data de Entrega	Quando o empreendimento foi entregue aos moradores
ID Tipo de Unidade	Número identificador da tipologia do imóvel
Área	Tamanho da tipologia do imóvel em m²
Quartos	Número de dormitórios no imóvel
Suítes	Número de dormitórios que tem banheiro exclusivo
Garagem	Número de vagas de garagem da tipologia do imóvel
Preço	Preço do anúncio ou tabela da construtora para o imóvel
R$/m²	Preço por m² do imóvel
Disponibilidade	Quantos imóveis estavam disponíveis daquela tipologia
Data do preço	A data em que aquele preço foi anunciado
Facilidades do empreendimento	Itens de lazer que são da área do condomínio
Facilidades da unidade tipo	Itens que são da tipologia do imóvel

Fonte: autoria própria.

Segundo estudo da DataZap+ (2022), os itens convencionais de condomínios que são de preferência dos consumidores são: churrasqueira, piscina, academia e salão de festas. Com base nesses itens, foi criada uma variável categórica binária para informar esses dados. Utilizando da mesma lógica foi criado duas variáveis categóricas para dizer se a tipologia do imóvel possuía, ou não, varanda gourmet e varanda com piscina.

Com as variáveis de “Data de Entrega”, criou-se uma variável de diferença de datas em relação à data atual. Desta forma, datas passadas ficaram com valor negativo e datas futuras com valor positivo.

A variável do “Bairro” foi utilizada para criar uma variável categórica onde os principais bairros da cidade ficaram com valor numérico de cinco e conforme a procura e a distância desses bairros o valor diminuía, podendo chegar até um, como a base exclui imóveis do estilo minha casa minha vida não se tem bairro com índice um. De outra forma de maior interesse para menor interesse. A Tabela 2 mostra as variáveis que serão consideradas para o estudo na próxima etapa.

Tabela 2 - Variáveis a serem consideradas

Variável	Descrição
Índice do bairro	Categórica de 1 a 5
Referência da data de entrega	Contínua com valor em meses de diferença da data atual
Área	Contínua em m²
Quartos	Contínua com valor inteiro
Suítes	Contínua com valor inteiro
Garagem	Contínua com valor inteiro
Preço	Contínua com valor real
R$/m²	Contínua com valor real
Varanda gourmet	Categórica binária, em relação à tipologia do imóvel
Varanda com piscina	Categórica binária, em relação à tipologia do imóvel
Salão de festas	Categórica binária, em relação à área de lazer do condomínio
Área de churrasqueiras	Categórica binária, em relação à área de lazer do condomínio
Espaço fitness / academia	Categórica binária, em relação à área de lazer do condomínio
Piscina	Categórica binária, em relação à área de lazer do condomínio

Fonte: autoria própria.

Definidas as variáveis que serão consideradas para a análise do estudo proposto, realizou-se limpeza de dados. Esta limpeza buscou remover dados inconsistentes, ajustes de dados errados e análise prévia dos extremos das variáveis. Após essa limpeza, a base de dados a ser utilizada permaneceu com 1.394 linhas para agosto de 2022, 411 linhas para agosto de 2021 e 298 linhas para agosto de 2020.

3.3 ANÁLISE EXPLORATÓRIA DOS DADOS

As variáveis independentes representam as características dos imóveis. Foram analisadas as relações entre as variáveis, estudando as dependências e correlações entre as variáveis selecionadas. Para realizar estas análises foi utilizado o *software* STATISTICA. Utilizou-se primeiro das variáveis contínuas para fazer a análise de correlações entre as variáveis.

O resultado da investigação para os dados de agosto de 2022 mostrou que quase todas as variáveis têm correlações significativas entre si, sendo a correlação entre preço e área a de maior valor. Por outro lado, cabe destacar que a variável “R$/m²” foi retirada das possíveis variáveis do modelo na segunda análise sobre os dados selecionados. Isto se deve ao fato de ela já ser uma variável modificada com a área, causando dependência.

Nesse sentido, as únicas variáveis que tiveram a análise de correlação como não significativas foram “referência de entrega” e “garagem”. A variável “referência de entrega” apresenta baixa correlação com as demais variáveis, como pode-se observar nos resultados apresentados na Tabela 3.

Tabela 3 - Correlações entre as variáveis contínuas de agosto de 2022

Variável	As correlações em vermelho são significativas em p < .05000 N=1394							
	Média	Desvio Padrão	Preço	Referência Entrega	Área	Quartos	Suítes	Garagem
Preço	973.241,7	864.015,3	1,0000	0,1543	0,9204	0,6024	0,6981	0,7672
Referência entrega	-48,5	109,8	0,1543	1,0000	-0,0624	-0,1845	0,0847	-0,0117
Área	130	84,6	0,9204	-0,0624	1,0000	0,7275	0,7525	0,8328
Quartos	2,8	0,8	0,6024	-0,1845	0,7275	1,0000	0,7582	0,7197
Suítes	2,2	1,2	0,6981	0,0847	0,7525	0,7582	1,0000	0,7620
Garagem	2,0	1,0	0,7672	-0,0117	0,8328	0,7197	0,7620	1,0000

Fonte: autoria própria.

O resumo das correlações determinadas é apresentado de forma decrescente na Tabela 4.

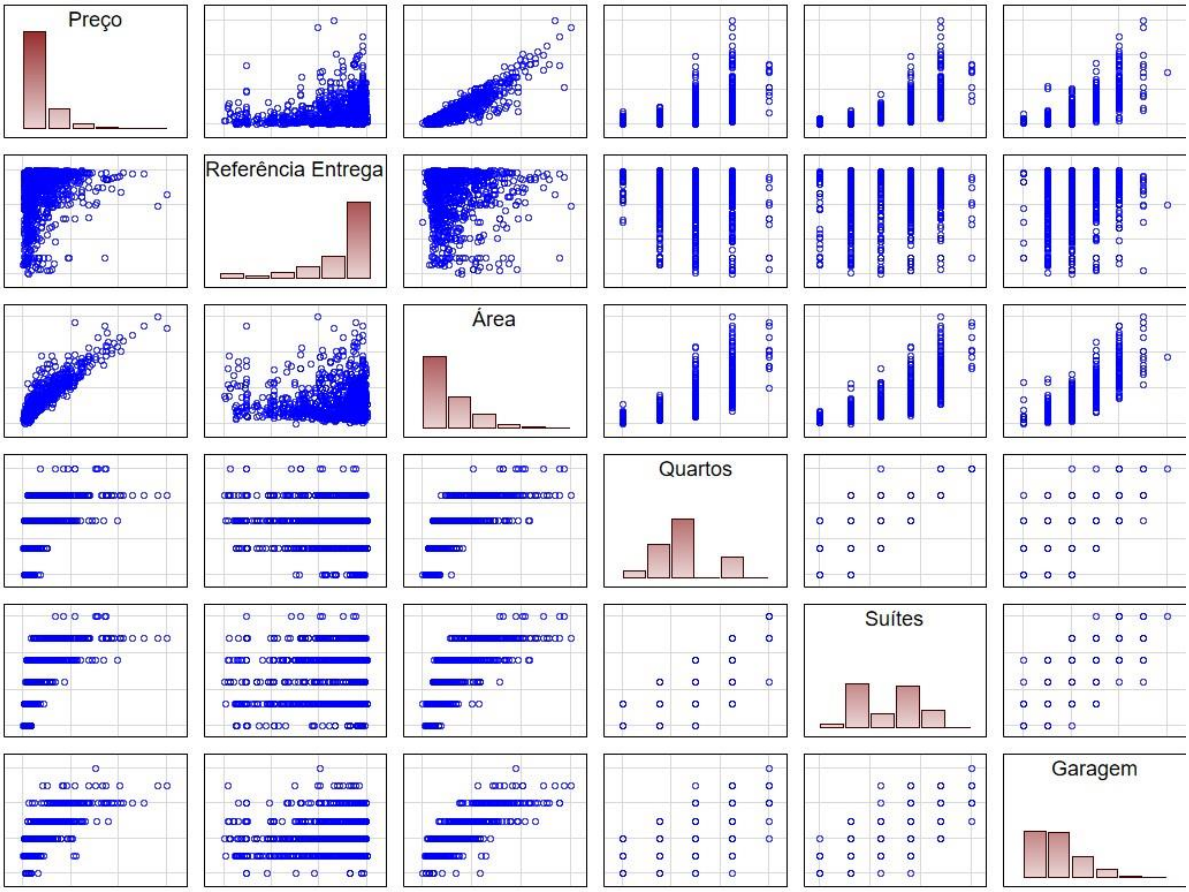
Tabela 4 - Correlações ordenadas de forma decrescente e valor absoluto com dados de agosto de 2022

Correlação	Variável	Variável
0,9204	Preço	Área
0,8328	Área	Garagem
0,7672	Preço	Garagem
0,7620	Suítes	Garagem
0,7582	Quartos	Suítes
0,7525	Área	Suítes
0,7275	Área	Quartos
0,7197	Quartos	Garagem
0,6981	Preço	Suítes
0,6024	Preço	Quartos
-0,1845	Referência entrega	Quartos
0,1543	Preço	Referência entrega
0,0847	Referência entrega	Suítes
-0,0624	Referência entrega	Área
-0,0116	Referência entrega	Garagem

Fonte: autoria própria.

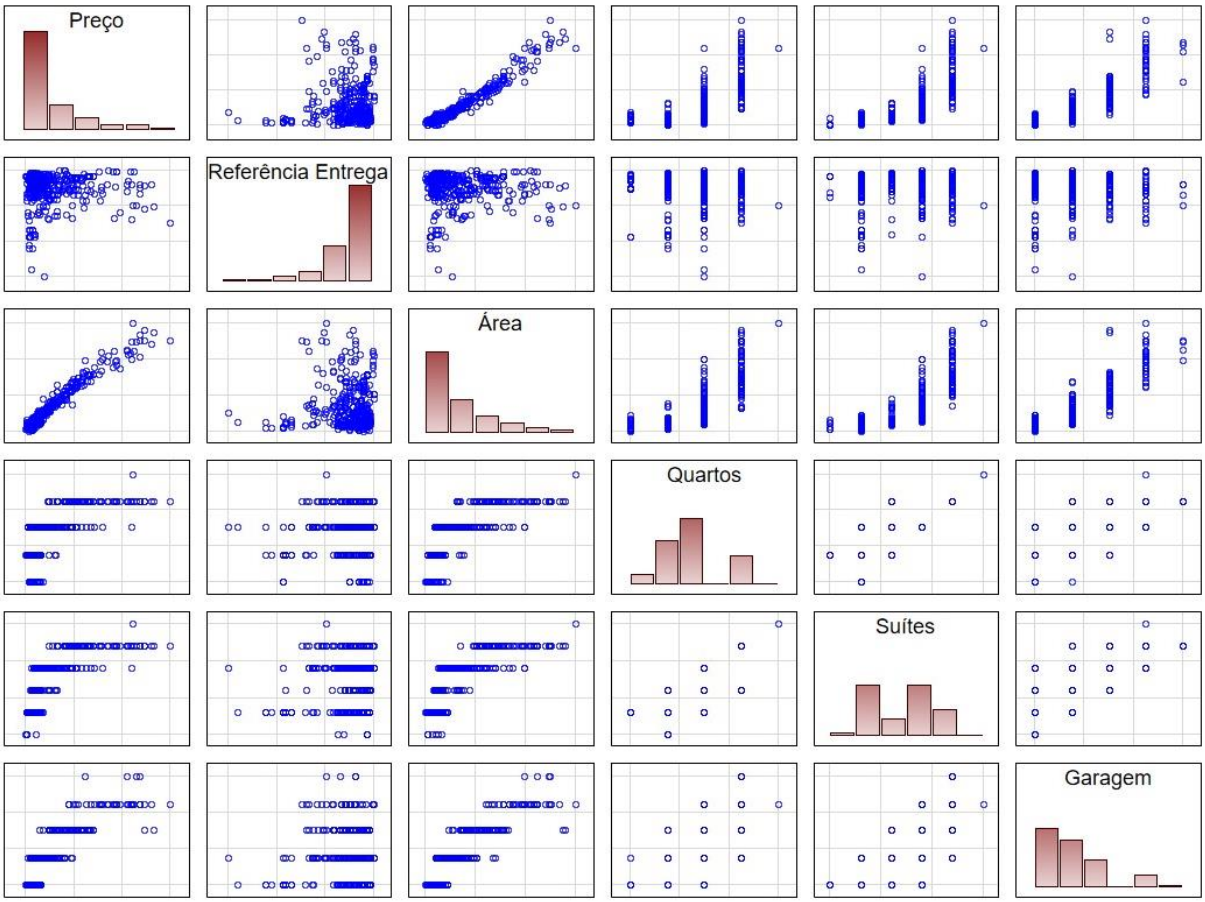
Nas Figuras 7, 8 e 9 são mostrados gráficos de dispersão e histogramas das variáveis, para os conjuntos de dados de agosto de 2022, agosto de 2021 e agosto de 2020, respectivamente. Como pode-se observar, semelhanças nestes gráficos são identificadas, apesar das distinções de tamanho da amostra e o conjunto de dados de agosto de 2022 conter empreendimentos que são de imóveis usados. Dentre os gráficos de dispersão, o único que visualmente demonstra tendência linear é o de “área” com o “preço”. No histograma, há concentração nos menores valores em “área” e “preço”. A variável “referência de entrega” tem concentração maior nos dados de entrega futura, o que condiz com a principal fonte da base de dados. Para a variável “quartos” observa-se uma maior concentração em imóveis de três quartos, enquanto, para a variável “suítes”, há uma maior concentração de imóveis com uma e três suítes. A variável “garagem” tem maior concentração em uma ou duas vagas de garagem.

Figura 7 - Variáveis e dispersão entre elas com dados de agosto de 2022



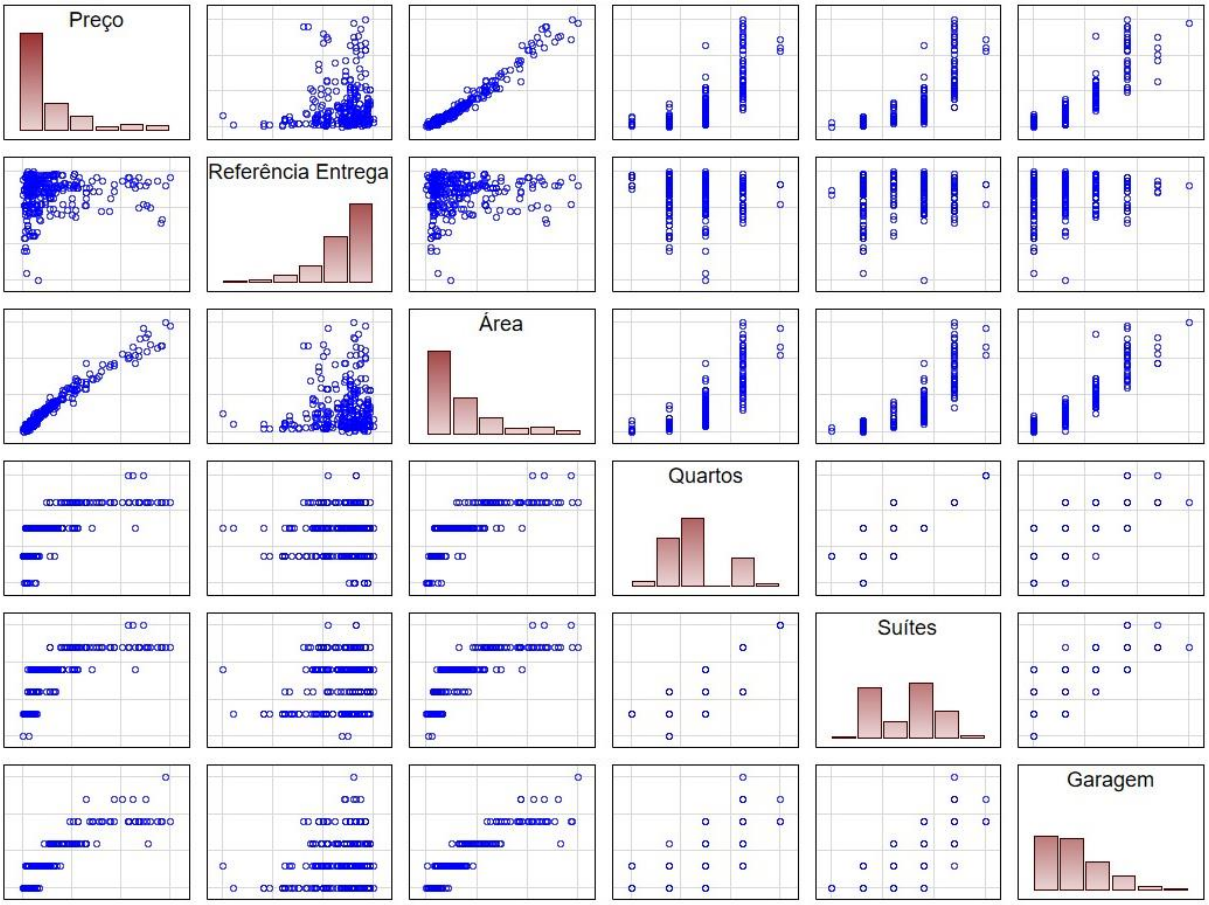
Fonte: autoria própria.

Figura 8 - Variáveis e dispersão entre elas com dados de agosto de 2021



Fonte: autoria própria.

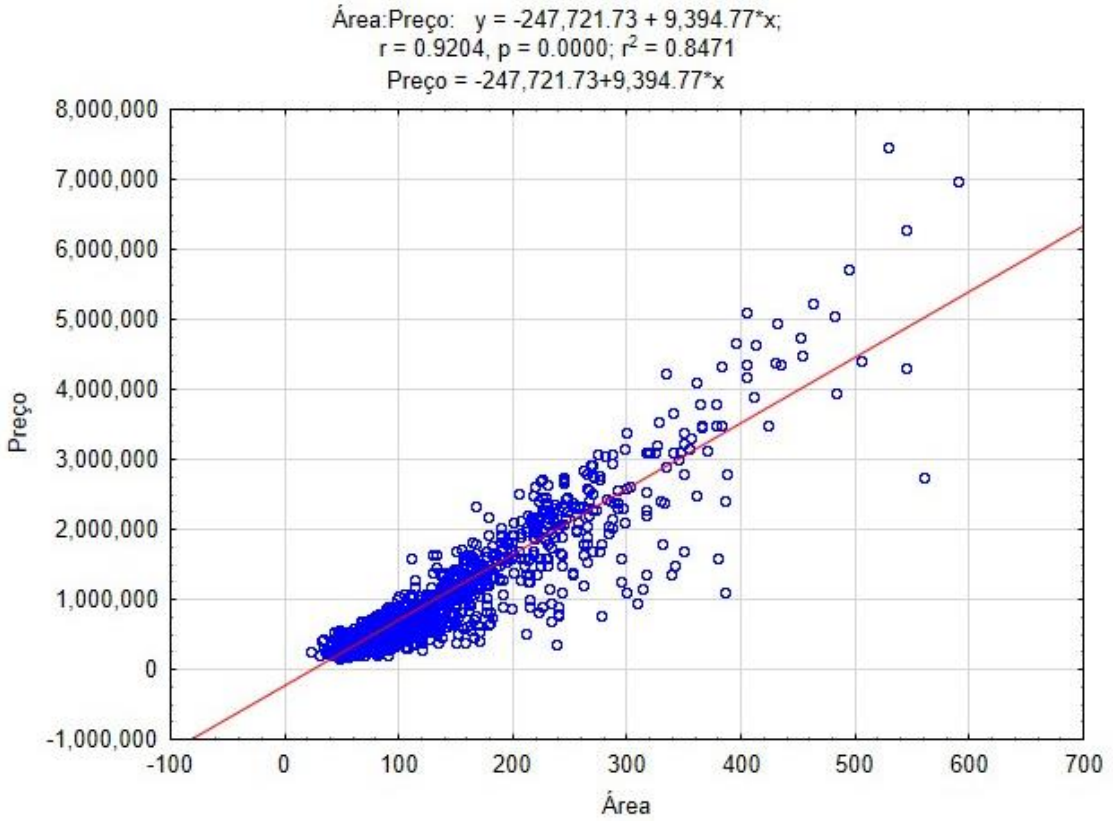
Figura 9 - Variáveis e dispersão entre elas com dados de agosto de 2020



Fonte: autoria própria.

A Figura 10 apresenta o recorte ampliado do gráfico de dispersão entre “preço” e “área”. Pode-se constatar que a variável “área” apresenta a partir do intervalo de 200 e 300 m², os pontos tendem a apresentar forte oscilação, enquanto, pontos que representam imóveis de menor área, ou seja, menores que 200 m² tendem a seguir a reta de regressão linear. Tal fenômeno pode ser explicado pela escassez de dados referentes a imóveis de maior área, ou outros fatores que influenciam no seu preço, já que são imóveis de padrão diferenciado. A regressão linear simples entre “preço” e “área”, como mostrado na Figura 10, apresenta coeficiente de determinação (R^2) de 0,8471.

Figura 10 - Dispersão Preço x Área



Fonte: autoria própria.

Após a primeira regressão, com a maior correlação existente com a variável “preço”, foi realizada a segunda regressão, desta vez adicionando ao modelo de regressão a variável “garagem”, a qual possui a segunda maior correlação com a variável “preço”. O coeficiente de determinação ajustado (R^2 ajustado) obtido para essa regressão foi de 0,8469. A variável “garagem” não apresentou nível de significância suficiente para ser considerada na regressão. Os resultados estão representados na Tabela 5.

Tabela 5 - Valores da regressão. Dependente: Preço e independentes: Área e Garagem

	Valor
R	0,9203
R²	0,8471
R² ajustado	0,8469
F(2,1391)	3.854,2974
p	0
Erro padrão de estimativa	338.053,89
b* da área	0,9185
b da área	9.376,00
p-valor da área	0,00
b* da garagem	0,002189
b da garagem	1.932,00
p-valor da garagem	0,9080

Fonte: autoria própria.

Na terceira regressão foi inserida a variável de número de "suítes" no modelo. Para esta situação obteve-se coeficiente de determinação ajustado (R² ajustado) o valor de 0,8468. Neste caso, a variável "área" foi a única considerada significativa para o modelo.

Uma quarta tentativa de regressão incluiu a variável de número de "quartos". Neste modelo, as variáveis "área", "quartos" e "suítes" foram estatisticamente significativas para o modelo. Esta regressão conduziu ao coeficiente de determinação ajustado (R² ajustado) de 0,8603. Os resultados são apresentados nas Tabelas 6 e 7.

Tabela 6 - Sumário estatístico da regressão. Dependente: Preço e independente: Área, Garagem, Quartos e Suítes

	Valor
R	0,9277
R²	0,8607
R² ajustado	0,8603
F(4,1389)	2.145,5986
p	0
Erro padrão de estimativa	322.937,90
Intercepto	47.687

Fonte: autoria própria.

Tabela 7 - Valores da regressão. Dependente: Preço e independente: Área, Garagem, Quartos e Suítes

	b*	**b**	**p-valor**
Área	0,9621	9821	0
Garagem	0,0313	27627	0,1127
Suítes	0,0971	70455	0
Quartos	-0,1937	-201493	0

Fonte: autoria própria.

A inclusão da variável "referência de entrega" conclui o processo de inserção contínua de variáveis contínuas disponíveis na análise de regressão linear múltipla. O modelo de regressão com todas as variáveis contínuas resultou para o R^2 ajustado o valor de 0,8950. Nesse modelo, as variáveis "área", "quartos" e "referência de entrega" são significativas para a regressão, como mostram as Tabelas 8 e 9.

Tabela 8 - Sumário estatístico da regressão. Dependente: Preço e independente: Área, Garagem, Referência entrega, Quartos e Suítes

	Valor
R	0,9462
R^2	0,8953
R^2 ajustado	0,8950
F(5,1388)	2. 375,53134
p	0
Erro padrão de estimativa	279.983,598
Intercepto	-32.869,6974

Fonte: autoria própria.

Tabela 9 - Valores da regressão. Dependente: Preço e independente: Área, Garagem, Referência entrega, Quartos e Suítes

	b*	b	p-valor
Área	0,9919	10.125,0644	0
Garagem	0,0109	9.708,41623	0,5208
Suítes	-0,0138	-10.065,587	0,3959
Quartos	-0,0791	-82.330,9974	0
Referência entrega	0,2029	1.596,6155	0

Fonte: autoria própria.

No intuito de seguir o Princípio da Parcimônia, foi elaborado um modelo de regressão linear múltipla envolvendo apenas as variáveis que se mostraram significativas no modelo descrito anteriormente. Neste modelo, que inclui as variáveis “preço”, “área”, “quartos” e “referência de entrega” apresentou coeficiente de determinação ajustado (R^2 ajustado) de 0,8950, como apresentado na Tabela 10.

Tabela 10 – Valores da regressão com as 3 variáveis significantes. Dependente: Preço e independente: Área, Quartos e Referência de entrega

	Valor
R	0,9462
R^2	0,8952
R^2 ajustado	0,8950
F(3,1390)	3. 961,97884
p	0
Erro padrão de estimativa	279.875,178
Intercepto	-27.031,70
b da área	10.141,4179
b de quartos	-86.629,6331
b da referência de entrega	1.581,1446

Fonte: autoria própria.

A título demonstrativo, como os dados dos anos de 2021 e 2020 não possuem imóveis usados, procedeu-se da mesma maneira com os dados do ano de 2022. A Tabela 11 apresenta um resumo das regressões para cada ano.

Tabela 11 - Comparativo entre regressões de anos diferentes

	Agosto de 2020	Agosto de 2021	Agosto de 2022
Tamanho da amostra	298	411	779
R	0,9869	0,9788	0,9792
R^2	0,9741	0,9581	0,9588
R^2 ajustado	0,9736	0,9576	0,9585
Erro padrão de estimativa	130.231,005	175.751,026	192.884,694
Intercepto	-9.927	-62.510	-94.805
b da área	9.396	10.332	11.699
b de quartos	-105.410	-107.532	-106.129
b de suítes	-34.637	-26.167	-19.087
b de garagem	-25.847	3.454	-9.825
b da referência de entrega	-65	379	1.405

Fonte: autoria própria.

O conjunto de 2020 teve o valor de R^2 ajustado melhor que os dos outros anos, o que mostra que seus dados estão distribuídos de forma mais homogênea. O conjunto de 2022 resultou no modelo em que somente “suítes” e “garagens” não são significativos para a regressão linear. A Tabela 11 mostra, em vermelho, a partir do intercepto, as variáveis que são significativas para o modelo que considera somente imóveis novos, enquanto, a Tabela 10 são apresentados os resultados que inclui imóveis novos e usados. Os resultados mostram que a exclusão de imóveis usados na regressão proporciona um incremento de 6,35% no coeficiente de determinação ajustado.

3.4 APLICAÇÃO E AVALIAÇÃO DO MODELO DE PREVISÃO DE PREÇOS

Nas análises desse tópico, foi utilizada a linguagem *Python* para testar o modelo com três variáveis significativas para a base de agosto de 2022, com e sem os imóveis usados para melhor compreensão e análise.

Para testar os erros do modelo com imóveis usados, utilizou-se de validação cruzada com uma divisão na base com 30% para dados de teste, sendo realizada uma nova regressão linear com os outros 70% da base. Testaram-se 50 estados aleatórios para cálculo dos erros, que na métrica EPAM variou de 19,8% a 25,6%. Estado aleatório é utilizado em *Python* para que os dados do conjunto se dividam de forma aleatória, de modo que cada estado aleatório difira do outro. É utilizado de um número de referência para cada estado aleatório testado, para que se possa repetir a mesma distribuição de dados quantas vezes forem necessárias.

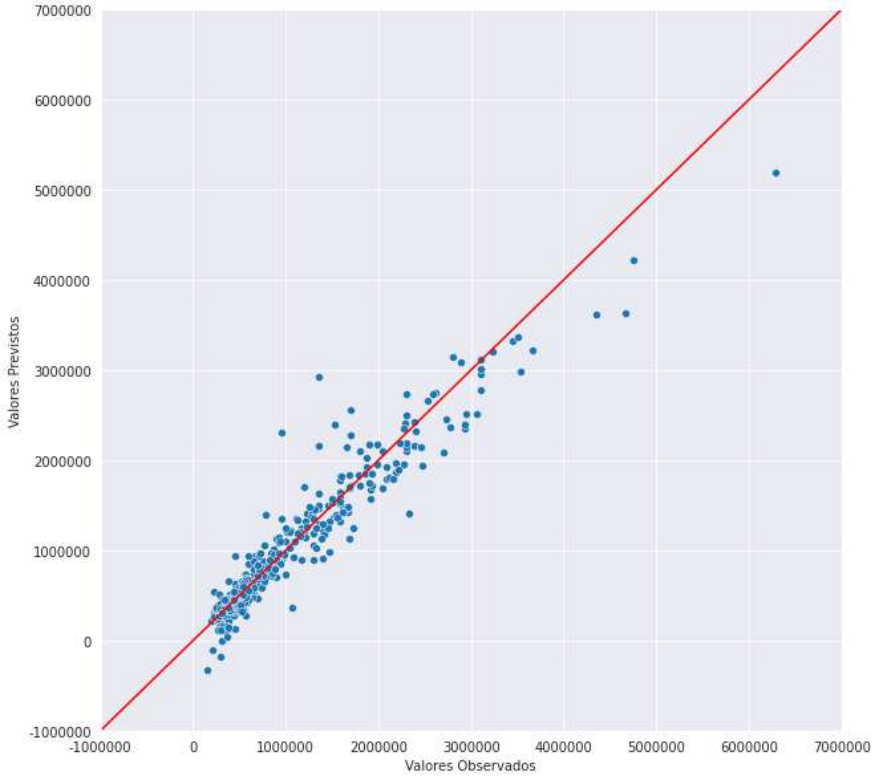
A Tabela 12 mostra os dados da regressão linear e os erros. A Figura 11 apresenta o gráfico de dispersão que compara valores observados e valores preditos para a divisão com o EPAM de 19,8%. Na regressão com dados que incluíam imóveis usados, é interessante ressaltar que apareceram valores preditos negativos, que na realidade não existem.

Tabela 12 – Dados da regressão de teste e erros do modelo com usados

Tamanho da amostra da regressão	975
Tamanho da amostra de teste	419
R^2	0,8861
R^2 ajustado	0,8858
Intercepto	6.654,2526
b da Referência de entrega	1.506,8774
b da Área	10.213,2015
b dos Quartos	-105.150,8343
EQM	R$² 58.660.403.955,76
REQM	R$ 242.199,09
EPAM	19,8%

Fonte: autoria própria.

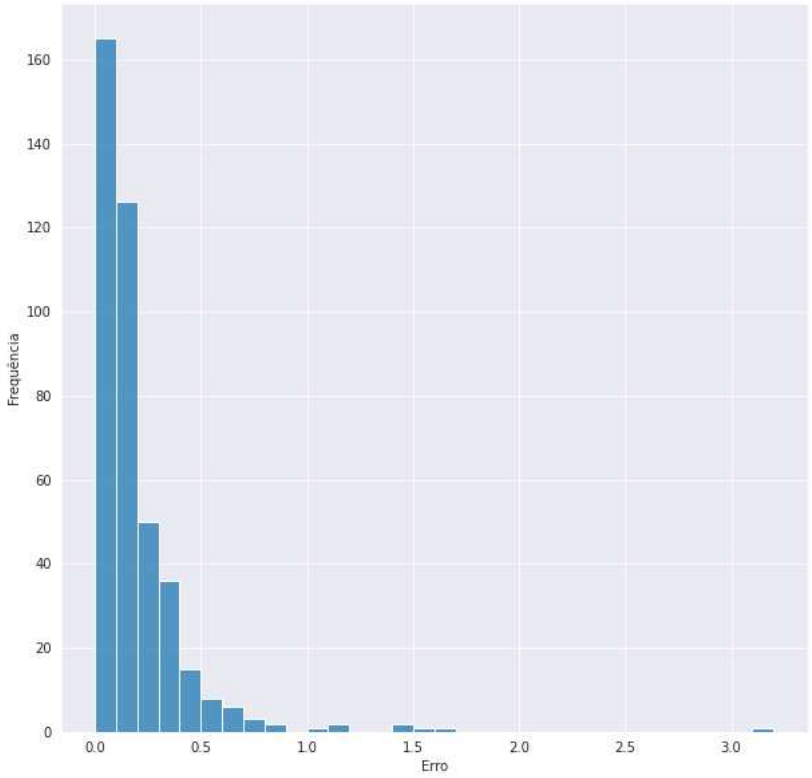
Figura 11 - Dispersão entre valor predito e valor observado com imóveis usados



Fonte: autoria própria.

A Figura 12 mostra o histograma com os erros de cada amostra de teste. Observa-se que 70% das amostras têm valor de EPAM menor que o EPAM médio. O que indica que na maioria dos casos o erro do modelo é baixo.

Figura 12 – Histograma com o EPAM (em porcentagem) de predição da base com usados



Fonte: autoria própria.

Para testar os erros do modelo sem imóveis usados foi feito uma divisão na base com 30% para dados de teste, e foi feita uma nova regressão linear com os outros 70% da base. Testou-se 50 estados aleatórios para cálculo dos erros, que na métrica EPAM variou de 12,8% a 16,7%.

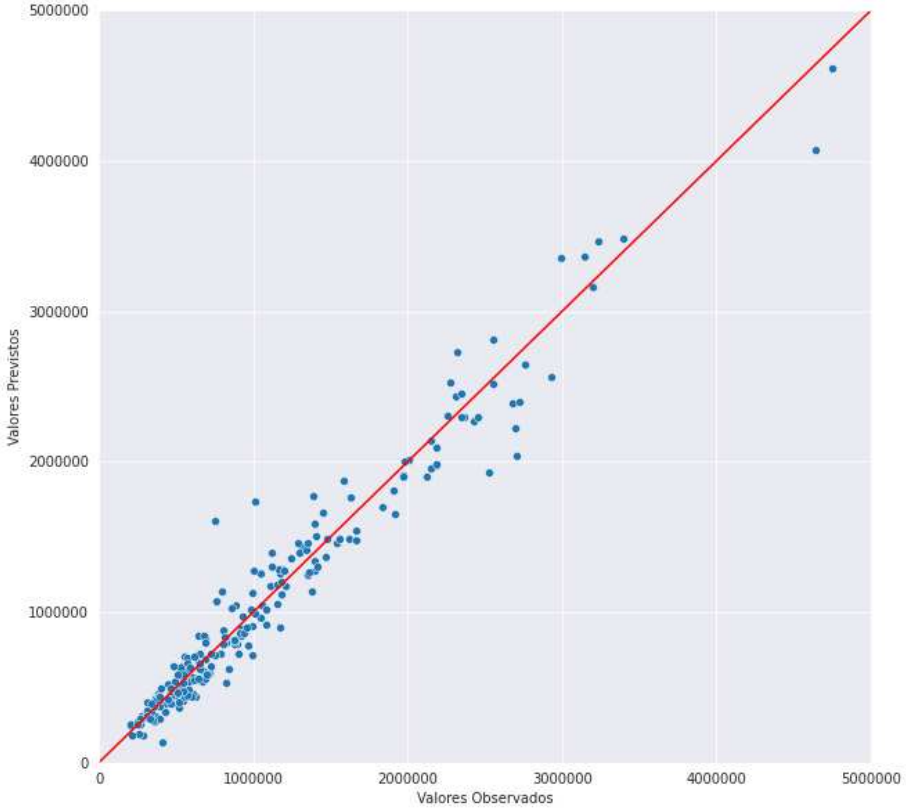
A Tabela 13 mostra os dados da regressão linear e os erros. A Figura 13 apresenta o gráfico de dispersão com as mesmas características da Figura 11, o que muda são os dados utilizados. Na regressão com a base de dados que excluiu os imóveis usados, não gerou na base de testes valores negativos e resultou em erro de 6,9% menor que na base que inclui imóveis usados.

Tabela 13 - Dados da regressão de teste e erros do modelo sem usados

Tamanho da amostra da regressão	542
Tamanho da amostra de teste	234
R^2	0,9562
R^2 ajustado	0,9560
Intercepto	-95.4007,4481
b da Referência de entrega	1.512,3167
b da Área	11.420,6132
b dos Quartos	- 118.483,6752
EQM	R$² 31.375.383.861,60
REQM	R$ 177.130,9793
EPAM	12,7934%

Fonte: autoria própria.

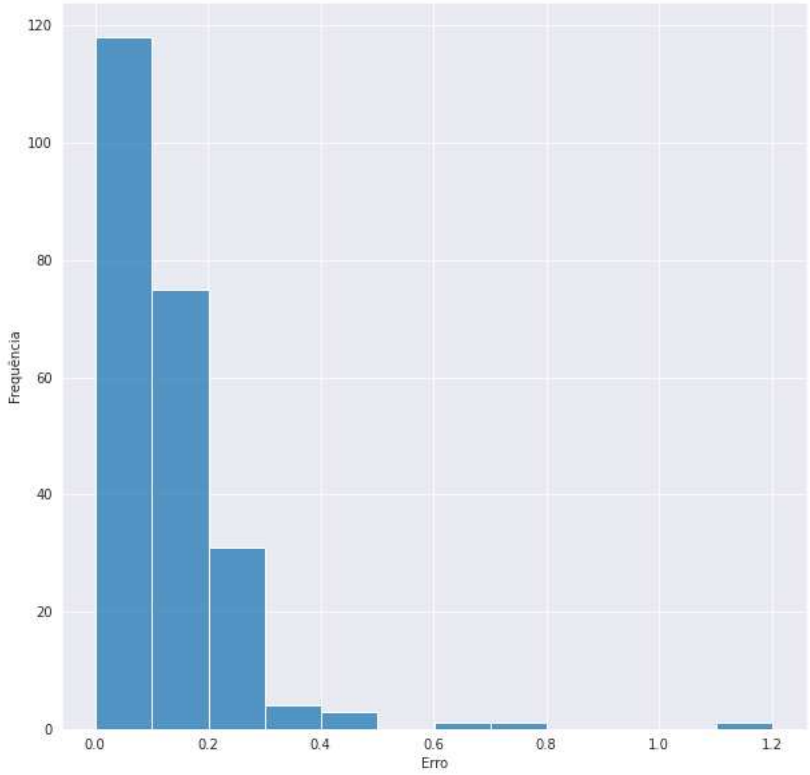
Figura 13 - Dispersão entre valor predito e valor observado sem imóveis usados



Fonte: autoria própria.

A Figura 14 apresenta o histograma com o erro de cada amostra de teste, dos dados sem imóveis usados. Diferentemente da Figura 10, esse histograma está balanceado. Para o valor EPAM dos dados de teste tem-se 61,3% dos valores EPAM com valor abaixo do EPAM médio.

Figura 14 - Histograma com o EPAM (em porcentagem) de predição da base sem usados



Fonte: autoria própria.

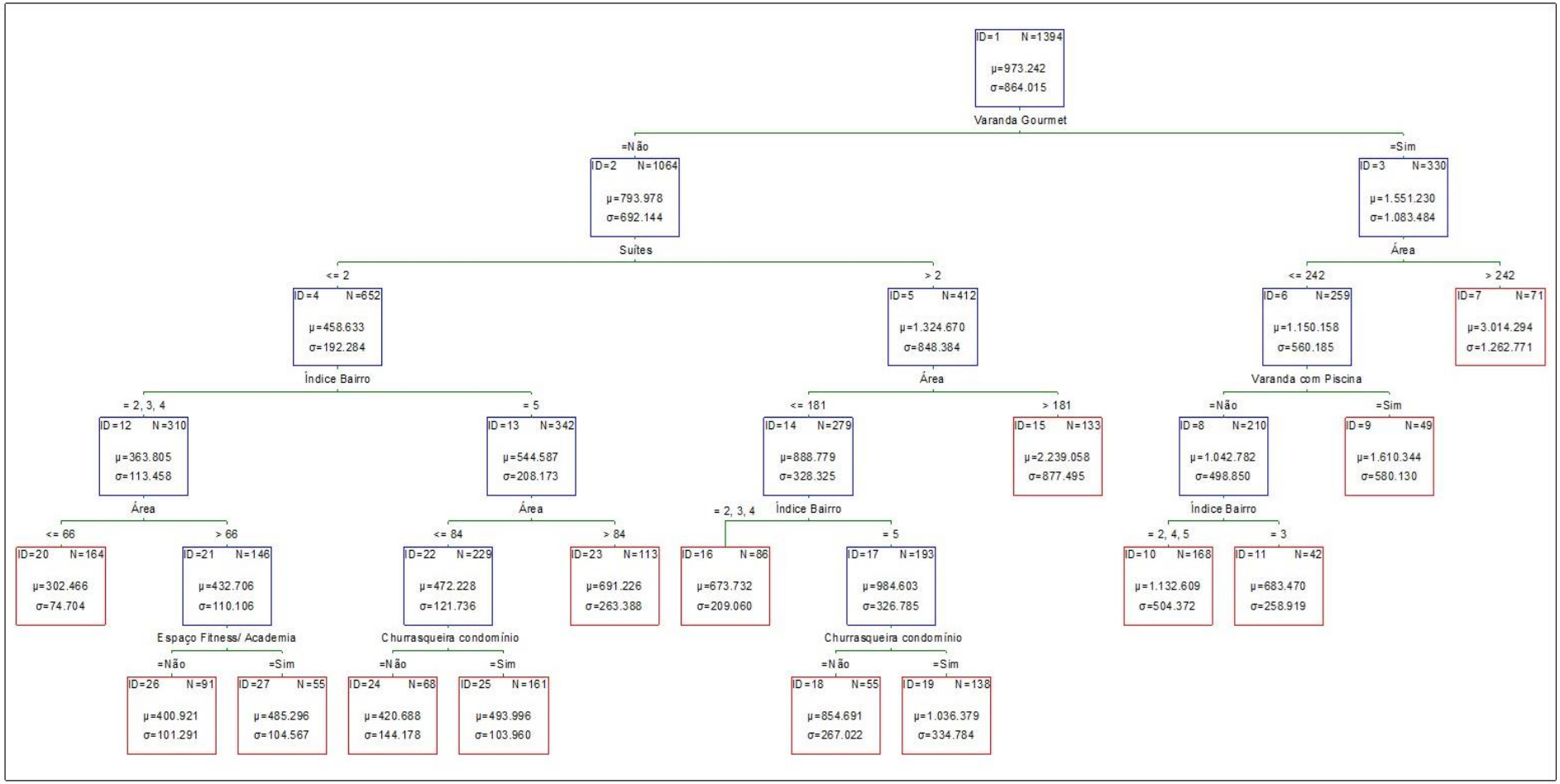
3.5 ANÁLISE DAS VARIÁVEIS CATEGÓRICAS

Nesta seção é abordada a inclusão das variáveis categóricas que estão presentes na base de dados, para verificar quais variáveis podem ou não influenciar nos preços dos imóveis. Para esta análise foi utilizado o *software* STATISTICA na construção da árvore de decisão CHAID e regressões. A linguagem *Python* foi utilizada para testar os erros das regressões. Para gerar o modelo da árvore de decisão foi utilizado do conjunto de dados que contém os imóveis usados.

A Figura 15 mostra a árvore de decisão obtida com a aplicação do método CHAID sobre a amostra de dados dos empreendimentos. Para o método do qui-quadrado a variável "varanda gourmet" é a variável que apresenta o melhor qui-quadrado para iniciar a divisão.

Apesar da árvore CHAID aceitar dividir em mais de dois nós, o resultado que se obteve com a nossa base de dados foi sempre de duas divisões.

Figura 15 –Árvore CHAID

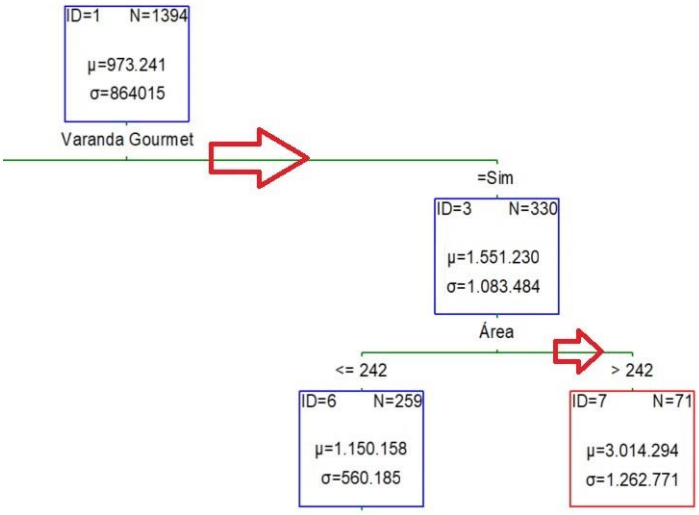


Fonte: autoria própria.

A Figura 16 apresenta a folha com as unidades que têm a maior média de preço dentre as folhas que foram divididas na árvore. Este subconjunto compreende 71 unidades, o qual caracteriza-se pelo caminho “sim” para “varanda gourmet” e “área” maior que 242 m². A unidades desse subconjunto apresentam valor médio do imóvel de R$3.014.294 e desvio padrão de R$1.266.771 (subconjunto identificado pela “folha” ID=7 da árvore).

As unidades desse subconjunto da amostra apresentam os maiores valores de imóveis. Esta classificação ditada pelo modelo CHAID considera a existência de “Varanda Gourmet” no imóvel assim como uma ampla área do imóvel como características decisivas para descriminar esta categoria de imóveis na amostra.

Figura 16 - Caminho da árvore para unidades com maiores preços



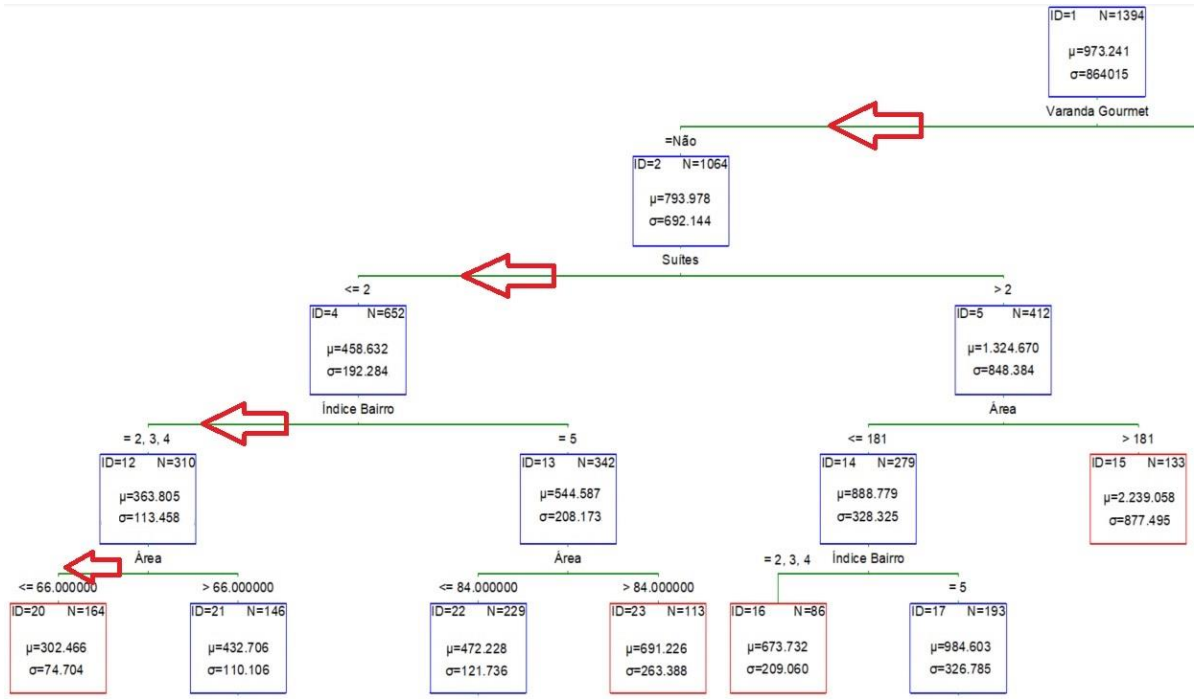
Fonte: autoria própria.

A Figura 17 apresenta o caminho para a folha com as unidades que têm a menor média de preços dentre as folhas que foram divididas na árvore. Este subconjunto compreende 164 unidades, o qual caracteriza-se pelo caminho “não” para “varanda gourmet”, com duas ou menos “suítes” (<=2), com bairro com índice que seja diferente de cinco (2,3 ou 4) e que a “área” seja menor ou igual a 66 m². A unidades desse subconjunto apresentam valor médio do imóvel de R$302.466 e desvio padrão de R$74.704 (subconjunto identificado pela “folha” ID=20 da árvore).

As unidades desse subconjunto da amostra apresentam os menores valores de imóveis. Esta classificação ditada pelo modelo CHAID considera a ausência de “varanda gourmet” no imóvel, que ele possua poucas “suítes”, que não esteja nos melhores bairros e que seja um

imóvel compacto como características decisivas para descriminar esta categoria de imóveis na amostra.

Figura 17 - Árvore com as decisões da folha para menor preço



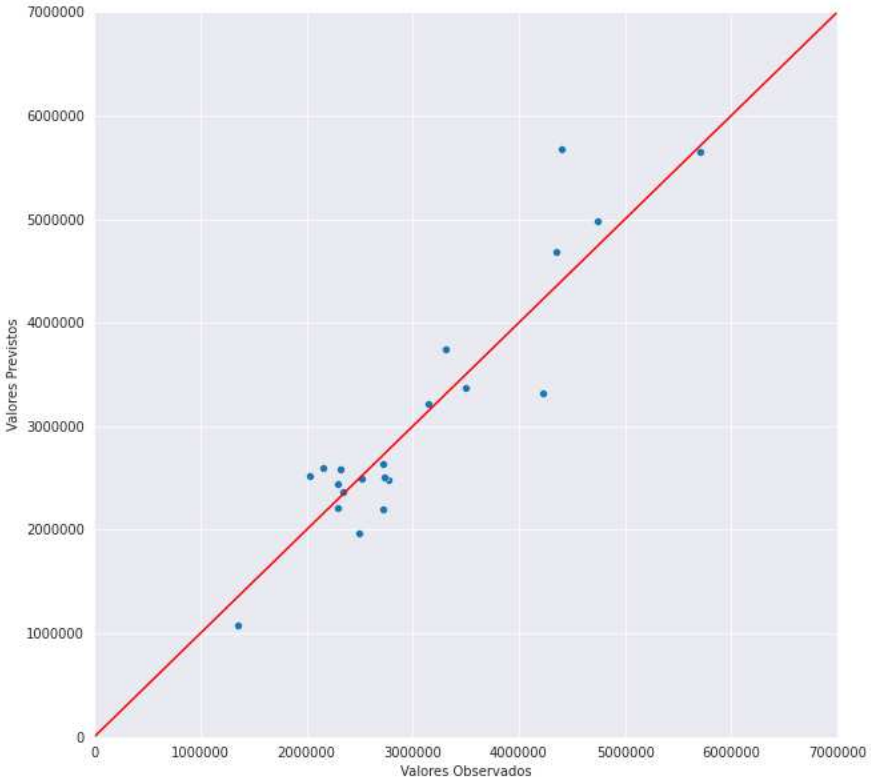
Fonte: autoria própria.

Por exemplo, se na Figura 15 utilizar-se as folhas com ID=26 e ID=27, que seguiram o caminho de “não” para “varanda gourmet”, com duas ou menos (<=2) “suítes”, nos bairros com índice dois, três ou quatro (2,3 ou 4), com “área” maior que 66 m² (>66), sendo a última decisão a ser tomada para chegar nas folhas a de “sim” ou “não” para “academia” no condomínio. A folha de ID=26 apresenta média de preços de R$400.921 enquanto a folha de ID=27, apresenta média de R$485.296. Pode-se dizer que, em média, a presença de academia no condomínio do imóvel reflete em aumento de preço médio de R$84.374.

Realizada as análises sobre a árvore de decisão, testou-se a possibilidade de realizar regressões lineares múltiplas para as duas folhas que foi demonstrado o caminho decisório na Figura 16 e Figura 17, com as folhas de ID=7 e ID=20. A regressão para a folha com ID=7, resultou em R^2 ajustado de 0,8733 e com “referência de entrega” e “área” como as variáveis significativas para o modelo. Este modelo foi testado na análise de erros em 10 estados aleatórios com divisão de 70% para o modelo e 30% para teste, obteve-se o valor do EPAM

entre 9,82% até 15,90%. A Figura 18 mostra o gráfico de dispersão entre valor observado e previsto para o erro com 9,82%.

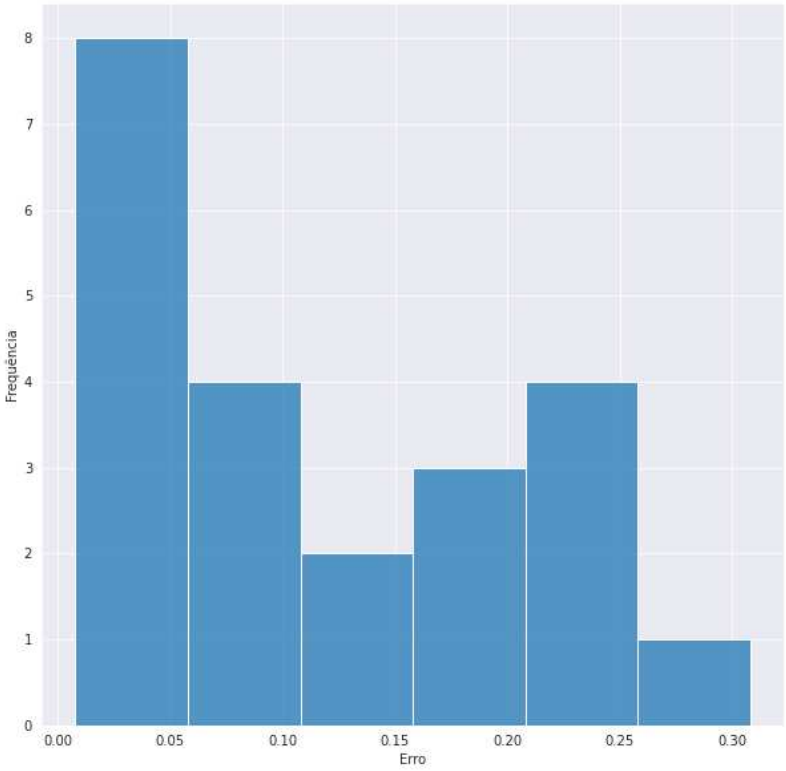
Figura 18 - Dispersão entre valor observado e valor real para a folha de ID=7



Fonte: autoria própria.

A Figura 19 apresenta o histograma de erro para cada imóvel no conjunto de teste da validação cruzada. Para o EPAM médio de 9,82%, tem-se 12 amostras de um total de 22 que apresentam EPAM médio inferior ao EPAM médio, ou seja, 54,5% dos dados de teste.

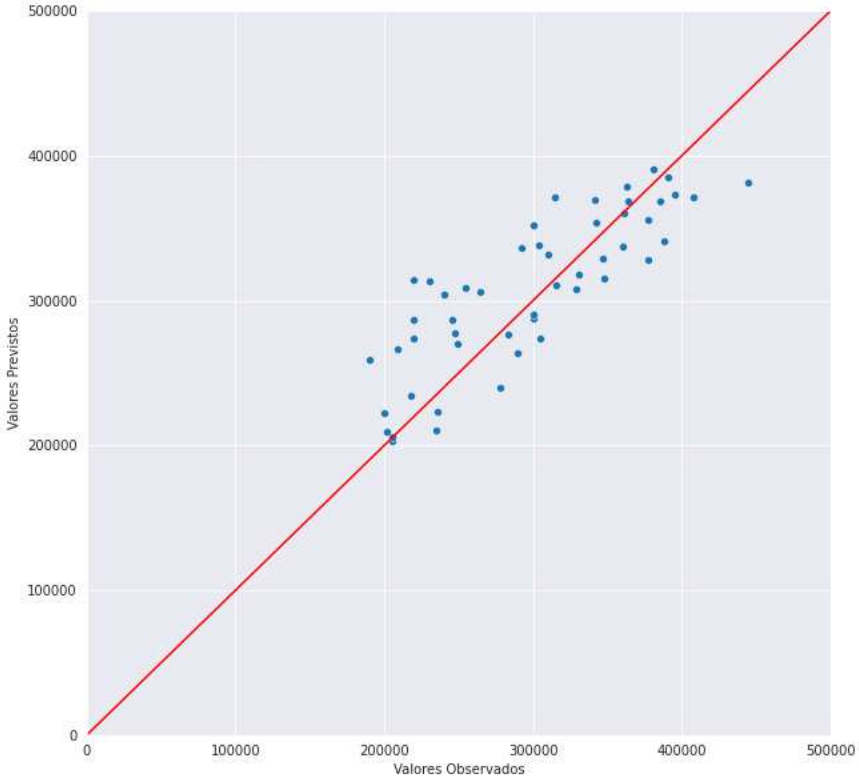
Figura 19 - Histograma de EPAM (em porcentagem) para a folha de ID=7



Fonte: autoria própria.

A regressão realizada para folha de ID=20 resultou em regressão linear múltipla com R^2 ajustado de 0,4827 com “referência de entrega”, “área” e “quartos” como variáveis significativas para o modelo. Testado este modelo para a análise de erros em 10 estados aleatórios com divisão de 70% para o modelo e 30% para teste, obteve-se resultado de EPAM variando entre 11,3% até 17,3%. A Figura 20 mostra o gráfico de dispersão entre valor observado e previsto para o EPAM com 11,3%.

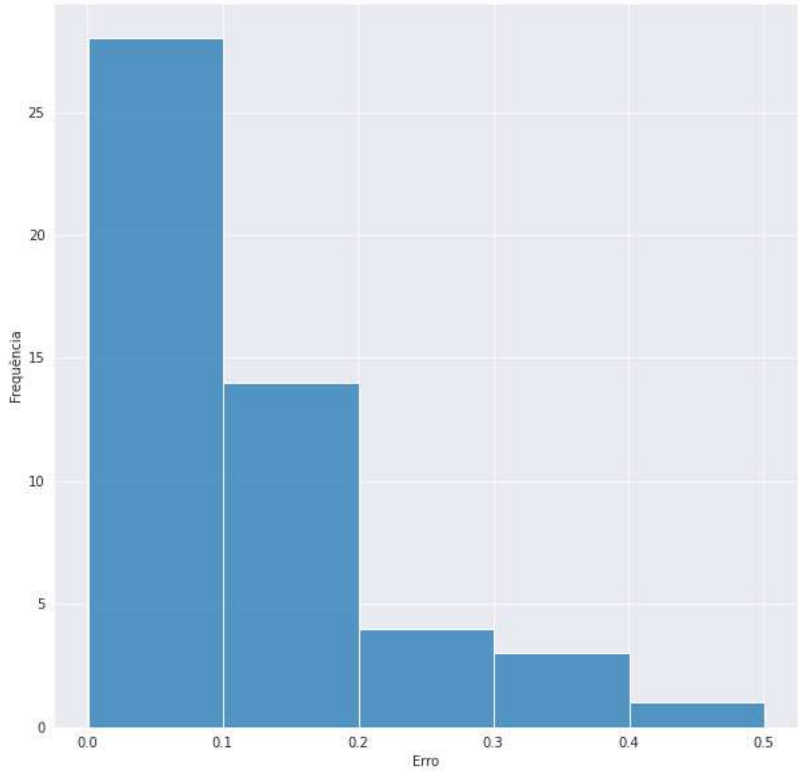
Figura 20 - Dispersão entre valor observado e valor real para a folha de ID=20



Fonte: autoria própria.

A Figura 21 apresenta o histograma de erro para cada ponto de teste, abaixo da média do EPAM tem-se 32 amostras que representam 64% dos dados de teste.

Figura 21 - Histograma de EPAM (em porcentagem) para a folha de ID=20



Fonte: autoria própria.

A compilação dos resultados relativos ao erro EPAM, obtidos para cada uma das folhas da árvore é apresentada na Tabela 14. Estes resultados mostram que tantos os erros mínimo, médio e máximo obtidos para cada uma das folhas são inferiores aos respectivos erros obtidos pela regressão múltipla, a qual considera todos os elementos da amostra.

Considerando valores médios dos erros da regressão para os subconjuntos que constituem as folhas da árvore, verifica-se que:

- EPAM mínimo houve uma melhoria de 19,8% da regressão múltipla para 10,8% para a média dos erros obtidos com a regressão aplicada aos subconjuntos na amostra;
- EPAM máximo houve uma melhoria de 25,6% da regressão múltipla para 19,0% para a média dos erros obtidos com a regressão aplicada aos subconjuntos na amostra;
- EPAM médio houve uma melhoria de 22,7% da regressão múltipla para 14,9% para a média dos erros obtidos com a regressão aplicada aos subconjuntos na amostra.

Por outro lado, o valor médio do R^2 ajustado da regressão múltipla considerando toda a amostra foi consideravelmente superior (=0,88) ao valor médio (=0,65) dos R^2 ajustados das regressões múltiplas considerando cada um dos subconjuntos da amostra. Este resultado

desfavorável à abordagem apresentada é, no entanto, enviesado pois esta comparação é realizada com amostras de diferentes tamanhos além de variâncias diferentes relativas aos subconjuntos da amostra.

Tabela 14 – Resumo de R² ajustado e EPAM para cada folha da Árvore CHAID

ID da folha	R² ajustado	EPAM mínimo	EPAM máximo	EPAM médio
ID=7 (n=71)	0.87	9.8%	15.9%	12.9%
ID=9 (n=49)	0.85	10.3%	19.8%	15.1%
ID=10 (n=168)	0.80	10.2%	18.5%	14.4%
ID=11 (n=42)	0.88	6.5%	17.4%	12.0%
ID=15 (n=133)	0.68	13.5%	25.6%	19.6%
ID=16 (n=86)	0.77	8.1%	14.5%	11.3%
ID=18 (n=55)	0.66	9.5%	22.2%	15.9%
ID=19 (n=138)	0.66	10.9%	15.8%	13.4%
ID=20 (n=164)	0.48	11.3%	17.3%	14.3%
ID=23 (n=113)	0.48	12.0%	23.9%	18.0%
ID=24 (n=68)	0.67	15.4%	22.9%	19.2%
ID=25 (n=161)	0.57	11.3%	14.3%	12.8%
ID=26 (n=91)	0.32	14.8%	19.3%	17.1%
ID=27 (n=55)	0.51	6.9%	18.0%	12.5%
Média	0.65	10.8%	19.0%	14.9%
Regressão múltipla (n=1394)	0.88	19.8%	25.6%	22.7%

Fonte: autoria própria.

Os resultados desta abordagem mostram que a utilização de árvore de decisão para classificar uma amostra de dados é um caminho promissor para obter estimativas oriundas de regressões relativas à precificação de imóveis.

Considerando que esses resultados aparentavam um caminho promissor, procedeu-se a análise dos erros residuais das regressões realizadas. O erro residual é a soma da diferença do valor observado do valor predito para cada amostra. Se esse valor for zero, significa que o modelo está equilibrado. Para a base completa de cada amostra os erros residuais possuíam valor próximo de zero.

Os valores do erro de resíduos ao utilizar-se do teste de validação cruzada com divisão de 70% e 30% para os menores valores de erro EPAM obtiveram uma melhora na soma dos erros das folhas em relação ao erro da regressão linear múltipla geral. Os erros residuais são apresentados na Tabela 15.

Tabela 15 - Erros residuais das folhas e da regressão linear

ID da folha (n amostra de teste)	Média do erro residual	Curtose	Assimetria
10 (n=50)	-8.378	2,22	0,69
19 (n=41)	-15.150	2,97	0,89
18 (n=17)	13.055	-0,99	0,15
23 (n=34)	32.429	6,98	2,18
25 (n=47)	19.884	-0,63	0,17
24 (n=21)	-12.095	0,52	0,39
9 (n=15)	-79.583	-0,28	0,34
27 (n=17)	6.849	-0,17	-0,06
26 (n=28)	6.782	-0,35	-0,23
20 (n=48)	-9.306	-0,64	-0,33
7 (n=22)	133.744	0,91	-0,29
11 (n=13)	41.748	-0,49	-0,18
15 (n=40)	-22.487	7,88	-1,82
16 (n=26)	6.206	0,82	1,13
Resíduos das folhas (n=419)	8.121	1,33	0,21
Resíduos da regressão (n=419)	15.702	9,62	-0,70

Fonte: autoria própria.

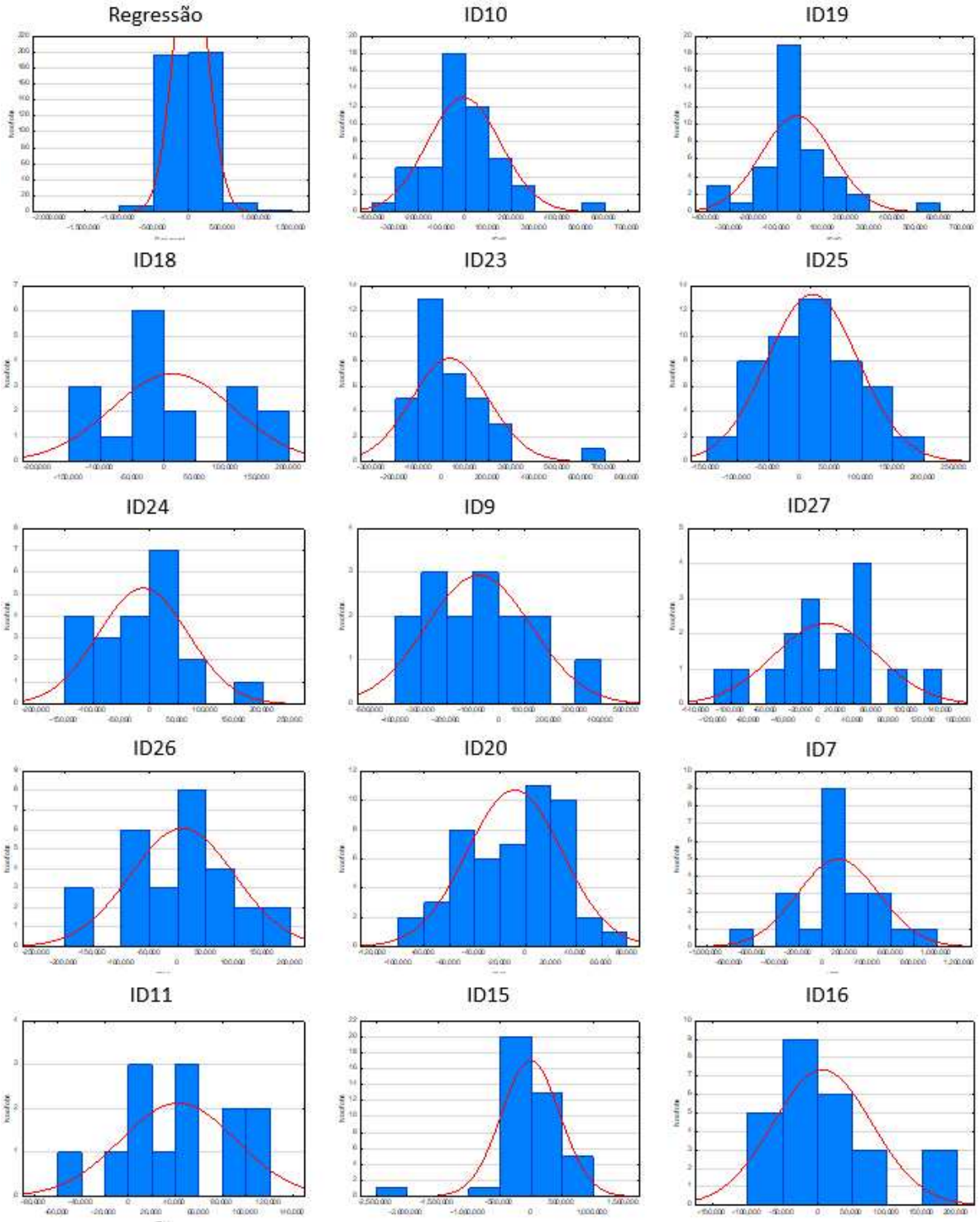
Dos dados apresentados na Tabela 15, observa-se que seis folhas têm valor médio de erro residual maior que a da regressão linear com todos os dados. Em seguida, nota-se que outras oito folhas possuem em valor média de erro residual menor do que a regressão linear com todos os dados.

Ao analisar os valores de curtose para cada folha e regressão linear com todos os dados, observa-se que a regressão linear com todos os dados e as folhas ID10, ID19, ID23, ID24, ID7, ID15 e ID16 apresentam uma curva de frequências mais aberta quando comparada com a distribuição normal. Por outro lado, as folhas ID18, ID25, ID9, ID27, ID26, ID20, ID11 apresentam uma curva de frequências mais fechada que a distribuição normal.

Na análise de assimetria, as folhas de ID16 e ID23 têm distribuição com assimetria positiva e a folha ID15 tem distribuição com assimetria negativa. As demais folhas e a regressão linear com todos os dados têm distribuição simétrica.

A Figura 22 apresenta os histogramas de cada um dos conjuntos de dados com a curva normal para uma análise visual.

Figura 22 - Histogramas com curva normal referentes aos erros residuais



Fonte: autoria própria.

Dos gráficos apresentados na Figura 22 observa-se que há tendência a seguir normalidade. No entanto, ocorrem exceções como o ID23 e o ID20 que têm um deslocamento do centro. No ID11 e ID18 têm-se uma maior amplitude. Essas exceções podem ocorrer pela consideração dos dados de teste, que são 30% do conjunto de dados causando maior variabilidade e poucos dados na amostragem.

Diante disso, a análise pelos erros residuais e normalidade apresenta resultados que seguem apontando a utilização da abordagem de árvore de decisão para classificar uma amostra de dados, sendo um caminho promissor para obter estimativas oriundas de regressões relativas à precificação de imóveis.

4 CONCLUSÕES

O objetivo geral do presente trabalho foi propor um modelo de precificação de imóveis em uma cidade do Centro-Oeste, considerando dados de imóveis disponíveis *online* e disponibilizados por uma empresa do ramo. Para tanto foram aplicadas técnicas estatísticas e matemáticas, tais como análise de correlação, regressão linear para as variáveis contínuas, análise de erro para a regressão e árvore de decisão com o método CHAID para incluir variáveis categóricas na análise.

A análise de regressão múltipla foi realizada com a inclusão sucessiva de variáveis que apresentaram maior correlação com o preço dos imóveis. Em cada rodada foram adicionadas variáveis ao modelo de regressão, podendo-se assim avaliar o comportamento destas variáveis na qualidade da regressão. Esta análise permitiu determinar quais variáveis foram significativas para o modelo de regressão no conjunto de dados de agosto de 2022 incluindo-se dados de imóveis usados. Com o resultado desse procedimento foram selecionadas as variáveis: “Área”, “Referência de Entrega” e “Número de Quartos”. O mesmo procedimento de regressão foi aplicado no conjunto de dados de agosto de 2022, excluindo dados de imóveis usados. Neste caso, obteve-se um modelo de regressão que apresentou R^2 ajustado superior ao modelo que incluiu imóveis usados. Salienta-se que, em ambos os modelos, as mesmas variáveis utilizadas foram estatisticamente significativas nas respectivas regressões.

A avaliação de performance de regressão considerou erros de predição baseados em três métricas: EPAM, EQM e REQM. Os erros foram determinados utilizando validação cruzada com 70% do banco de dados para a regressão propriamente dita e os restantes 30% para testes. A constituição de uma amostra de erros considerou 50 estados aleatórios na divisão do conjunto de dados.

Os resultados da validação cruzada com o modelo de regressão do conjunto de dados de agosto de 2022, com imóveis usados, mostram EPAM variando de 19,8% a 25,6% com 70% dos erros abaixo da média. Os erros obtidos com o modelo de regressão com o conjunto de dados de agosto de 2022, sem imóveis usados, variaram entre 12,8% e 16,7% sendo que aproximadamente 60% da amostra apresentou erro menor que o valor do EPAM. Conclui-se, portanto que o modelo que considera imóveis usados tem uma variação de erro maior do que a do modelo que não inclui imóveis usados.

A aplicação do método CHAID, combinando variáveis contínuas e categóricas do conjunto de dados de agosto de 2022 com imóveis usados conduziu a uma árvore de decisão com 14 nós terminais (ou folhas), os quais configuram o caminho de tomada de decisão, ou

classificação. O método de regressão linear múltipla foi aplicado em cada uma das 14 folhas da árvore, pois considerou-se que os imóveis contidos nos subconjuntos do banco de dados caracterizados pelas folhas apresentam maior grau de homogeneidade, resultando assim em uma série de modelos de regressão múltipla com erros médios de predição menores. O EPAM médio obtido com o modelo de regressão linear múltiplo considerando somente variáveis contínuas foi de 22,7%, enquanto o EPAM médio obtido com a combinação da árvore de decisão e os múltiplos modelos de regressão linear múltipla foi de 14,9%. Portanto pode-se considerar que esta abordagem apresenta maior assertividade na precificação de imóveis.

Os objetivos específicos propostos no trabalho foram alcançados ao aplicar-se os métodos e algoritmos citados. A determinação das variáveis importantes para precificar o modelo tanto no modelo de regressão quanto na árvore de decisão CHAID utilizam de técnicas que determinam as variáveis significativas para o conjunto de dados.

Os resultados do trabalho mostram que a combinação de modelos de regressão com a técnica de árvores de decisão permite considerar simultaneamente variáveis contínuas e categóricas, obtendo-se desta forma, resultados bastante promissores com relação aos erros de predição na precificação de imóveis.

REFERÊNCIAS

FAVERO, Luiz Paulo Lopes. **Modelos de preços hedônicos aplicados a imóveis residenciais em lançamento no município de São Paulo.** 2003. Dissertação (Mestrado em Administração) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2003. doi:10.11606/D.12.2003.tde-05052004-152353. Acesso em: 20 set. 2022.

APTO. Entendendo a participação da construção civil no PIB brasileiro ao longo dos anos. **Blueprint**, 2022. Disponível em: https://blueprint.apto.vc/entendendo-a-participacao-da-construcao-civil-no-pib-brasileiro-ao-longo-dos-anos. Acesso em: 9 set. 2022.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-1**: Avaliação de bens - Parte 1: Procedimentos Gerais**.** Rio de Janeiro: ABNT, 2019.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2**: Avaliação de bens - Parte 2: Imóveis urbanos. Rio de Janeiro: ABNT, 2010.

AZANK, F. Como avaliar seu modelo de regressão: As principais métricas para avaliar seus modelos de regressão. **Medium.** 2020. Disponível em: https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regress%C3%A3o-c2c8d73dab96. Acesso em: 24 set. 2022.

BERTI, S. M.; FOSSATTI, E. C.; MANOSSO, T. W. S.; PEREIRA, A. S. **Regressão Linear Múltipla:** Como simplificar por meio do Excel e SPSS. Passo Fundo, 2019. Disponível em: https://www.upf.br/_uploads/Conteudo/cepeac/textos-discussao/texto-01-2019.pdf. Acesso em: 19 set. 2022.

BRAULE, R. **Estatística Aplicada com Excel**: para cursos de administração e economia. Rio de Janeiro: Campus, 2001.

CHEIN, FLÁVIA. **Introdução aos modelos de regressão linear:** um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas. Brasília: Enap, 2019.

CHOI, S.; YI, M. Y. Computational Valuation Model of Housing Price Using Pseudo Self Comparison Method. **Sustainability (Switzerland)**, *[s. l.]*, v. 13, n. 20, p. 11489, 2021. DOI 10.3390/su132011489. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cbt&AN=EIS153345550&lang=pt-br&site=eds-live&scope=site. Acesso em: 15 set. 2022.

DATAZAP+. Tendências de Moradia: Compra. **Datazap+**, 21 set. 2022. Disponível em: https://www.datazap.com.br/tendencias-de-moradia/. Acesso em: 15 out. 2022.

DOROW, A.; MACEDO JUNIOR, J. S. **Heurística da ancoragem na estimativa de preços de imóveis por corretores profissionais. [dissertação].** *[S. l.: s. n.]*. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat07205a&AN=uls.270113&lang=pt-br&site=eds-live&scope=site. Acesso em: 9 set. 2022.

DOUMPOS, M. et al. Developing automated valuation models for estimating property values: a comparison of global and locally weighted approaches**. ANNALS OF OPERATIONS RESEARCH**, *[s. l.]*, 2020. DOI 10.1007/s10479-020-03556-1. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edswsc&AN=000516934000002&lang=pt-br&site=eds-live&scope=site. Acesso em: 15 set. 2022.

ECCONIT CONSULTORIA ECONÔMICA. **Estudo técnico dedicado à atualização das necessidades habitacionais 2004-2030.** São Paulo, 2020.

FARIAS, DIEGO M.; KONZEN, PEDRO H. A.; SOUZA, RAFAEL R. **Álgebra Linear:** Um Livro Colaborativo. Porto Alegre, 2020.

FERNANDO R. PELLEGRINI; FLÁVIO S. FOGLIATTO. Passos para implantação de sistemas de previsão de demanda: técnicas e estudo de caso. **Production**, *[s. l.]*, v. 11, n. 1, p. 43–64, 2001. DOI 10.1590/S0103-65132001000100004. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsdoj&AN=edsdoj.63924fcc3aac4380bee4da86079175f3&lang=pt-br&site=eds-live&scope=site. Acesso em: 15 out. 2022.

Hoffmann, Rodolfo. **Análise de regressão**: uma introdução à econometria [recurso eletrônico] / Rodolfo Hoffmann. - - 5. ed. Piracicaba: O Autor, 2016. Disponível em: https://www.esalq.usp.br/biblioteca/sites/default/files/Analise_Regress%C3%A3o.pdf. Acesso em: 22 set. 2022.

ISHIJIMA, H. (1); MAEDA, A. (2)**.** Real Estate Pricing Models: Theory, Evidence, and Implementation. **Asia-Pacific Financial Markets**, *[s. l.]*, v. 22, n. 4, p. 369-396–396, 2015. DOI 10.1007/s10690-013-9170-7. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edselc&AN=edselc.2-52.0-84945476668&lang=pt-br&site=eds-live&scope=site. Acesso em: 15 set. 2022.

KASS, G. V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, *[s. l.]*, v. 29, n. 2, p. 119–27, 1980. DOI 10.2307/2986296. Disponível em: https://doi.org/10.2307/2986296. Acesso em 27 nov. 2022.

KUMARI, KAJAL. Implement of decision tree using CHAID. **Analytics Vidhya**. 5 ago. 2022. Disponível em: https://www.analyticsvidhya.com/blog/2021/05/implement-of-decision-tree-using-chaid/#. Acesso em 25 nov. 2022.

LANCASTER, K. J. A New Approach to Consumer Theory. **Journal of Political Economy**, *[s. l.]*, v. 74, n. 2, p. 132-157, 1966. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsjsr&AN=edsjsr.1828835&lang=pt-br&site=eds-live&scope=site. Acesso em: 16 set. 2022.

LOPES, R.D**. Previsão de Autopeças Um Estudo de Caso em Uma Concessionária de Veículos. [dissertação].** *[s. l.]*, 2002. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=edsndl&AN=eds

ndl.OCLC.oai.xtcat.oclc.org.OCLCNo.181043864&lang=pt-br&site=eds-live&scope=site. Acesso em: 9 dez. 2022.

MAIA, A. G. **Econometria: Conceitos e Aplicações.** Editora Saint Paul. São Paulo, 2017.

MALPEZZI, S. Hedonic Pricing Models: A Selective and Applied Review. *In*: O'SULLIVAN, T.; GIBB, K. **Housing Economics and Public Policy**: Essays in Honor of Duncan Maclennan. Oxford: Blackwell Science, 2002. p. 67-89. Disponível em: https://doi.org/10.1002/9780470690680.ch5. Acesso em: 16 set. 2022.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses**: The PRISMA Statement**. Disponível em: www.prisma-statement.org. Acesso em: 28 set. 2022.

NAKAMURA, E. Aprendizagem da máquina: uma introdução ao erro quadrático médio e linhas de regressão. **Freecodecamp.** 2022. Disponível em: https://www.freecodecamp.org/portuguese/news/aprendizagem-da-maquina-uma-introducao-ao-erro-quadratico-medio-e-linhas-de-regressao/#:~:text=Explica%C3%A7%C3%A3o%20geral&text=Em%20estat%C3%ADstica%2C%20o%20erro%20quadr%C3%A1tico,e%20o%20que%20%C3%A9%20estimado. Acesso em: 30 de nov. 2022

OWUSU-ANSAH, A. A Review of Hedonic Pricing Models in Housing Research. **Journal of International Real Estate and Construction Studies**, *[s. l.]*, v. 1, n. 1, p. 19-38, 2011. Disponível em: https://www.researchgate.net/publication/287232776_A_review_of_hedonic_pricing_models_in_housing_research. Acesso em: 16 set. 2022.

PEREIRA, MÔNICA A. T., PEREIRA, PAULO J. **Notas de Aula de Estatística Aplicada à Engenharia.** Disponível em: https://pemd.univasf.edu.br/arquivos/estatistica.pdf. Acesso em: 11 set. 2022.

ROSEN, S. Hedonic prices and implict markets: production differentiation in pure competition. **Journal of Political Economy**, Chicago, v. 82, n. 1, p. 34-55, 1974. Disponível em: https://www.jstor.org/stable/1830899. Acesso em: 20 set. 2022.

SARDÁ, A.; HOCHHEIM,N. **Aplicação de técnicas de análise multivariada em avaliações de imóveis**. *[S. l.: s. n.]*. Disponível em: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat07205a&AN=uls.358909&lang=pt-br&site=eds-live&scope=site. Acesso em: 19 set. 2022.

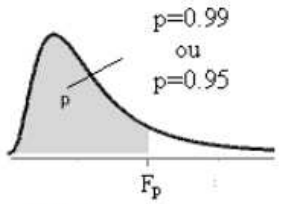
TURE M.; TOKATLI F.; KURT I. Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 AND ID3) in determining recurrence-free survival of breast cancer patients. **Expert System with Applications: An International Journal**, v. 36, n. 2, p. 2017-2026. 2009. Disponível em: http://dl.acm.org/citation.cfm?id=1465032. Acesso em 20 nov. 2022.

VERAS, André Duarte. **Uma Proposta de Utilização de Redes Neurais Recorrentes na Previsão de Preços de Imóveis no Distrito Federal.** 2019. 82 f. Trabalho de Conclusão de

Curso (Bacharelado em Administração) – Universidade de Brasília, Brasília, 2019. Disponível em: https://bdm.unb.br/handle/10483/25643. Acesso em: 17 set. 2022

Anexo A – Tabela F de Fischer-Snedecor

Tabela III: Distribuição F de Fischer-Snedecor



Fornece os quantis $F_{0,95}$ (em cima) e $F_{0,99}$ (em baixo) em função do nº de g.l. numerador ν_1 (coluna) e do nº de g.l. denominador ν_2 (linha)
F tem distribuição F com ν_1 g.l. no numerador e ν_2 g.l. no denominador $P(F \leq F_{0,95}) = 0,95$ e $P(F \leq F_{0,99}) = 0,99$

$\nu_2 \setminus \nu_1$	1	2	3	4	5	6	7	8	9	10	20	40	60	120	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	248,01	251,14	252,20	253,25	254,31
	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47	6055,85	6208,73	6286,78	6313,03	6339,39	6365,76
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,45	19,47	19,48	19,49	19,50
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,45	99,47	99,48	99,49	99,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,59	8,57	8,55	8,53
	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	26,69	26,41	26,32	26,22	26,13
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,72	5,69	5,66	5,63
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,02	13,75	13,65	13,56	13,46
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,46	4,43	4,40	4,37
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,55	9,29	9,20	9,11	9,02
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,77	3,74	3,70	3,67
	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,40	7,14	7,06	6,97	6,88
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,34	3,30	3,27	3,23
	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,16	5,91	5,82	5,74	5,65
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	3,04	3,01	2,97	2,93
	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,36	5,12	5,03	4,95	4,86
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,83	2,79	2,75	2,71
	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,81	4,57	4,48	4,40	4,31
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,66	2,62	2,58	2,54
	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,41	4,17	4,08	4,00	3,91
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	1,99	1,95	1,90	1,84
	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	2,94	2,69	2,61	2,52	2,42
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84	1,69	1,64	1,58	1,51
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,37	2,11	2,02	1,92	1,81
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75	1,59	1,53	1,47	1,39
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,20	1,94	1,84	1,73	1,60
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,66	1,50	1,43	1,35	1,25
	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,03	1,76	1,66	1,53	1,38
∞	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,57	1,39	1,32	1,22	1,02
	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	1,88	1,59	1,47	1,33	1,03

Obs.: O quantil F_p correspondente a ν_1 g.l. no numerador e ν_2 g.l. no denominador coincide com o inverso do quantil F_{1-p} correspondente a ν_2 g.l. no numerador e ν_1 g.l. no denominador.

Anexo B – Exemplificação da classificação da árvore CHAID

No anexo B se encontra a evolução do passo a passo de como é definido cada característica importante na árvore de decisão utilizando do método CHAID.

Após definir a primeira variável da árvore de decisão é feito a análise do qui-quadrado para cada um dos caminhos da árvore de decisão. Abaixo é demonstrado o segundo passo para o caminho da primeira decisão, neste caso seguindo pelo caminho *Sunny* para a variável *Outlook*.

Outlook = Sunny branch

This branch has 5 examples. Presently, we search for the most predominant feature. By The Way, we will disregard the outlook feature now since they are altogether the same. At the end of the day, we will find out the most predominant columns among temperature, humidity, and wind.

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes

Humidity feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
High	0	3	3	1.5	1.225	1.225
Normal	2	0	2	1	1	1

Chi-square value of humidity feature for sunny outlook is

= 1.225 + 1.225 + 1 + 1

= 4.449

Wind feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	1	2	3	1.5	0.408	0.408
Strong	1	1	2	1	0	0

Chi-square value of wind feature for sunny outlook is

= 0.408 + 0.408 + 0 + 0

= 0.816

Temperature feature for when the outlook is Sunny

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Hot	0	2	2	1	1	1
Mild	1	1	2	1	0	0
Cool	1	0	1	0.5	0.707	0.707

So, the chi-square value of temperature feature for sunny outlook is

= 1 + 1 + 0 + 0 + 0.707 + 0.707

= 3.414

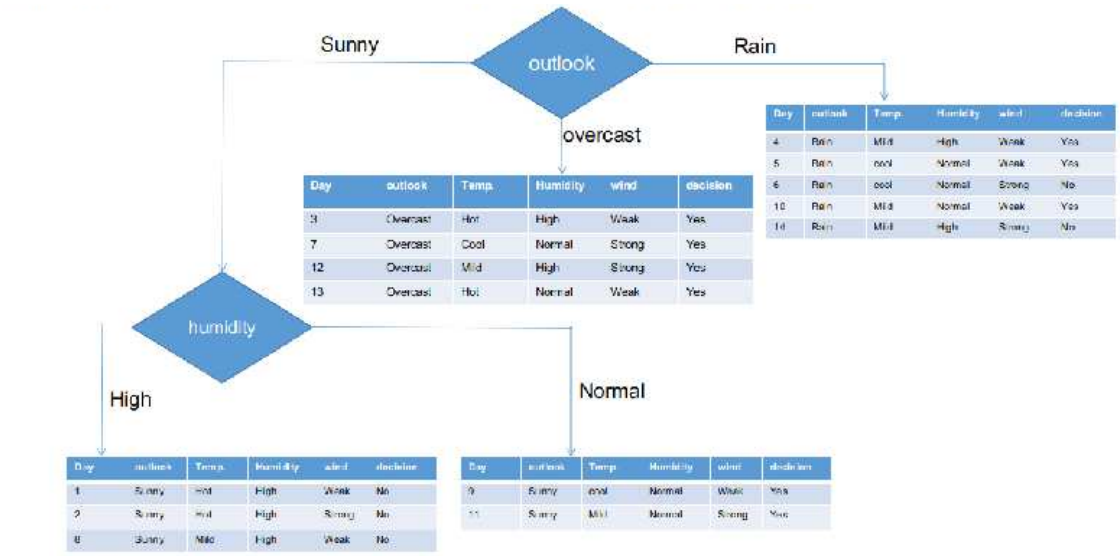
We have found chi-square values for sunny is outlook. Let's see them all at a table.

Feature	Chi-square
Temperature	3.414
Humidity	4.449
Wind	0.816

Presently, humidity is the most predominant feature for the sunny outlook branch. We will put this feature as a decision rule.

No caminho *Sunny,* após realizar os cálculos de qui-quadrado para as variáveis *Temperature, Humidity* e *Wind*, pelo valor do qui-quadrado definiu-se que a próxima decisão a ser realizada é quanto à variável *Humidity*¸ com a decisão se é: *High* ou *Normal.*

No próximo passo é apresentado os cálculos para definir o caminho a se seguir com a primeira decisão sendo o *Outlook = Rain.* Abaixo é apresentado o cálculo do qui-quadrado para cada possível decisão de variável.



Presently, both humidity branches for sunny outlook have only one decision as delineated previously. CHAID tree will return NO for sunny outlook and high humidity and it will return YES for sunny outlook and normal humidity.

Rain outlook branch

This branch actually has both yes and no decisions. We need to apply the chi-square test for this branch to find out an accurate decision. This branch has 5 distinct instances as demonstrated in the accompanying sub informational collection dataset. How about we find out the most predominant feature among temperature, humidity and wind.

Day	Outlook	Temp.	Humidity	Wind	Decision
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
10	Rain	Mild	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Wind feature for rain outlook

There are two types of a class present in wind feature for rain outlook such that weak and strong.

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Weak	3	0	3	1.5	1.225	1.225
Strong	0	2	2	1	1	1

So, the chi-square value of wind feature for rain outlook is

= 1.225 + 1.225 + 1 + 1

= 4.449

Humidity feature for rain outlook

There are two types of a class present in humidity feature for rain outlook such that high and normal.

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
High	1	1	2	1	0	0
Normal	2	1	3	1.5	0.408	0.408

Chi-square value of humidity feature for rain outlook is

= 0 + 0 + 0.408 + 0.408

= 0.816

Temperature feature for rain outlook

There are two types of a class present in temperature features for rain outlook such that mild and cool.

	Yes	No	Total	Expected	Chi-square Yes	Chi-square No
Mild	2	1	3	1.5	0.408	0.408
Cool	1	1	2	1	0	0

Chi-square value of temperature feature for rain outlook is

= 0 + 0 + 0.408 + 0.408

= 0.816

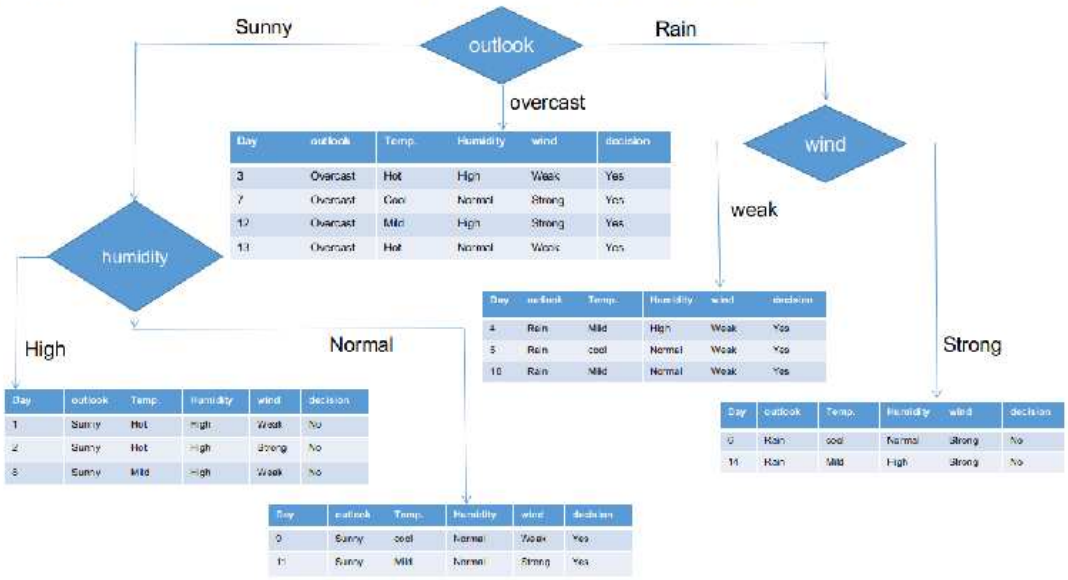
We have found all chi-square values for rain is outlook branch. Let's see them all at a single table.

Feature	Chi-squared
Temperature	0.816
Humidity	0.816
Wind	4.449

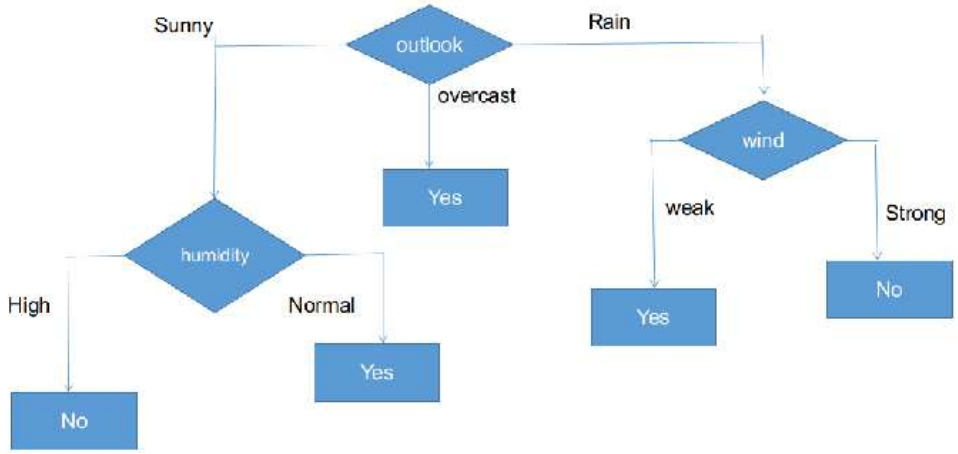
Thus, the wind feature is the victor for the rain is the outlook branch. Put this column in the connected branch and see the corresponding sub informational dataset.

No caminho para *Outlook=Rain*, depois de realizado o cálculo de qui-quadrado para cara variável, a que obteve resultado com mais significado no teste foi a variável *Wind,* que é o próximo caminho a ser seguido. As decisões para a variável *Wind*, neste caminho é: *Weak* ou *Strong.*

A variável *Outlook=Overcast* sempre retorna a decisão *yes* para o exemplo de comprar leite. Por este motivo não é calculado o valor de qui-quadrado.



As seen, all branches have sub informational datasets having a single decision such that yes or no. In this way, we can generate the CHAID tree as illustrated below.



A árvore acima explicita o caminho de decisão para *Yes* ou *No* para comprar leite. Esta árvore, juntamente com a sua lógica esta apresentada na seção 3.3.