UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

GUILHERME HENRIQUE DE CAMARGO

CLASSIFICAÇÃO DE PREÇOS DE IMÓVEIS UTILIZANDO A TÉCNICA DE FLORESTA ALEATÓRIA

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA
2020

GUILHERME HENRIQUE DE CAMARGO

CLASSIFICAÇÃO DE PREÇOS DE IMÓVEIS UTILIZANDO A TÉCNICA DE FLORESTA ALEATÓRIA

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do título de Bacharel em Ciência da Computação, do Departamento Acadêmico de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. André Koscianski

PONTA GROSSA
2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Ponta Grossa
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Bacharelado em Ciência da Computação

TERMO DE APROVAÇÃO

CLASSIFICAÇÃO DE PREÇOS DE IMÓVEIS
UTILIZANDO A TÉCNICA DE FLORESTA ALEATÓRIA

por

GUILHERME HENRIQUE DE CAMARGO

Este Trabalho de Conclusão de Curso (TCC) foi apresentado em 26 de outubro de 2020 como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

_____
Prof. Dr. André Koscianski
Orientador

_____
Prof. Dr. André Pinz Borges
Membro titular

_____
Prof. Dr. Richard Duarte Ribeiro
Membro titular

_____
Prof. MSc. Geraldo Ranthum
Responsável pelo Trabalho de Conclusão de Curso

_____
Profa. Dra. Mauren Louise Sguario
Coordenador do curso

“Sem dados você é apenas mais
uma pessoa com uma opinião.”
(DEMING, William Edwards, 1945)

RESUMO

CAMARGO, Guilherme Henrique de. **Classificação de preços de imóveis utilizando a técnica de floresta aleatória**. 2020. 44p.Trabalho de Conclusão de Curso Bacharelado em Ciência da Computação - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2020.

A precificação de imóveis é relevante para diversos assuntos, como investimento imobiliário, simulação e até no planejamento urbano. Sendo assim, investidores, compradores e vendedores de imóveis são pessoas potencialmente interessadas. Contudo a obtenção de informações a partir de bases eletrônicas não é trivial. Com isso processos para a descoberta de conhecimento foram desenvolvidos, como é o caso do Knowledge Discovery in Databases (KDD). Este processo possui, dentre suas etapas, as etapas de Mineração de Dados (MD) e avaliação do modelo gerado. Para realizar a mineração é utilizado um algoritmo, que nesta aplicação em específico, é a Random Forest (RF). Para a avaliação, um método deve ser empregado sobre a descoberta gerada, como a matriz de confusão. Por sua vez a previsão de preços pode ser um objetivo do KDD. Comparações podem ser realizadas utilizando o modelo gerado pelo trabalho. Neste trabalho a modelagem foi realizada com um conjunto de dados geográficos, a classificação final sendo um intervalo de preço.

Palavras-chave: Mineração de Dados. Floresta Aleatória. Descoberta de Conhecimento em Bancos de Dados.

ABSTRACT

CAMARGO, Guilherme Henrique de. **Real estate price classification using the random forest technique**. 2020. 44p. Work of Conclusion Course Graduation in Bachelor of Science in Computer Science – Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2020.

Real estate pricing is relevant to several issues, such as real estate investment, simulation and even urban planning. Therefore, investors, buyers and sellers of real estate are potentially interested people. Information gathering from electronic databases is nontrivial; this motivates the development of Knowledge Discovery in Databases (KDD). This process has, among its steps, the Data Mining (MD) and evaluation steps. To carry out mining, some algorithms are used, in this specific application, the Random Forest (RF). For the evaluation, a method must be used on the generated discovery, such as the confusion matrix. Price forecasting, in turn, can be a goal of KDD. Comparisons can be made using the model generated by the work. In this work, the modeling was performed with a set of geographic data, the final classification being a price range.

Keywords: Data Mining. Random Forest. Knowledge Discovery in Databases.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

RF	*Random Forest* (Floresta Aleatória)
KDD	*Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados)
TB	Terabyte
IA	Inteligência Artificial
RNA	Rede Neural Artificial
SE	Sistema Especialista
AG	Algoritmo Genético
AD	Árvore de Decisão
HTML	*HyperText Markup Language* (Linguagem de Marcação de Hipertexto)
OSM	OpenStreetMap (Mapa Aberto de Ruas)
BD	Banco de Dados
MD	Mineração de Dados

SUMÁRIO

1 INTRODUÇÃO

No mercado imobiliário a previsão de preços é interessante para compradores e vendedores (YU; WU, 2016), pois auxilia na melhoria de decisões comerciais dado o fato de que o valor de uma região ou imóvel pode vir a ser valorizado ou desvalorizado, em função de fatores como decisões políticas (BERTONCELLO *et al.*, 2019) ou índice de criminalidade contra patrimônio (SANTOS, 2018). Previsões de preço também são usadas no mercado de investimentos (MORO, 2017), para garantir que todo o dinheiro investido seja utilizado da melhor maneira possível e tenha um retorno mais provável. Estudos referentes a mercado imobiliário estão presentes também na área de simulações urbanas (ROTH, 2019; KAMUSOKO; GAMBA, 2015; SHAFIZADEH-MOGHADAM *et al.*, 2017).

O aumento da utilização dos meios virtuais para armazenamento, nos últimos vinte anos, atingiu números expressivos nas áreas comercial e científica (GOLDSCHMIDT; PASSOS, 2005). Os tamanhos ultrapassam os terabytes (TB) e existem grandes bases de dados referentes à cidades e mercado imobiliário. A análise realizada por um ser humano seria lenta, custosa e mesmo impossível ao explorar um volume tão grande de dados (HAN; PEI; KAMBER, 2011).

Dessa forma, recorrer a algoritmos e/ou métodos de processamento para que informações úteis sejam extraídas é essencial. Algumas técnicas incluem regressão gaussiana (CROSBY *et al.*, 2016) e modelos estatísticos (LI, 2016) ou Inteligência Artificial (IA), utilizada neste trabalho. Em específico, um processo de extração de conhecimento, denominado *Knowledge Discovery in Databases* (KDD), no qual foi aplicado o algoritmo *Random Forest* (RF) para precificação de imóveis.

Embora exista uma enorme quantidade de dados disponíveis, o que se tem na verdade é uma escassez de informações (HAN; PEI; KAMBER, 2011). Os dados são representações computacionais de modelos e atributos que são referentes a entidades reais ou simuladas, enquanto as informações são resultados de um processo computacional, como análises estatísticas, que dão significados aos dados ou a transcrição de alguns significados atribuídos por seres humanos (CHEN *et al.*, 2008).

Para realizar a extração de informações pode-se utilizar algoritmos de mineração de dados presentes no processo de KDD, que baseando-se em seus resultados pode-se realizar análises dos dados, detectar irregularidades e extrair modelos, relações e estruturas importantes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

É interessante que a etapa de mineração de dados utilize grandes volumes de dados para que ajuste seus algoritmos e faça com que as conclusões apresentadas sejam as mais precisas possíveis (GOLDSCHMIDT; PASSOS, 2005). Porém, como na modelagem não se tem certeza de nada, alguns modelos com uma quantidade menor de dados ainda pode ser satisfatório. Para a predição de preço, este trabalho utilizou o algoritmo chamado RF, por trazer uma aplicação pouco utilizada que por sua vez se tornar um modelo de comparação para trabalhos futuros de

outros autores, mas principalmente porque a RF é citada na literatura e sua aplicação depende de fatores como dados disponíveis e decisões de tratamento dos mesmos. Assim, o trabalho teve a intenção de investigar o uso da técnica em uma base de dados disponível, verificando dificuldades de trabalhar, taxas de acerto, possíveis desvantagens, dentre outros. E com isso, poderiam ser testados outros métodos e feitas novas comparações.

A partir do estudo realizado, a previsão do preço do mercado imobiliário poderá ser combinada a outras ferramentas, bem como uma forma de auxílio no planejamento urbano, mercado imobiliário e/ou de investimentos imobiliários. Neste trabalho os dados, previamente padronizados pelo trabalho de Roberto (2019) e devidamente tratados, foram inseridos na RF, onde passaram pelo processo do KDD e foram analisados gerando um modelo preditivo, cuja precisão deste modelo foi inferida mediante um método de avaliação. O método de avaliação utilizado foi o de matriz de confusão, responsável por aferir os acertos e erros da classificação. O método será explicado com maiores detalhes no item 3.8.2.

Dessa forma, este trabalho faz parte de um conjunto de pesquisas, do qual o trabalho de Roberto (2019) pertence e por sua vez é usado como base para este trabalho, que posteriormente será utilizado como base para trabalhos futuros.

1.1 OBJETIVOS

Nesta seção serão apresentados o objetivo geral e os específicos.

1.1.1 Objetivo Geral

O presente trabalho tem como objetivo geral a classificação de preços de imóveis por meio de um algoritmo de mineração de dados denominado *Random Forest*.

1.1.2 Objetivos Específicos

Definem-se como objetivos específicos deste trabalho:

- Realizar um levantamento bibliográfico de trabalhos similares para estudar como outros e até o mesmo algoritmo é utilizado em problemas parecidos ou iguais;
- Encontrar os parâmetros que melhor se adaptam ao conjunto de dados gerando modelos e comparando-os;
- Analisar o banco de dados de imóveis obtido;

- Aplicar os métodos e algoritmos selecionados;
- Definir um critério de análise;
- Analisar os dados.

1.2 JUSTIFICATIVA

A previsão de preços é relevante por ser a base de investimentos, nas questões de compra e venda de imóveis. Sendo também um auxiliador do planejamento urbano que trás como consequência o decrescimento dos índices de criminalidade (CHIODI, 2016). A RF é também realizada em diferentes situações, como: sensoriamento remoto (BELGIU; DRĂGUŢ, 2016) ou sistema de detecção de intrusão de rede (FARNAAZ; JABBAR, 2016).

Os resultados obtidos poderão ser aproveitados em trabalhos de simulação e projeção de crescimento urbano. O trabalho por sua vez teve a intenção de investigar o uso da técnica em uma base de dados disponível, verificando dificuldades de trabalhar, taxas de acerto, possíveis desvantagens, dentre outras coisas. E a partir do trabalho, outros métodos podem ser testados e realizadas novas comparações.

1.3 ORGANIZAÇÃO DO TRABALHO

A organização deste trabalho se dá sobre os assuntos: precificação de imóveis e descoberta de conhecimento em banco de dados. Sendo composto por seis capítulos.

No Capítulo 2, o mercado imobiliário é explorado, no qual um contexto sobre a precificação de imóveis é descrito, assuntos como qual são as formas de calcular o preço de um imóvel, como os valores são obtidos, como os riscos são avaliados, apresentando equações matemáticas, alguns modelos existentes para determinar o valor de um imóvel e por fim o que outros autores acham sobre a precificação e os tipos de precificações.

No Capítulo 3, o processo de descoberta de conhecimento em banco de dados é estudado informando suas etapas e como cada uma é feita, trabalhos relacionados e as considerações finais.

No Capítulo 4, todos os processos realizados no desenvolvimento do trabalho são apresentados, cada um com sua particularidade e por fim as considerações finais do capítulo.

No Capítulo 5, os resultados da modelagem são apresentados e brevemente discutidos.

No Capítulo 6, a conclusão do trabalho é apresentada, discutindo os objetivos e se os mesmos foram atingidos.

2 MERCADO IMOBILIÁRIO

Estudar os preços de imóveis abrange a exploração dos riscos na predição imobiliária, dos cálculos utilizados na precificação, de como os preços são obtidos, dos modelos para predição de preços e da relevância dos preços com os atributos relevantes. O sucesso que esse tema vem ocasionando nos últimos quarenta anos é incontestável, principalmente ao se observar os progressos técnicos aplicados em todas as áreas adjacentes a economia financeira (WHEATON *et al.*, 2001). Há riscos envolvidos na predição que serão apresentados neste trabalho, cujo interesse é prever com a maior precisão possível o valor monetário futuro de propriedades por meio de valores anteriores.

Para calcular o preço de um imóvel vários fatores estão envolvidos. No ponto de vista financeiro o preço é referente ao rendimento do imóvel e sua rentabilidade total desejada (LAIA, 2007). Outro fatores como: as características dos imóveis, a qualidade de vida nas proximidades, a acessibilidade e as especificações de uso e ocupação do solo também influenciam no valor do imóvel (GOMES; MACIEL; KUWAHARA, 2012). Além desses, o bairro onde o imóvel se encontra é utilizado em trabalhos para determinar o preço (FURTADO, 2011). Outra forma de precificação é a feita por um corretor de imóveis profissional, na qual o profissional pode controlar o preço com o intuito de desenvolver o processo de venda para interesse próprio, já que as decisões em investimentos são complexas e com cargas emocionais elevadas (DOROW, 2012). Além desses, modelo com preços médios mensais de imóveis utilizando redes neurais também é empregado (VERAS, 2019). Bem como técnicas de agrupamento (MA *et al.*, 2020).

A obtenção do preço pode ser feita por regressão linear, um método estatístico que relaciona uma variável alvo com outras variáveis (PEREIRA; GARSON; ARAÚJO, 2012; BRASIL, 2019). Utilizando estimação através de modelos baseados em dados, os modelos baseados em aprendizagem de máquina (MALERE; ALMEIDA; SANO, 2019). Ou também, coletar os preços de imóveis colocados a venda na internet (AMARAL, 2018).

Existem diversos modelos para determinação de preço, como o *cap rate* ou *yield*, que são modelos estimativos, no qual sabendo do rendimento de um imóvel e de sua rentabilidade total desejada utiliza-se uma das fórmulas para fazer a estimativa do preço (LAIA, 2007). A aplicação da fórmula *cap rate* para um determinado imóvel depende, por sua vez, da verificação prévia da *cap rate* para imóveis semelhantes, com as mesmas características (LAIA, 2007). Outro modelo existente é o *Fair Market Value* (FMV), onde uma equação é dada para formar o preço de um imóvel (DOROW, 2012):

$$V = (S \times \overline{P}) + C + (F_1 + F_2 + ... + F_n) \tag{2.1}$$

Dos quais V é a avaliação de valor, S é o tamanho da residência usando como cálculo o preço por metro quadrado, P é a média de preço das propriedades vizinhas, C é o estado da casa e F sendo características significantes em comparação as propriedades vizinhas (DOROW, 2012).

O principal risco, e mais significante, é o futuro indeterminado, pois com ele definimos os riscos posteriores. Economistas costumam debater este assunto (BODIE; KANE; MARCUS, 2014), porém, independente do debate entre teóricos da economia a atenção maior é voltada a aferir os riscos, não defini-los. Assim, o resultado da previsão se torna mais confiável. Modelos de previsão por sua vez captam os itens temporais, que influenciam na certeza da previsão. Essa influência acontece por conta de um item interferir no outro. Uma previsão mais confiável torna o resultado mais provável e por sua vez mais correto (WHEATON *et al.*, 1999).

O cálculo do risco pode ser feito por um modelo, como a Teoria da Utilidade Esperada (TUE), que busca descrever um modelo de tomada de decisão sobre o risco, que resumidamente afirma que a determinação do valor de um item deve ser baseada na utilidade que esse item traz (DOROW, 2012). Outros aspectos como a região do imóvel e sua futura valorização são importantes na decisão de investimentos (VERAS, 2019).

Trabalhos existentes de previsão de preços utilizam métodos de regressão para encontrar valores concretos, porém, no mundo real, a faixa de preço pode ser mais prática (MA *et al.*, 2020). O processo de estimar preço é subjetivo, já que inclui saber profundamente sobre o comportamento do mercado em diferentes faixas de preço, locais e tipos de imóveis (VERAS, 2019).

2.1 CONSIDERAÇÕES FINAIS

Neste capítulo apresentou-se um contexto da precificação de imóveis, como são calculados os preços, como são obtidos, modelos existentes para determiná-los, cálculos dos riscos e como outros autores entendem o processo de precificação.

Com essas informações verifica-se que existem alguns atributos essenciais e que é possível se basear em preços anteriores para estimar valores futuro. Assim como utilizar uma faixa de preço é mais prático e estimar preço é um processo subjetivo. Ressaltando que riscos também estão presentes em processos de precificação como esses e identificá-los pode ser interessante. Por fim, entende-se que todo esse domínio auxiliará no processo de KDD.

3 DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS

A Descoberta de Conhecimento em Banco de Dados, do inglês *Knowledge Discovery in Database* (KDD), nada mais é do que um processo. Seu objetivo é realizar uma ou várias identificações de padrões ocultos. Um padrão pode ser a relevância entre dois atributos. Padrões previamente desconhecidos são interessantes para o domínio dos dados. O principal objetivo é que ao final das etapas do KDD esses padrões sejam inteligíveis à pessoa que utilizar este processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MARKUSOSKI *et al.*, 2019).

A principal etapa deste processo é a Mineração de Dados (MD), na qual a escolha de um algoritmo de mineração é realizada. Posteriormente essa escolha resultará na concepção de um modelo que produzirá padrões anteriormente ocultos. O modelo gerado terá como propósito a melhor absorção dos fatos referentes ao conjunto de dados, a análise do domínio e a previsão de valor (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MARKUSOSKI *et al.*, 2019).

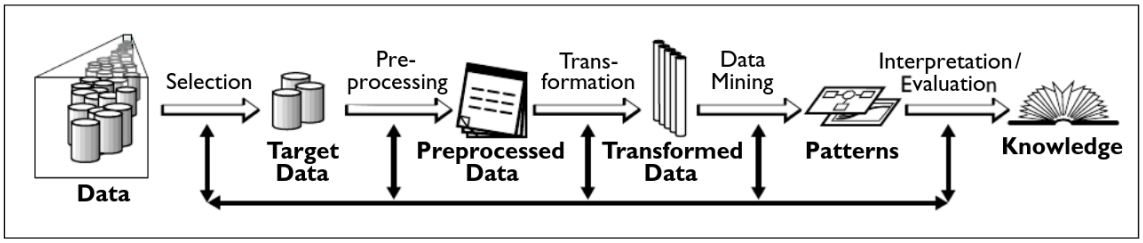
O KDD, representado pela Figura 1, é um processo iterativo e interativo, que pode ser resumido em nove etapas que serão descritas em detalhes nos capítulos seguintes. As etapas são iterativas, ou seja, é possível voltar a qualquer etapa em qualquer momento da aplicação do processo, porém, sendo necessária a aplicação das etapas sucessoras. A aplicação das etapas posteriores à etapa que se retornou pode ser realizada ou mantida, dependendo da conclusão do responsável pelo processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; MARKUSOSKI *et al.*, 2019).

A existência de um modelo pré-definido para um conjunto de dados particular é inexistente, fazendo assim a aplicação e reaplicação do processo ser realizada diversas vezes. Cada nova rodada trará um resultado que será definido como bom ou ruim pelo responsável pelo domínio do processo (MARKUSOSKI *et al.*, 2019).

Pela inexistência de um modelo pré-definido, ou seja, por não dispor de fórmulas ou regras exatas para cada uma das etapas, o processo é intuitivo com base na análise das respostas em cada modelo realizado. Dominar o processo é essencial, pois quando se sabe o que se possui, na questão de dados, e em qual objetivo se quer chegar, a realização de um modelo ideal é mais provável (MARKUSOSKI *et al.*, 2019).

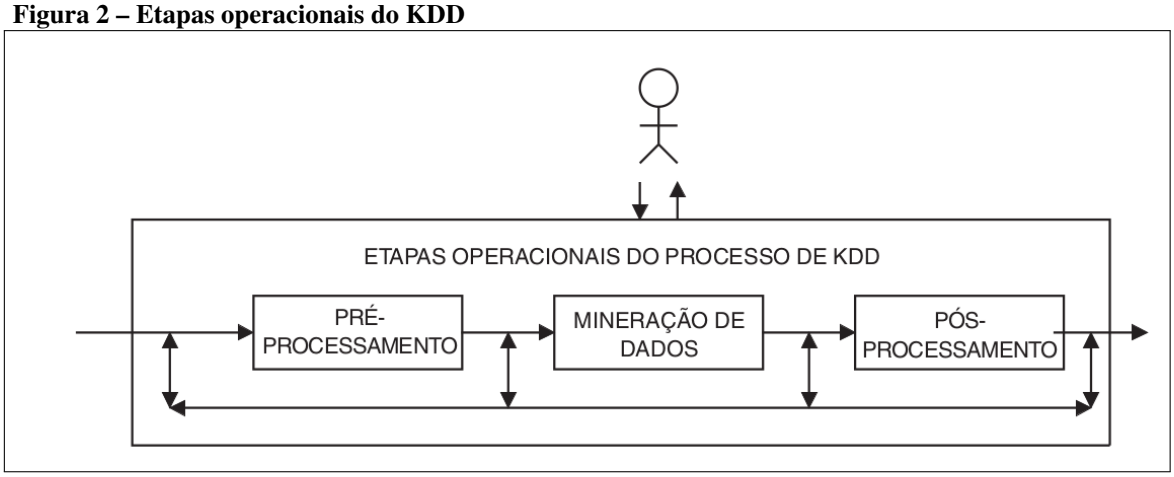
O KDD possui algumas abstrações, como apresentado na Figura 2, realçando a questão

Figura 1 – Passos do processo de KDD



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

Figura 2 – Etapas operacionais do KDD



Fonte: Goldschmidt e Passos (2005)

de não haverem padrões específicos para a busca de conhecimento, levando à abstrações dos processos (GOLDSCHMIDT; PASSOS, 2005).

Como pode ser observado na Figura 2, alguns dos passos existentes no KDD são reunidos em outros passos fazendo com que o processo padrão do KDD seja realizado em etapas menores.

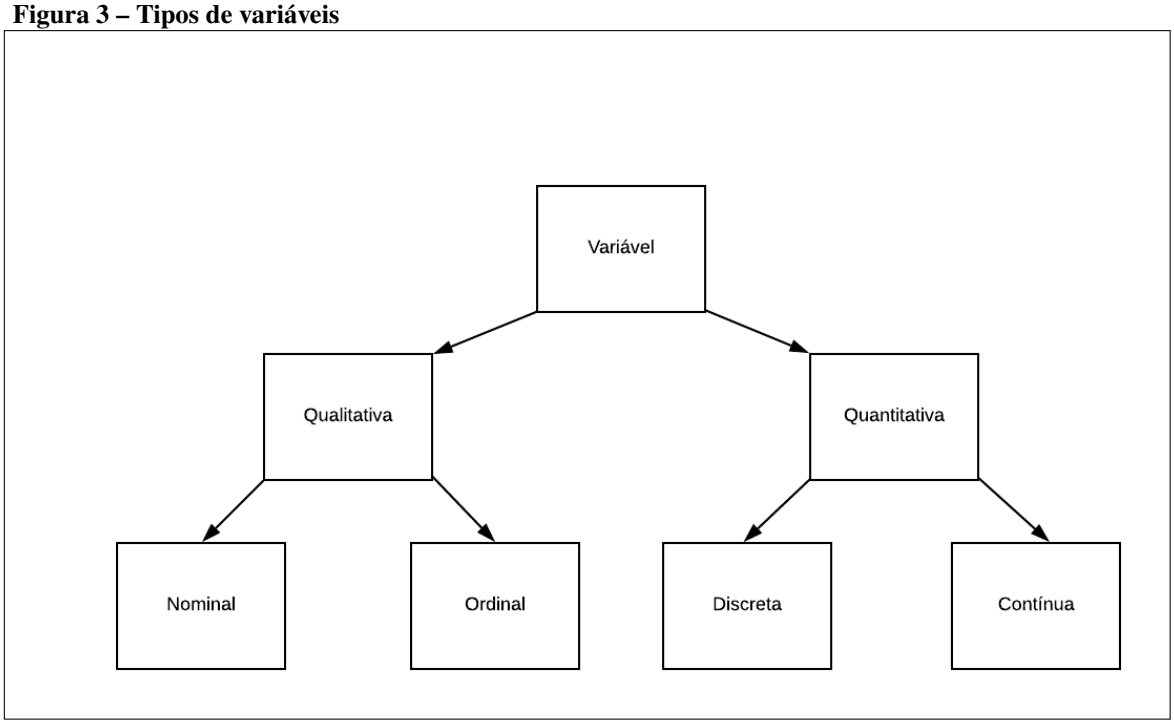
3.1 TIPOS DE DADOS E VARIÁVEIS

Os dados possuem classificações dependentes de seus aspectos ou da natureza da informação. O primeiro aspecto é referente à representação dos valores, denominado tipo de dado. Tal aspecto indica quais as maneiras de informar como as variáveis, ou atributos, do conjunto ao qual se está manipulando estão armazenadas. O segundo aspecto é referente à natureza da informação, também chamado de tipo de variável. Ele é representado por variáveis quantitativas, subdividas em discretas ou contínuas, e qualitativas, subdivididas em nominais ou ordinais (GOLDSCHMIDT; PASSOS, 2005), apresentadas visualmente na Figura 3.

As variáveis então, podem ser classificadas em (PYLE, 1999):

- Nominais – São variáveis que não podem ser mensuradas quantitativamente, mas sim de forma categórica. Não existe ordenação, apenas uma rotulação dos dados. Alguns exemplos deste tipo de dado ocorrem nas classificações de: sexo, cor dos olhos, dependente ou não, dentre outros;
- Ordinais – São variáveis categóricas que possuem uma ordem, porém não é possível quantificar a diferença entre elas. Alguns exemplos são níveis de escolaridade, classes sociais, dentre outros;
- Discretas – São variáveis que representam um conjunto inteiro finito ou enumerável de valores, que representam uma contagem. Alguns exemplos são número de falhas em equi-

Figura 3 – Tipos de variáveis



Fonte: Autoria Própria

pamentos, quantidade de clientes de um serviço, dentre outros;

- Contínuas – São variáveis que possuem valores reais expressos por intervalos abertos ou fechados. Exemplos de variáveis contínuas são: peso, altura, massa, dentre outros.

3.2 AQUISIÇÃO DOS DADOS

Os dados deste trabalho foram obtidos por meio da extração de anúncios de venda de imóveis em *sites* de imobiliárias em uma cidade, no Paraná, a aproximados 115 quilômetros de distância de Curitiba denominada Ponta Grossa representada pela Figura 4.

A aquisição dos dados é um assunto que geralmente está incluso no processo do KDD, mas neste trabalho foi tratado separadamente por outro trabalho Roberto (2019).

Para tal extração foram utilizados três *sites* imobiliários, Conceito Imóveis, Procure Imóvel e Tavarnaro, por serem os lugares em que existe uma maior quantidade de anúncios. Todos os anúncios de venda foram selecionados um a um e, utilizando um seletor. Retiraram-se as informações, ou características, do imóvel de dentro do código *HyperText Markup Language* (HTML). Com o resultado dessa extração tais informações foram utilizadas para procurar no mapeamento colaborativo denominado *OpenStreetMap* (OSM). Posteriormente essas informações passaram por tratamentos, descritos em Roberto (2019), que disponibilizaram 65.909 registros, no formato *Tab-Separated Values* (TSV). Este conjunto possuindo 14 atributos, apresentados nos próximos itens. Alguns registros são apresentados pela Figura 5.

Figura 4 – Visão da cidade de Ponta Grossa

Fonte: OpenStreetMaps (2020)

3.3 DOMÍNIO DE APLICAÇÃO

O passo de preparação do processo de KDD acontece quando o domínio da aplicação é entendido e desenvolvido, isso é necessário para que se compreenda os dados sendo utilizados (MARKUSOSKI *et al.*, 2019).

Este é o responsável por auxiliar o compreendimento do cenário que dará o curso das medidas a serem tomadas nos passos posteriores, assim como a seleção dos dados que são definidos como essenciais, quais pré-processamentos e transformações serão realizados nos dados dependendo do algoritmo ou algoritmos selecionados para o passo de mineração de dados e posteriormente quais métodos de avaliação serão realizados (MARKUSOSKI *et al.*, 2019).

Os responsáveis pelo processo de KDD estarão incubidos de entender e determinar as finalidades do usuário final, bem como o ambiente onde todo o processo acontecerá. Geralmente os responsáveis são pessoas que dominam o conhecimento sobre o cenário dos dados (MARKUSOSKI *et al.*, 2019).

O KDD, por ser um processo interativo e iterativo, tem como adjunta a possibilidade

Figura 5 – Registros com os atributos principais

data	valor	latitude	longitude	bairro
09/02/2019	1700000	-251.004	-501.582	Centro
09/02/2019	165000	-250.925	-501.608	Centro
09/02/2019	950000	-250.833	-501.605	Jardim Carvalho
09/02/2019	550000	-251.119	-501.618	Oficinas
09/02/2019	175000	-250.907	-501.414	Uvaranas
09/02/2019	5250000	-250.655	-501.592	Órfãs
09/02/2019	185798.4	-250.974	-501.539	Centro
09/02/2019	4000000	-25.119	-501.748	Estrela
09/02/2019	1100000	-250.807	-501.831	Nova Rússia
09/02/2019	394000	-251.007	-501.775	Ronda
09/02/2019	365000	-251.108	-501.715	Estrela
09/02/2019	190000	-250.869	-501.502	Centro
09/02/2019	590000	-250.794	-501.574	Jardim Carvalho
09/02/2019	345000	-250.829	-501.573	Jardim Carvalho
09/02/2019	330000	-250.594	-501.675	Centro
09/02/2019	325000	-250.891	-501.502	Uvaranas
09/02/2019	950100	-251.119	-501.618	Oficinas
09/02/2019	1340000	-250.816	-501.677	Órfãs
09/02/2019	160000	-250.934	-501.616	Centro
09/02/2019	1800000	-251.121	-501.509	Oficinas
09/02/2019	200000	-251.046	-501.257	Uvaranas
09/02/2019	550000	-250.974	-501.601	Centro
09/02/2019	1150000	-250.818	-501.775	Nova Rússia
09/02/2019	578000	-250.789	-501.544	Jardim Carvalho
09/02/2019	2080000	-250.928	-50.167	Centro
09/02/2019	590000	-251.121	-501.509	Oficinas
09/02/2019	384000	-250.962	-501.636	Centro
09/02/2019	199000	-250.979	-501.732	Ronda
09/02/2019	1100000	-251.119	-501.618	Oficinas

Fonte: Autoria Própria

de retornar a esta etapa e rever todas as conclusões (MARKUSOSKI *et al.*, 2019).

Neste trabalho o cenário foi estudado e demonstrado no capítulo 2. A finalidade deste processo é precificar os imóveis. O ambiente onde o processo ocorrerá, um *software*, é chamado de *RapidMiner*. Este programa de ciência de dados fornece diversas funcionalidades, como: preparação de dados, aprendizado de máquina, aprendizado profundo, mineração de texto e análise preditiva. Portanto, o *software* foi totalmente preparado para o conjunto de dados selecionados juntamente com o planejamento de todo o processo de KDD.

3.4 SELEÇÃO DOS DADOS

O passo de seleção e criação do conjunto de dados, do KDD, acontece após identificados os propósitos e afins já apresentados no item anterior. Os dados que serão utilizados deverão ser selecionados neste passo. Para que a seleção seja finalizada é necessário que os dados sejam analisados. Se determinado que o conjunto necessita de mais dados, ou seja, é necessário juntar dados de outros locais, isso será resolvido neste estágio. Além dos dados a serem considerados, os atributos vistos como relevantes também são selecionados neste processo (MARKUSOSKI *et al.*, 2019).

Esta etapa tem um valor muito relevante para a mineração de dados, pois um determinado conjunto de dados escolhido pode ajudar ou atrapalhar todo o KDD, por ser a fonte de toda a estruturação dos modelos (MARKUSOSKI *et al.*, 2019).

Os dados podem ser unidos por junção direta, na qual todos são incluídos da mesma forma que são adquiridos, ou seja, sem uma análise prévia para verificar como os atributos e dados serão úteis no processo. Outra união que pode ser realizada é a por junção orientada, em que os dados e seus atributos são selecionados em conjunto com um especialista em KDD ou especialista do domínio (GOLDSCHMIDT; PASSOS, 2005).

Uma das tarefas da seleção é a redução de dados verticais, na qual a escolha de uma das combinações disponíveis, deve ser realizada de forma que o menor número de atributos seja utilizado e a maior consistência seja persistida. O principal dilema envolvido nesta escolha é que quanto maior for n, maior será o número de combinações possíveis. Para conseguir uma redução há a possibilidade de realizar a eliminação direta de atributos, que tem como requisito único e essencial que a pessoa que for realizar tal eliminação tenha o maior conhecimento possível sobre o domínio (GOLDSCHMIDT; PASSOS, 2005).

A eliminação em questão, por sua vez, possui duas heurísticas que podem ser utilizadas para a execução da mesma, uma de remoção de atributos com valores constantes, pois valores que não se alteram não apresentam utilidade para o KDD; e a outra de remoção de atributos que sejam identificadores, como código de um determinado item, que não são úteis para o processo por buscar-se a generalização e não a especificação dos dados (GOLDSCHMIDT; PASSOS,

2005).

3.5 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento é a parte do processo de KDD em que os dados passam por adaptações e aprimoramentos. Os dados nesta etapa passam por tratamento de falhas, como preenchimento de informações ausentes, remoções de ruídos ou dados atípicos, denominados *outliers* (MARKUSOSKI *et al.*, 2019).

As opções realizadas nesta etapa vão de métodos brutos, em que a própria percepção da pessoa que está realizando o modelo é levada em conta, até métodos estatísticos ou a aplicação de algoritmos de MD para o preenchimento dos mesmos. As decisões do que utilizar e como utilizar sempre ficarão a cargo da pessoa que estiver modelando todo o processo de KDD, portanto, ficará a ela a opção de qual metodologia utilizar (MARKUSOSKI *et al.*, 2019).

3.5.1 Limpeza dos Dados

O passo de limpeza resume-se ao tratamento de todas as informações que possam prejudicar a modelagem de alguma forma, por exemplo, baixa consistência, erros, ruídos, valores desconhecidos e afins (GOLDSCHMIDT; PASSOS, 2005).

Sua realização se justifica no aperfeiçoamento do conjunto de dados para que não ocorra o denominado *GIGO* (*Garbage in, Garbage out*) e para que a execução dos algoritmos de MD não seja lesada. Ressalta-se que a atuação de especialistas é uma ótima escolha. Um exemplo simples na limpeza dos dados é a definição dos valores mínimos e máximos do atributo, fazendo com que todos os dados que possuírem valores fora deste intervalo estabelecido sejam descartados pelo fato de serem *outliers* (GOLDSCHMIDT; PASSOS, 2005).

A eliminação de dados com informações ausentes é uma das subtarefas do processo de limpeza de dados, em que valores faltantes podem existir, e caso existam, os dados que não possuírem os atributos considerados relevantes são tratados com um dos meios possíveis para que não hajam problemas futuros (GOLDSCHMIDT; PASSOS, 2005).

Alguns dos meios utilizados podem ser desde preenchimento manual, que requer muito tempo e recurso, até os preenchimentos pela utilização de estatística ou de métodos de MD (GOLDSCHMIDT; PASSOS, 2005).

3.6 TRANSFORMAÇÃO DOS DADOS

A etapa de transformação de dados serve para adaptar os dados originais de maneira que possam ser manipulados dentro da MD. Por exemplo, um intervalo contínuo de valores pode ser particionado para definir categorias, necessárias para construir uma árvore de decisão. Uma vez que esse tipo de modificação altera os resultados, pode ser necessário testar diferentes transformações de dados conjuntas buscando melhorar as conclusões do processo de mineração. Esta seção mostra diferentes possibilidades de transformação de dados, explicando porque elas são necessárias e como podem ser aplicadas (MARKUSOSKI *et al.*, 2019).

3.6.1 Agregação

A agregação é responsável por juntar dados ou até mesmo atributos afim de diminuir o conjunto e manter uma qualidade aceitável das informações, ou algumas vezes perder alguns detalhes que as vezes podem nem alterar a qualidade dos resultados dados pelo modelo (GOLDSCHMIDT; PASSOS, 2005).

3.6.2 Normalização

A normalização é uma forma de representação que ajusta proporcionalmente os valores de um atributo em um intervalo. Por exemplo, um atributo que tem valores de 0 até 10.000 pode ser ajustado em um intervalo de -1 à 1 (GOLDSCHMIDT; PASSOS, 2005).

Existem diversos métodos para realizar a normalização, os quatro principais são (GOLDSCHMIDT; PASSOS, 2005):

- *Z-Transformation*, também chamado de normalização estática, em que subtrai-se a média dos dados de todos os valores e os divide pelo desvio padrão. A distribuição então é finalizada com uma média igual a zero e uma variação igual a um. Sendo a equação dada por:

$$x' = (x - \mu)/\sigma \tag{3.1}$$

  No qual x' é o novo valor, x é o valor atual, μ é a média e σ é o desvio padrão.

- *Range Transformation* normaliza todos os valores em um intervalo especificado pelo modelador, e, posteriormente os dados são ajustados a este intervalo. Sua equação se dá da seguinte forma:

$$x' = (x - x_{min})/(x_{max} - x_{min}) \tag{3.2}$$

Dos quais x' é o novo valor, x é o valor atual, x_{min} é o menor valor do intervalo e x_{max} é o maior valor do intervalo.

- *Proportion Transformation* realiza a normalização de maneira que cada valor de um atributo é dividido pela soma total de valores desse mesmo atributo. Assim, expressa-se na seguinte equação:

$$x \in X, \frac{x}{\sum_{i=1}^{n} a_i} \tag{3.3}$$

Do qual x é o valor de um atributo, X é o conjunto de valor de um atributo, n é a quantidade de valores de um atributo e a_i é o valor de cada atributo.

- *Interquartile Range* em que utiliza-se do intervalo interquartil. Esse intervalo se dá na diferença entre o primeiro quartil (25%), chamado de quartil inferior e o último quartil (75%), denominado quartil superior. Na normalização os dados são classificados obtendo o primeiro ou último quartil dos dados restantes. O quartil do meio (50%) é o valor que separada os valores classificados pela metade. Para realizar a normalização então, utiliza-se o valor do último quartil subtraindo com o valor do primeiro quartil. Sua equação se dá pela seguinte forma:

$$IQR = Q_3 - Q_1 \tag{3.4}$$

No qual IQR é o valor, Q_3 é o valor do último quartil e Q_1 é o valor do primeiro quartil.

3.6.3 Codificação

A etapa de codificação é responsável por refatorar os dados de uma forma que sejam melhor interpretados em todo o processo subsequente do KDD. Algumas limitações de algoritmos específicos de MD podem fazer com que a codificação se torne uma etapa obrigatória. Por exemplo, algoritmos que não conseguem tratar dados categóricos, como o caso das RNAs, precisam imprescindivelmente do apoio da codificação. A forma como a codificação será realizada também é importante no momento em que há a busca por conhecimento. Existem dois tipos de codificação, o primeiro apresentado pela conversão dos dados categóricos em numéricos e o segundo, inverso do primeiro (GOLDSCHMIDT; PASSOS, 2005).

3.6.3.1 Discretização

A discretização, uma forma de codificação numérica para categórica e que também é denominada mapeamento em intervalos, é a etapa na qual os valores numéricos são convertidos em intervalos. Tais intervalos podem ser de comprimentos definidos pelo responsável pela modelagem, de comprimentos iguais ou até por meio de agrupamento (GOLDSCHMIDT; PASSOS, 2005).

3.6.3.2 Representação Discreta Padrão

Nesta representação os valores categóricos são associados a valores discretos de 1 até N (GOLDSCHMIDT; PASSOS, 2005).

3.6.4 Filtering de Texto

A etapa de *filtering* se torna responsável por remover todos os caracteres de pontuação, apesar desses caracteres terem a possibilidade de alterar o sentido da palavra, geralmente o sentido como um todo é mantido (HACK *et al.*, 2013).

3.7 MINERAÇÃO DE DADOS

A etapa de mineração, em que foram processadas todas as etapas anteriores do KDD, é segmentada em três subetapas: definir a tarefa, escolher o algoritmo e mineração de dados (TENFEN, 2003; MARKUSOSKI *et al.*, 2019).

Na primeira subetapa, de definição da tarefa, é determinada a tarefa de mineração de dados, com base nos objetivos e nas etapas anteriores do KDD. Existem dois objetivos na MD, a seleção de um exclui o outro, são eles: a previsão chamada de MD supervisionada, em que há uma classe alvo; ou a previsão denominada de MD não supervisionada, na qual se infere sobre um domínio. As técnicas de mineração de dados por sua vez utilizam a aprendizagem indutiva, em que a modelagem é realizada a partir da generalização de dados de treinamento. Tal aprendizagem é utilizada pois, em sua essência, o modelo já treinado poderá ser utilizado com dados desconhecidos, ou seja, dados novos (MARKUSOSKI *et al.*, 2019).

A segunda subetapa, responsável pela escolha do algoritmo, é uma etapa estratégica, já que a tática foi escolhida, os dados estão formatados, supostamente, da melhor maneira. Nesta subetapa a compreende-se das melhores circunstâncias nas quais determinado algoritmo é mais apropriado, para assim se tornar um candidato. Os algoritmos de mineração de dados possuem particularidades, como ter suporte a determinados tipos de dados que outros algoritmos não tenham, serem supervisionados ou não, dentre outros parâmetros (MARKUSOSKI *et al.*, 2019).

A terceira, e última, subetapa trata da implementação ou utilização do algoritmo selecionado. A modificação dos parâmetros do algoritmo pode ser realizada inúmera vezes até que o resultado seja o mais satisfatório possível (MARKUSOSKI *et al.*, 2019).

3.7.1 Tarefas de Mineração de Dados

As tarefas são divididas em cinco principais, cada uma com seu objetivo específico, na qual uma será escolhida com base em seu objetivo principal. As divisões são mostradas no Quadro 1.

Com o objetivo de precificação, juntamente com as informações fornecidas pelo Quadro 1, a tarefa é escolhida. As tarefas podem ser confusas caso não sejam totalmente conhecidas. Com o objetivo de precificação a tarefa a ser utilizada será classificação, pois a previsão se dará por intervalos de preços e não um preço específico. O resultado da classificação por sua vez pode ser utilizado no auxílio de itens referentes ao mercado imobiliário ou na comparação com outros métodos.

3.7.2 Técnicas de Mineração de Dados

As técnicas são primeiramente filtradas com base na tarefa definida, pois cada técnica possui sua singularidade. Posteriormente, com base no domínio e na preferência do modelador a técnica é finalmente definida. O Quadro 2 apresenta algumas das técnicas de mineração de dados, uma breve descrição das mesmas e a quais tarefas elas estão associadas.

Este trabalho se concentrou no uso de árvores de decisão, mais especificamente em RF, buscando confirmar a adequação do modelo em questão aos dados dispostos, e assim, criando uma base inicial para comparação futura com outras técnicas. Portanto, deixando para trabalhos futuros o uso de outras técnicas e comparações entre elas.

3.7.3 Árvore de Decisão (AD)

Árvore de decisão (AD) é uma opção interessante para classificação e uso posterior como modelo de comparação, graças a sua grande utilização para tal tarefa. Para uma maior compreensão deste tipo de modelagem, segue abaixo algumas informações essenciais, juntamente com uma ilustração, Figura 6.

Os nós de uma árvore podem ser representados como o começo de ramos. Em uma árvore de decisão existem três tipos de nós. O nó raiz, representando um teste sobre um atributo que resultará na ramificação de dois ou mais nós respectivamente exclusivos. O(s) nó(s) interno(s), representando o restante dos testes sobre atributos disponíveis, ou até mesmo algum atributo já utilizado. O(s) nó(s) folha, que resulta(m) em uma classe – por se tratar de classificação – derivada do conjunto de condições selecionadas dos nós anteriores ao nó folha (GOLDSCHMIDT; PASSOS, 2005; SONG; YING, 2015).

Quadro 1 – Tarefas realizadas por técnicas de mineração de dados

Tarefa	Descrição	Exemplo
Classificação	Usada para construir um modelo para ser aplicado a dados não classificados a fim de categorizar dados em classes, com o objetivo de relacionar a classe alvo a um conjunto de atributos dos dados.	Classificar pedidos de crédito. Esclarecer pedidos de seguros fraudulentos. Identificar a melhor forma de tratamento de um paciente.
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida.	Estimar o número de filhos ou a renda total de uma família. Estimar o valor em tempo de vida de um cliente. Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos. Prever a demanda de um consumidor para um novo produto.
Associação	Usada para determinar quais itens tendem a ser adquiridos juntos em uma mesma transação.	Determinar que produtos costumam ser colocados juntos em um carrinho de supermercado.
Segmentação (ou *Clustering*)	Usada para particionar uma população heterogênea em vários subgrupos ou grupos mais homogêneos.	Agrupar clientes por região do país. Agrupar clientes com comportamento de compra similar. Agrupar seções de usuários *web* para prever comportamento futuro de usuário.
Sumarização	Usada com diferentes métodos para encontrar uma descrição compacta para um subconjunto de dados.	Tabular o significado e desvios padrão para todos os itens de dados. Derivar regras de síntese.

Fonte: Adaptado de Dias (2002)

Quadro 2 – Técnicas de mineração de dados

Técnica	Descrição	Tarefas
Descoberta de Regras de Associação	Utilizada para estabelecer uma correlação estatística entre atributos de dados e conjuntos de dados.	Associação
Árvores de Decisão	Utilizada para realizar a hierarquização dos dados, baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.	Classificação Regressão
Raciocínio Baseado em Casos ou MBR	Utilizado para estabelecer uma hierarquia de semelhança, baseada no método do vizinho mais próximo, combinando e comparando atributos.	Classificação Segmentação
Algoritmos Genéticos	Utilizados para busca e otimização, inspirados na Teoria da Evolução, onde a cada nova geração, soluções melhores têm mais chance de terem "descendentes".	Classificação Segmentação
Redes Neurais Artificiais	Utilizadas na busca de conhecimento, inspirando-se na fisiologia do cérebro, sendo resultante por meio do mapa das conexões neurais e dos pesos das mesmas.	Classificação Segmentação

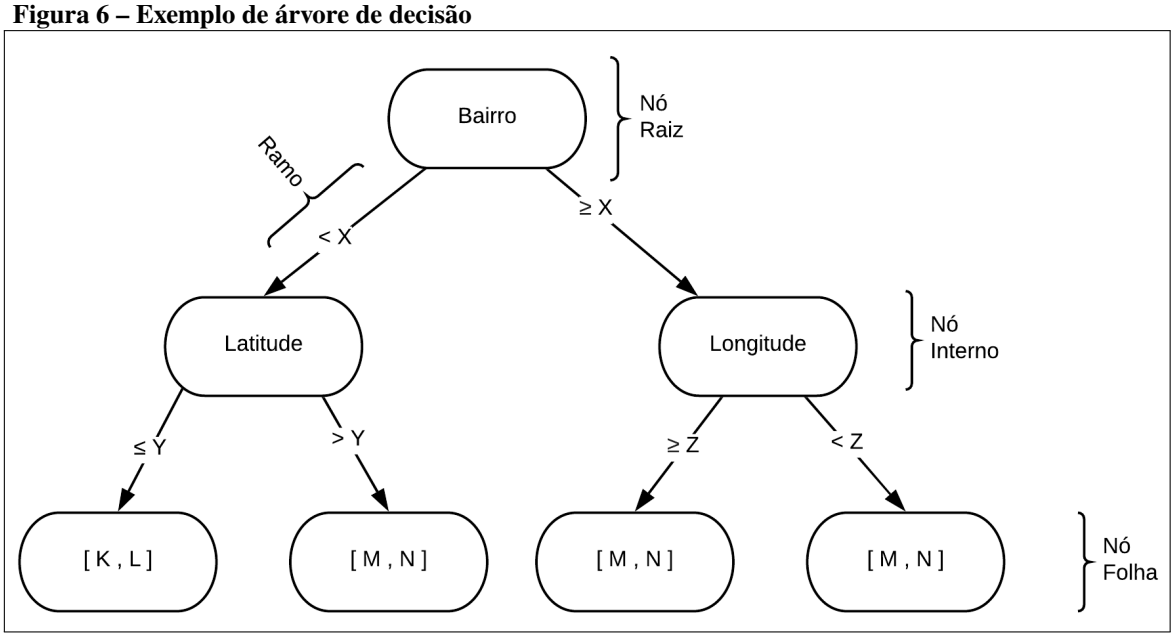
Fonte: Adaptado de Dias (2002)

A conexão entre nós se dá pelos denominados ramos. Cada ramo por sua vez representa uma decisão realizada por uma condição "se-então", por essa razão a árvore de decisão é facilmente implementada (GOLDSCHMIDT; PASSOS, 2005; SONG; YING, 2015).

Para criar uma AD particiona-se de forma recursiva o conjunto de dados de treinamento até que as amostras fiquem classificadas. Durante a construção da AD duas operações são realizadas, são elas: a avaliação do melhor ponto de separação e a criação das partições referentes ao ponto de separação encontrado (GOLDSCHMIDT; PASSOS, 2005).

Para avaliar um ponto de separação, cada nó da árvore é analisado, utilizando três etapas (GOLDSCHMIDT; PASSOS, 2005):

Figura 6 – Exemplo de árvore de decisão



Fonte: Autoria Própria

1. Ganho de informação levando em conta a partição da qual o nó analisado pertence.

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) bits \tag{3.5}$$

   Dos quais S é referente a partição, $freq(C_j, S)$ é a quantidade da classe C_j acontece em S, $|S|$ representa o número de casos do conjunto S e k indica a quantidade de classes existentes.

2. Ganho de informação de cada atributo levando em conta a partição da qual o nó analisado pertence.

$$info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \tag{3.6}$$

   Dos quais T representa a quantidade total de vezes que todas as classes aparecem e T_i a quantidade de vezes que uma classe aparece no conjunto T.

   Que se dá pelo ganho de informação obtido pela equação:

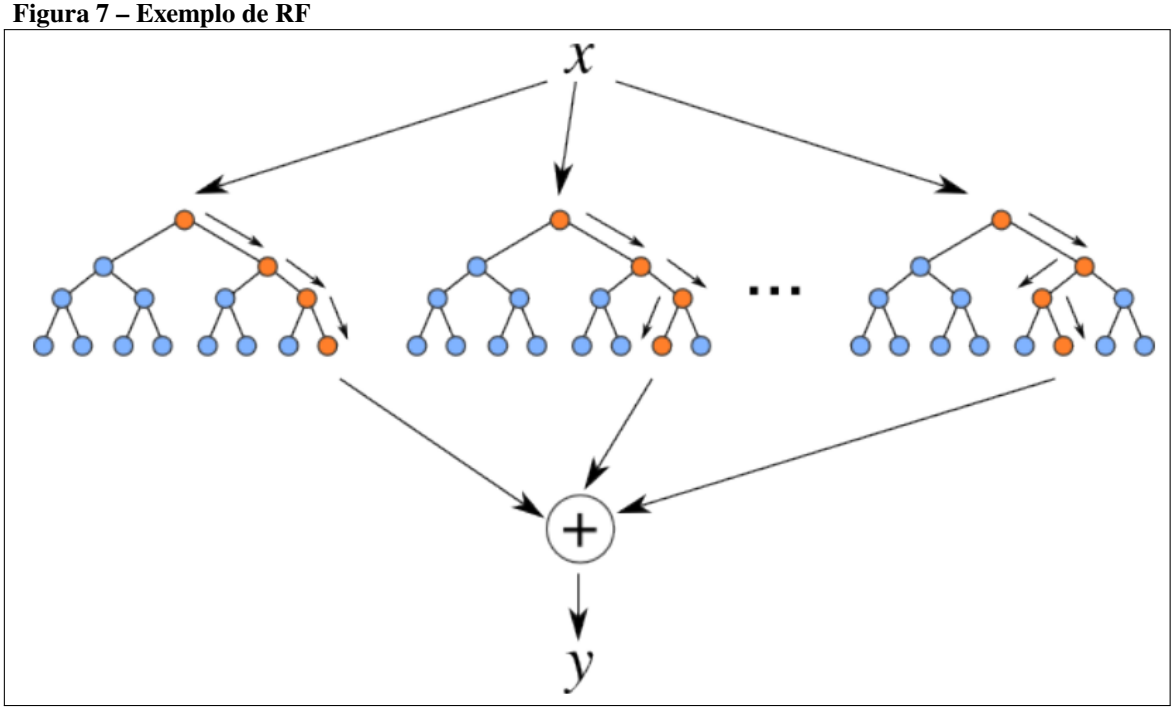
$$gain(X) = info(T) - info_x(T) \tag{3.7}$$

3. E por fim, o atributo com o maior ganho é selecionado como ponto de separação.

3.7.4 *Random Forest* (RF)

A *Random Forest* (RF) nada mais é que uma estrutura contendo uma quantidade especificada de AD. A quantidade n é especificada pela pessoa a realizar a modelagem. Um conjunto de

Figura 7 – Exemplo de RF



Fonte: Teloken (2016)

dados é fornecido como entrada para essas n árvores. Cada árvore processa os dados e os rotula individualmente. A rotulação para cada dado é dada pela classe com maior frequência. Caso a frequência entre classes seja igualitária a classe será escolhida aleatoriamente (BREIMAN, 2001). Abaixo segue a ilustração de RF feita pela Figura 7.

Na Figura 7, X representa um conjunto de dados. As setas que saem deste conjunto indicam a direção, ou fluxo, dos dados. Posteriormente, os dados servem como entrada em cada uma das árvores previamente determinadas. As setas dentro das árvores, juntamente com os nós destacados de laranja, indicam qual a decisão tomada em cada uma das árvores. As setas abaixo das árvores, direcionadas ao símbolo de soma, indicam que os dados passarão por uma classificação, onde a classe com maior frequência consequentemente será atribuída.

3.8 AVALIAÇÃO E INTERPRETAÇÃO DOS RESULTADOS

A etapa de avaliação é realizada em relação aos resultados da etapa de MD. O modelo gerado é então analisado em relação à sua utilidade, ou seja, se o objetivo foi alcançado e, conforme isso, analisado também sobre sua interpretação. Por conta do processo de KDD ser iterativo o conhecimento descoberto fica armazenado para que possa ser utilizado futuramente (MARKUSOSKI *et al.*, 2019).

3.8.1 Validação Cruzada com K Conjuntos (*K-Fold CrossValidation*)

Esta validação divide o conjunto de dados em K subconjuntos (*folds*) com N elementos cada. A quantidade de elementos costuma ser a mais igualitária possível, sendo assim a equação $\frac{N}{K}$ é utilizada. Cada um dos K subconjuntos são utilizados como conjunto de teste enquanto os $K-1$ subconjuntos viram conjuntos de treinamento. Esse processo é realizado K vezes até que todos os K subconjuntos sejam analisados como conjunto de teste (GOLDSCHMIDT; PASSOS, 2005).

3.8.2 Matriz de Confusão

A matriz de confusão é utilizada para avaliação do modelo gerado e é geralmente aplicada nos aprendizados supervisionados. A estrutura de uma matriz de confusão se dá pela classificações incorretas e, consequentemente corretas. A utilização desta técnica se dá pela inspeção dos erros para cada classificação, tornando possível realizar ajustes nos parâmetros do algoritmo de MD e comparação das versões de cada modelo (BEAUXIS-AUSSALET; HARDMAN, 2014).

A matriz de confusão, representada pelo Quadro 3, é uma matriz gerada baseada em um modelo com duas classes. Cada linha corresponde a uma classe e cada coluna corresponde as mesmas classes das linhas. Como apresentado no Quadro 3 a matriz apresenta quatro valores principais, são eles (HAN; PEI; KAMBER, 2011):

- Verdadeiro Positivo (VP): São as tuplas que foram classificadas corretamente, por exemplo, levando em consideração onde a classe real fosse A, a classe prevista seria A também.
- Verdadeiro Negativo (VN): Igualmente ao Verdadeiro Positivo (VP), porém, alterando a classe A, para por exemplo, classe B.
- Falso Positivo (FP): São as tuplas que foram classificadas incorretamente, por exemplo, levando em consideração onde a classe real fosse A, a classe prevista seria B.
- Falso Negativo (FN): Igualmente ao Falso Positivo (FP), porém, alterando a classe A, para por exemplo, classe B e vice-versa.

Quadro 3 – Exemplo de uma Matriz de Confusão

		Classe Predita	
		Classe A	Classe B
Classe Real	Classe A	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Classe B	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Autoria Própria

Sabendo as classificações corretas e incorretas algumas informações podem ser extraídas, como (HAN; PEI; KAMBER, 2011):

- Acurácia: É referente a quantidade de classes, por exemplo classes A, B e assim sucessivamente, que foram classificadas corretamente, expressa-se pela função: $(VP+VN)/(VP+VN+FP+FN)$
- Precisão: É referente a quantidade que uma classe, por exemplo a classe A, foi classificada corretamente, expressa-se por: $VP/(VP+FP)$
- *Recall*: É referente a taxa de valores classificados como X comparada com quantos deveriam ser, expressando-se por: $VP/(VP+FN)$
- *F-score*: É a combinação da precisão com o *recall*, expressando-se por: $(2 * precisao * \textit{recall})/(precisao + \textit{recall})$

3.9 TRABALHOS RELACIONADOS

A partir de uma busca realizada na internet alguns artigos e trabalhos foram selecionados por serem relacionados a proposta do presente trabalho, trazendo um acrescimento no referencial teórico juntamente com alguns trabalhos realizados anteriormente.

Silva (2019) propõem encontrar o analisar alguns modelos e encontrar o melhor que faça o cálculo das unidades habitacionais unifamiliares da cidade de Fortaleza utilizando de aprendizagem de máquina.

Veras (2019) utiliza redes neurais recorrentes para realizar previsões sobre os preços de imóveis no Distrito Federal, por meio do uso de uma série temporal com preços médios mensais e a localização dos imóveis transacionados entre os anos de janeiro de 1997 a dezembro de 2011.

3.10 CONSIDERAÇÕES FINAIS

Este capítulo iniciou-se falando sobre o processo de KDD seguido das diferenças entre tipos de dados e variáveis, no qual cada tipo é descrito. Posteriormente são descritas as etapas presentes no processo de KDD, como a seleção dos dados, pré-processamento, transformação, mineração de dados e avaliação. Na etapa de mineração de dados são apresentados os algoritmos de AD e RF. Por fim, são apresentados alguns trabalhos que possuem relação com o presente TCC, nos temas de Mineração de Dados, Inteligência Artificial, mercado imobiliário e valores imobiliários.

4 DESENVOLVIMENTO

Neste Capítulo todos os métodos utilizados para o desenvolvimento do trabalho são apresentados levando em consideração as etapas do KDD.

4.1 BASE DE DADOS

Os dados utilizados são referentes a uma base de dados do trabalho de Roberto (2019), onde tais dados encontram-se previamente padronizados. Esses dados por sua vez foram padronizados com o propósito de serem continuados por este trabalho.

A base de dados consiste de aproximadamente 65 mil dados retirados do código fonte de páginas *web*, com 14 atributos. Dois atributos dos quatorze sendo retirados utilizando georreferenciamento (ROBERTO, 2019).

4.2 LIMPEZA E SELEÇÃO DE DADOS

Neste trabalho uma variação da junção orientada foi utilizada. O motivo de se utilizar uma variação se dá pela falta de um especialista do assunto, cabendo assim apenas ao autor analisar os atributos a serem selecionados, juntamente com os estudos feitos sobre o mercado imobiliário. Conjuntamente com tal abstração a eliminação direta de atributos foi realizada. Os dados brutos, ou seja, sem nenhuma modificação, possuem os atributos apresentados no Quadro 4 a seguir.

Os dados brutos precisaram passar pela seleção e remoção de atributos para que pudessem ser melhor aproveitados posteriormente pelo algoritmo selecionado no passo de mineração de dados, tornando assim essencial os tratamentos realizados nos mesmos.

Como na modelagem a eliminação direta dos atributos se dá pelo estudo e escolha do modelador. Neste caso alguns atributos, apresentados pelo Quadro 5, foram removidos pela justificativa de não serem relevantes para a proposta de precificação a ser realizada por este trabalho, baseando-se no estudo realizado sobre o mercado imobiliário. Portanto, removendo os atributos considerados irrelevantes para a modelagem do problema, os atributos mantidos foram: *data*, *valor*, *latitude*, *longitude* e *bairro*.

Quadro 4 – Atributos existentes nos dados brutos

Atributo	Tipo	Descrição
data	Texto	A data na qual o anúncio ficou disponível
enderecoAnuncio	Texto	O endereço adquirido no anúncio
enderecoMatch	Texto	O endereço encontrado no georreferenciamento
link	Texto	O endereço URL do anúncio extraído
ref	Inteiro	ID do anúncio no site
valor	Real	O valor do imóvel (em reais)
latitude	Real	O valor da latitude do imóvel
longitude	Real	O valor da longitude do imóvel
tipo	Texto	Especifica o tipo da coordenada extraída – na maioria dos casos o valor deste atributo é "ponto" por se tratar apenas de uma coordenada.
bairro	Texto	O bairro do qual o imóvel pertence
cidade	Texto	A cidade onde o imóvel se encontra
logradouro	Texto	Nome referente a rua ou avenida
numero	Inteiro	O número de identificação do imóvel
vila	Texto	Atributo vazio

Fonte: Autoria Própria

4.3 PRÉ-PROCESSAMENTO DOS DADOS

Na etapa de pré-processamento foi utilizada a limpeza dos dados por meio da eliminação de dados com informações ausentes.

Sendo o método mais simples da subtarefa de limpeza, consequentemente o consumo de tempo excessivo ou alto não acontece neste momento. Este meio se dá pela exclusão dos dados do domínio que possuam alguma informação ausente nos atributos relevantes selecionados na etapa de seleção. A utilização deste método se dá pela facilidade tornando-o um baixo custo computacional para o KDD, quantidade muito baixa de valores faltantes, quantidade relativamente grande de dados após a utilização.

Quadro 5 – Atributos excluídos dos dados brutos

Atributo	Justificativa
enderecoAnuncio	O endereço é utilizado para comparar com o enderecoMatch e verificar se eles batem, caso sejam iguais, é utilizado somente o endereço
enderecoMatch	Utilizado somente um endereço, este atributo é utilizado para comparação, caso haja, o endereço será utilizado
link	A URL não ajuda na precificação
tipo	Sempre o mesmo valor, fazendo com que haja generalização.
cidade	Sempre o mesmo valor.
logradouro	Obtem-se pelas coordenadas geográficas (*latitude* e *longitude*)
numero	O número de identificação do imóvel não auxilia na precificação do mesmo
vila	O atributo sempre está vazio, portanto foi removido

Fonte: Autoria Própria

4.4 TRANSFORMAÇÃO DOS DADOS

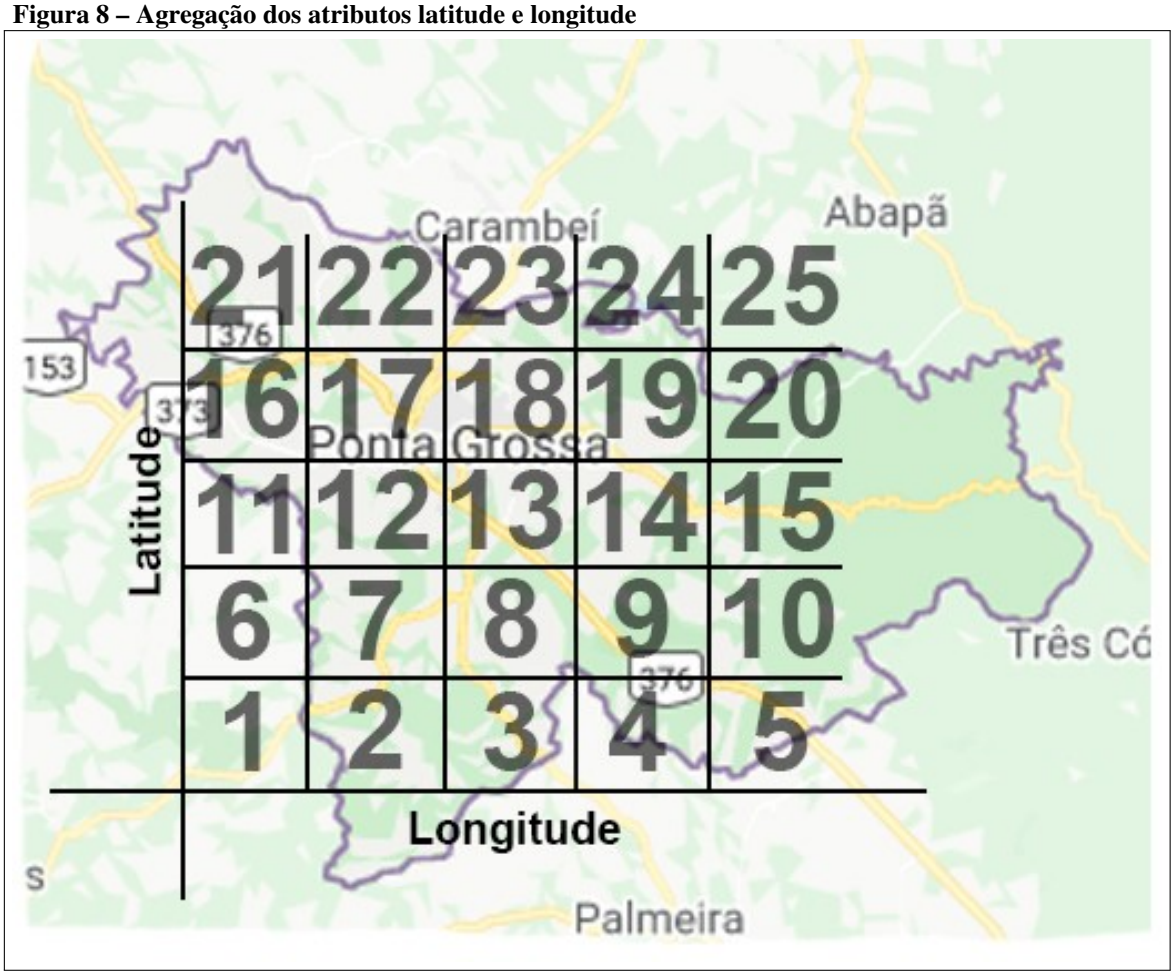
Na etapa de transformação dos dados alguns métodos foram utilizados, são eles: agregação, normalização, codificação e *filtering* de texto.

Na tentativa de amenizar os erros recorrentes da utilização de dois atributos, *latitude* e *longitude*, o mapa da cidade de Ponta Grossa foi abstratamente projetado em um plano cartesiano, dividido em quadrados. O plano por sua vez delimitado pela *latitude* e *longitude*, como na Figura 8.

Esperava-se que a agregação pudesse auxiliar na modelagem, fazendo a repartição dos valores por regiões, trazendo assim uma precificação mais específica de partes de Ponta Grossa. Porém, por não trazer resultados positivos ao modelo a utilização deste método foi removida do trabalho e substituída pela utilização da discretização.

A normalização foi aplicada utilizando todos os métodos, que foram testados em busca do que se comportasse melhor ao domínio.

Figura 8 – Agregação dos atributos latitude e longitude



Fonte: Adaptado de Google Maps (2020)

Na codificação foram utilizados os métodos de discretização e representação discreta padrão.

A discretização foi realizada nos atributos *valor*, *latitude* e *longitude*. A quantidade de partições do atributo preço se tornou estática em duas. Os atributos *latitude* e *longitude* por sua vez tiveram, durante o trabalho, a quantidade alterada algumas vezes com foco na melhora da acurácia do projeto.

A representação discreta padrão foi aplicada no atributo *bairro*, ou seja, cada bairro tornou-se um valor inteiro único, como apresentada a correspondência de alguns bairros pelo Quadro 6.

E para finalizar, o *filtering* de texto, que neste trabalho a utilização se deu pelas diversas maneiras que os bairros, valores do atributo *bairro*, estão redigidos. Com isso, primeiramente todos os bairros ficaram em caixa alta, para que posteriormente fosse mais fácil a remoção dos caracteres de acentuação. Em seguida todas as letras que antes possuíam acentos, após esse processo, ficaram sem. Por exemplo, o bairro “Órfãs” pode ser escrito de diferentes formas, corretas ou não, como: “Órfãs”, “Orfãs”, “Orfas”, “Órfas”, “orfas” “órfãs”, “orfãs”, “órfas”, dentre outras possibilidades.

Quadro 6 – Bairros com seus inteiros correspondentes

Bairro	Inteiro correspondente
Centro	0
Jardim Carvalho	1
Oficinas	2
Uvaranas	3
Órfãs	4
Estrela	5
Nova Rússia	6

Fonte: Autoria Própria

4.5 MINERAÇÃO DE DADOS

A etapa de mineração de dados possui três subetapas, como apresentado anteriormente, a primeira por sua vez é onde a definição da tarefa é feita, no caso deste trabalho, utilizando aprendizagem supervisionada, que busca uma classe alvo que por sua vez utiliza da tarefa de classificação, já que a resposta final é um intervalo de preços.

A segunda subetapa, de escolha do algoritmo, é onde encontra-se um candidato mais apropriado. Para a proposta deste trabalho, escolheu-se a RF que consegue tratar todos os tipos de dados e consegue realizar a tarefa de classificação.

A terceira e última subetapa, trata da modificação dos parâmetros do algoritmo escolhido, onde na questão da RF existe o número de árvores. Após alguns testes utilizando diversos valores de árvores e procurando uma acurácia acima de 60% conseguiu-se a número 10, que trouxe o modelo para resultados próximos aos 65%.

4.6 CONSIDERAÇÕES FINAIS

Neste capítulo todas as etapas do KDD apresentadas no Capítulo 3 são executadas, apresentando métodos testados que não tiveram êxito, como é o caso da Agregação, bem como métodos que apresentaram um acréscimo à qualidade do projeto. A utilização do *software* foi realizada aplicando todos os métodos e tornando prática toda a modelagem. O algoritmo de mineração de dados RF também é executado e ajustado em sua quantidade de árvores tendo como objetivo diminuir seus erros de classificação. Todas as formatações de dados foram realizadas no intuito de melhorar a classificação de preços.

5 RESULTADOS

Vários resultados foram adquiridos para que uma melhor comparação fosse feita. Configurações diversas foram utilizadas, como apresentadas no Quadro 7. A validação cruzada foi utilizada com o número K equivalente a 10, após testes que apresentaram o modelo mais factível a este valor. O resultado da acurácia da RF de cada configuração é apresentado em porcentagem no Quadro 9. Os melhores resultados de cada configuração estão em negrito. Os dois melhores resultados estão detalhados no final deste tópico.

Quadro 7 – Cenários dos conjuntos de normalizações

Cenários das Normalizações	**Normalização utilizada no atributo valor**	**Normalização utilizada nos atributos latitude e longitude**
Norm1	*Z-Transformation*	*Z-Transformation*
Norm2		*Range Transformation*
Norm3		*Proportion Transformation*
Norm4		*Interquartile Range*
Norm5	*Range Transformation*	*Z-Transformation*
Norm6		*Range Transformation*
Norm7		*Proportion Transformation*
Norm8		*Interquartile Range*
Norm9	*Proportion Transformation*	*Z-Transformation*
Norm10		*Range Transformation*
Norm11		*Proportion Transformation*
Norm12		*Interquartile Range*
Norm13	*Interquartile Range*	*Z-Transformation*
Norm14		*Range Transformation*
Norm15		*Proportion Transformation*
Norm16		*Interquartile Range*

Fonte: Autoria Própria

Se tratando de discretizações, os atributos *latitude* e *longitude* tiveram modificações na quantidade de partições, sendo testadas de 2 à 5 partições para cada atributo, representadas pelo Quadro 8.

As colunas do Quadro 9 tem como referência as configurações, representadas por cada linha, no Quadro 7. Por exemplo, a coluna *Norm3* do Quadro 9 significa que a normalização aplicada, respectivamente aos atributos *valor*, *latitude* e *longitude*, foi a *Z-Transformation* e *Proportion Transformation* (adquirida consultando o Quadro 8). A seguir o Quadro 9 com as seis primeiras configurações de normalização.

Quadro 8 – Cenários dos conjuntos de partições

Cenários dos conjuntos de partições	Número de partições do atributo latitude	Número de partições do atributo longitude
Part1	2	2
Part2	2	3
Part3	2	4
Part4	3	2
Part5	3	3
Part6	3	4
Part7	4	2
Part8	4	3
Part9	4	4
Part10	5	2
Part11	5	3
Part12	5	4

Fonte: Autoria Própria

Quadro 9 – Acurácia das configurações

Resultados	Norm1	Norm2	Norm3	Norm4	Norm5	Norm6	Norm7	Norm8	Norm9	Norm10	Norm11	Norm12	Norm13	Norm14	Norm15	Norm16
Part1	63,86%	64,13%	63,95%	63,97%	63,75%	63,75%	63,83%	63,80%	63,81%	63,98%	63,79%	63,81%	63,97%	63,83%	63,75%	64,20%
Part2	64,57%	64,66%	64,71%	64,64%	64,87%	64,95%	64,75%	64,79%	64,74%	64,71%	64,64%	64,96%	64,73%	64,73%	64,74%	64,93%
Part3	65,18%	65,17%	65,26%	65,23%	65,13%	65,18%	65,14%	65,06%	65,30%	65,05%	65,19%	64,99%	65,10%	65,20%	65,17%	65,16%
Part4	65,38%	65,44%	65,10%	65,30%	64,95%	65,31%	64,98%	65,32%	65,27%	65,40%	65,35%	65,13%	65,31%	65,32%	64,98%	64,97%
Part5	66,03%	66,21%	66,03%	66,25%	66,27%	65,80%	66,10%	65,80%	66,19%	66,13%	66,13%	66,06%	66,21%	66,29%	65,82%	66,17%
Part6	66,29%	66,39%	66,23%	66,38%	66,39%	**66,39%**	66,32%	66,31%	66,62%	**66,61%**	66,47%	66,41%	66,32%	66,31%	66,27%	66,60%
Part7	64,83%	65,01%	64,88%	64,92%	65,09%	64,91%	64,89%	65,09%	65,09%	64,77%	64,77%	65,12%	64,92%	64,75%	64,91%	64,77%
Part8	65,69%	65,58%	65,67%	65,58%	65,72%	66,01%	65,68%	65,81%	65,50%	65,52%	65,67%	65,81%	65,69%	65,55%	65,54%	65,57%
Part9	65,95%	66,05%	66,21%	66,07%	66,12%	66,17%	65,86%	65,97%	66,13%	66,13%	66,13%	66,01%	66,04%	65,87%	65,84%	65,89%
Part10	65,31%	65,39%	65,48%	65,40%	65,22%	65,35%	65,10%	65,31%	65,29%	65,36%	65,36%	65,36%	65,42%	65,10%	65,36%	65,60%
Part11	66,45%	66,35%	66,37%	66,45%	**66,45%**	66,01%	**66,49%**	66,06%	66,48%	66,47%	**66,67%**	**66,47%**	**66,49%**	**66,47%**	**66,48%**	66,47%
Part12	**66,72%**	**66,65%**	**66,74%**	**66,66%**	66,39%	66,37%	66,38%	**66,37%**	**66,78%**	66,43%	66,39%	66,37%	66,44%	66,43%	66,37%	**66,68%**

Fonte: Autoria Própria

Focando nos dois resultados mais interessantes tem-se na coluna *Norm3* do Quadro 9 o valor de 66,74% e na coluna Norm7 do Quadro 9 o valor de 66,78%. O Quadro 10 apresenta a matriz de confusão do resultado referente a configuração *Norm3*, utilizada na modelagem, que resultou em 66,74% de acurácia e 66,40% de precisão.

Quadro 10 – Matriz de confusão

	[-∞ à 0]	**[0 à ∞]**
[-∞ à 0]	5030	2467
[0 à ∞]	2548	5035

Fonte: Autoria Própria

O Quadro 11 apresenta a matriz de confusão resultante da aplicação da configuração *Norm9* que resultou em 66,78% de acurácia e 66,43% de precisão do modelo.

Quadro 11 – Matriz de confusão

	[-∞ à 0]	**[0 à ∞]**
[-∞ à 0]	5032	2464
[0 à ∞]	2546	5038

Fonte: Autoria Própria

6 CONCLUSÃO

O presente trabalho teve como objetivo classificar preços de imóveis por meio do KDD, que consiste nas etapas de seleção, pré-processamento, transformação, mineração de dados e avaliação, aplicado a uma base de dados com informações imobiliárias obtida a partir do trabalho de Roberto (2019). Para a etapa de mineração foi utilizada a RF que constituí-se de várias AD, e, para o processo de avaliação, a matriz de confusão, que aponta a acurácia do modelo.

Todo o desenvolvimento deste trabalho se deu por investigar o uso das técnicas selecionadas com base nos dados disponíveis verificando diversos fatores. Sendo assim, também podendo ser útil no planejamento urbano, tarefas de simulação e projeção de crescimento urbano, como também servindo como base para testes com trabalhos futuros.

Analisando os resultados, no Quadro 9, pode-se notar que todos os testes realizados obtiveram acurácia igual, ou superior, a 60%. Vale ressaltar que os testes continham diferentes configurações para normalização e discretização. Assim, o presente trabalho apresentou um resultado satisfatório, podendo ser considerado um modo de comparação positivamente interessante.

Por fim nota-se que este trabalho concluiu com os objetivos definidos, sendo possível dar sequência ao estudo acerca de simulação urbana por meio de pesquisas futuras, do qual ao observar todo o processo realizado por este trabalho sugere-se as seguintes tarefas:

- Aumentar a quantidade de dados afim de investigar o comportamento do modelo;
- Comparar os resultados com outras metodologias;
- Aumentar a quantidade de atributos para verificar a influência desses no modelo;
- Utilizar outras métricas de avaliação;
- Aumentar a quantidade de intervalos de preços para que a classificação seja mais precisa;

REFERÊNCIAS

AMARAL, André Sampaio do. **Uma Metodologia Orientada a Dados para Precificação de Imóveis**. Dissertação (B.S. thesis) — Universidade Federal do Rio Grande do Norte, 2018.

BEAUXIS-AUSSALET, Emma; HARDMAN, Lynda. Visualization of confusion matrix for non-expert users. In: **IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings**. [S.l.: s.n.], 2014.

BELGIU, Mariana; DRĂGUŢ, Lucian. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 114, p. 24–31, 2016.

BERTONCELLO, Alexandre Godinho *et al.* Loop econômico: Mercado imobiliário influencia e é influenciado pelas condições socioeconômicas. loop econômico e o mercado imobiliário. In: **Colloquium Socialis. ISSN: 2526-7035**. [S.l.: s.n.], 2019. v. 3, n. 3, p. 35–44.

BODIE, Zvi; KANE, Alex; MARCUS, Alan. **Fundamentos de investimentos**. [S.l.]: AMGH Editora, 2014.

BRASIL, Centro Universitário FEI. Modelagem empírica de rentabilidade no mercado de locação de imóveis na cidade de são paulo/sp utilizando modelos hedônicos e de regressão. 2019.

BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CHEN, Min *et al.* Data, information, and knowledge in visualization. **IEEE computer graphics and applications**, IEEE, v. 29, n. 1, p. 12–19, 2008.

CHIODI, Sarah Isabella. Crime prevention through urban design and planning in the smart city era. **Journal of Place Management and Development**, Emerald Group Publishing Limited, 2016.

CROSBY, Henry *et al.* A spatio-temporal, gaussian process regression, real-estate price predictor. In: **Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. [S.l.: s.n.], 2016. p. 1–4.

DIAS, Maria Madalena. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. **Acta Scientiarum. Technology**, v. 24, p. 1715–1725, 2002.

DOROW, Anderson. Heurística da ancoragem na estimativa de preços de imóveis por corretores profissionais. 2012.

FARNAAZ, Nabila; JABBAR, MA. Random forest modeling for network intrusion detection system. **Procedia Computer Science**, Elsevier, v. 89, n. 1, p. 213–217, 2016.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The kdd process for extracting useful knowledge from volumes of data. **Communications of the ACM**, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996.

FURTADO, Bernardo Alves. **Análise quantílica-espacial de determinantes de preços de imóveis urbanos com matriz de bairros: evidências do mercado de Belo Horizonte**. [S.l.], 2011.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data mining: um guia prático**. [S.l.]: Gulf Professional Publishing, 2005.

GOMES, Alexandre Esberard; MACIEL, Vladimir Fernandes; KUWAHARA, Mônica Yukie. Determinantes dos preços de imóveis residenciais verticais no município de são paulo. **Anais do XL Encontro Nacional de Economia**, p. 1–19, 2012.

HACK, A. F. *et al.* **Text Mining**, Florianópolis: Universidade Federal de Santa Catarina, 2013. Disponível em: <http://www.inf.ufsc.br/~luis.alvares/INE5644/G2_texto.pdf>. Acesso em: 03 jan. 2020.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

KAMUSOKO, Courage; GAMBA, Jonah. Simulating urban growth using a random forest-cellular automata (rf-ca) model. **ISPRS International Journal of Geo-Information**, Multidisciplinary Digital Publishing Institute, v. 4, n. 2, p. 447–470, 2015.

LAIA, AMARO NAVES. Avaliação de imóveis pelo método da cap rate ou yield (ii). 2007.

LI, Heng. **A price prediction method In real estate market**. Tese (Doutorado) — Massachusetts Institute of Technology, 2016.

MA, Chao *et al.* Cost-sensitive deep forest for price prediction. **Pattern Recognition**, Elsevier, v. 107, p. 107499, 2020.

MALERE, João Pedro P.; ALMEIDA, Publio; SANO, Humberto. Predição de preços de imóveis através de aprendizagem de máquina. 07 2019.

MARKUSOSKI, Ljupce *et al.* Knowledge discovery databases (kdd) process in data mining. In: FACULTY OF ECONOMIC PRILEP. **INTERNATIONAL CONFERENCE PROCEEDING**. [S.l.], 2019. p. 529–539.

MORO, Matheus Fernando. Modelo híbrido de séries temporais para previsão de demanda do mercado imobiliário de são paulo. Universidade Federal de Santa Maria, 2017.

PEREIRA, Júlio César; GARSON, Salomão; ARAÚJO, Elton Gean. Construção de um modelo para o preço de venda de casas residenciais na cidade de sorocaba-sp. **Revista Gestão da Produção Operações e Sistemas**, n. 4, p. 153, 2012.

PYLE, Dorian. **Data preparation for data mining**. [S.l.]: morgan kaufmann, 1999.

ROBERTO, Matheus Aparecido da Silva. **Uma solução de extração e georreferenciamento de anúncios imobiliários da internet**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2019.

ROTH, Ellen Cristina Wolf. **Urban growth forecast using segmented and complete maps with the SLEUTH simulator**. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2019.

SANTOS, Joebson Maurílio Alves dos. **A violência urbana e o preço dos imóveis: evidências de como a criminalidade afeta o mercado imobiliário**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2018.

SHAFIZADEH-MOGHADAM, Hossein *et al.* Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. **Computers, Environment and Urban Systems**, Elsevier, v. 64, p. 297–308, 2017.

SILVA, Gustavo Henrique Pinheiro da. Modelos de aprendizagem de máquina para precificação de imóveis na cidade de fortaleza. 2019.

SONG, Yan-Yan; YING, LU. Decision tree methods: applications for classification and prediction. **Shanghai archives of psychiatry**, Shanghai Mental Health Center, v. 27, n. 2, p. 130, 2015.

TELOKEN, Alex. Estudo comparativo entre os algoritmos de mineração de dados random forest e j48 na tomada de decisão. **Simpósio de Pesquisa e Desenvolvimento em Computação**, v. 2, n. 1, 2016.

TENFEN, EMERSON. A técnica de knowledge discovery in databases (kdd) aplicada nas ocorrências atendidas pela polícia militar. UNIVERSIDADE REGIONAL DE BLUMENAU, 2003.

VERAS, André Duarte. Uma proposta de utilização de redes neurais recorrente na previsão de preços de imóveis no distrito federal. 2019.

WHEATON, William C *et al.* Evaluating risk in real estate. **Real Estate Finance**, INSTITUTIONAL INVESTOR, INC., v. 16, p. 15–22, 1999.

____. Real estate risk: a forward-looking approach. **Real Estate Finance**, INSTITUTIONAL INVESTOR, INC., v. 18, n. 3, p. 20–28, 2001.

YU, Hujia; WU, Jiafu. Real estate price prediction with regression and classification. **CS229 (Machine Learning) Final Project Reports**, 2016.