UNIVERSIDADE NOVE DE JULHO - UNINOVE
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

MARCELO FERREIRA ALBANO

ARQUITETURA HÍBRIDA INTELIGENTE PARA CLASSIFICAÇÃO DE LIQUIDEZ IMOBILIÁRIA URBANA EM LEILÕES

SÃO PAULO
2020

MARCELO FERREIRA ALBANO

ARQUITETURA HÍBRIDA INTELIGENTE PARA CLASSIFICAÇÃO DE LIQUIDEZ IMOBILIÁRIA URBANA EM LEILÕES

Dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, como requisito parcial para obtenção do título de Mestrado.

Professor Dr. Domingos Marcio Rodrigues Napolitano - Orientador

SÃO PAULO

2020

UNINOVE
Universidade Nove de Julho

ATA DE DEFESA DA DISSERTAÇÃO

Ao décimo sétimo dia do mês de dezembro de dois mil e vinte, às 15h00, do programa de Pós-Graduação, desta Universidade, reuniu-se em sessão pública a Comissão Julgadora da dissertação de Mestrado de **Marcelo Ferreira Albano** sob o título "ARQUITETURA HÍBRIDA INTELIGENTE PARA CLASSIFICAÇÃO DE LIQUIDEZ IMOBILIÁRIA URBANA EM LEILÕES".

Integraram a comissão os professores: Prof. Dr. Domingos Marcio Rodrigues Napolitano (UNINOVE), o Prof. Dr. Jesús Pascual Mena-Chalco (UFABC), Prof. Dr. Marcos Antonio Gaspar (UNINOVE), e o Prof. Dr. Renato José Sassi (UNINOVE) sob a presidência do primeiro, orientador da dissertação. A banca examinadora, tendo decidido aceitar a dissertação, passou à argüição pública do candidato. Encerrados os trabalhos, os examinadores deram parecer final sobre a dissertação.

	Parecer
Prof. Dr. Domingos Marcio Rodrigues Napolitano	Aprovado
Prof. Dr. Jesús Pascual Mena-Chalco	Aprovado
Prof. Dr. Marcos Antonio Gaspar	Aprovado
Prof. Dr. Renato José Sassi	Aprovado

Parecer:

Aprovado

Em conclusão, o candidato foi considerado aprovado, no grau de Mestre em Informática e Gestão do Conhecimento. E, para constar, eu, Prof. Dr. André Felipe Henriques Librantz, diretor do Programa de Mestrado e Doutorado em Informática e Gestão do Conhecimento, lavrei a presente ata que assino juntamente com os membros da banca examinadora.

São Paulo, 17 de dezembro de 2020.

Prof. Dr. Domingos Marcio Rodrigues Napolitano

Prof. Dr. Jesús Pascual Mena-Chalco

Prof. Dr. Marcos Antonio Gaspar

Prof. Dr. Renato José Sassi

Prof. Dr. André Felipe Henriques Librantz

DEDICATÓRIA

Aos meus pais, que sempre se sacrificaram para que eu pudesse estudar.

A minha prima Gabriela Ferreira pela fonte de inspiração.

AGRADECIMENTOS

Um mestrado é um projeto de pesquisa que envolve muito tempo solitário. Na verdade, não se trata de um esforço pessoal, pois requer interação com várias pessoas, cada uma delas contribuindo para uma fase específica do projeto.

É difícil listar todos os apoios e contribuições, mas aqui estão aqueles a quem acredito ser indispensável o agradecimento neste momento.

Em primeiro lugar, agradeço a Deus por me dar força e convicção para concluir a tarefa que me confiou. Obrigado por me guiar sem hesitar. Através de muitos obstáculos em meu caminho. E por me manter determinado quando o mundo parecia perdido. Agradeço sua proteção e seus sinais ao longo do caminho.

As minhas filhas, Helena e Heloísa, meu afilhado Vinícius, minhas irmãs Marisa e Mônica e à minha querida Mel, por fazerem parte da minha vida, por acreditarem em mim e ajudarem a manter-me motivado mesmo nas horas mais difíceis.

A CAPES e ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho, pela concessão da bolsa de estudos, que viabilizaram a realização deste mestrado.

Agradeço Dr. Andre Zalcman da Zukerman pelo apoio e concessão de uso dos dados que fundamentaram todo o experimento dessa pesquisa. Para mim é uma honra se quer mencionar em meu trabalho essa empresa impar no setor de leilão.

Ao meu orientador, Prof. Dr. Domingos Marcio Rodrigues Napolitano, a quem agradeço pela parceria.

Aos meus colegas Fernando, Hugo, Gustavo, Paola, Renan, Róger, Augusto, Érika e Renata pela paciência, pela parceria e pelas muitas horas de conversa e de estímulo, que foram essenciais nesta jornada.

Ao melhor elenco técnico que já trabalhei, Victor, Naiade e Rafael, cujo apoio foi de muita importância para a realização desse trabalho.

Aos membros da banca de qualificação, Prof. Dr. Renato José Sassi e Prof. Dr. Marcos Antonio Gaspar, cujas preciosas contribuições foram essenciais para a realização deste trabalho.

E finalmente agradeço aos demais membros do corpo docente, discente e aos funcionários do Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho.

RESUMO

O papel de liquidez, no mercado imobiliário, tem atraído atenção na literatura financeira por conta de seu forte impacto na economia e pelos setores em que ela abrange. A liquidez de um imóvel é um indicador de velocidade ou o grau de facilidade com que as propriedades são negociadas, comercializadas e convertidas em valor monetário. Boa parte das informações desses imóveis estão disponíveis em grandes bases de dados pela internet. Se por um lado, o acesso a dados de imóveis não é um problema, extrair conhecimento dessas bases é. Os sistemas de descoberta de conhecimento *Knowledge Discovery in Data Bases* (KDD) são aplicados como uma solução para que a extração do conhecimento na tomada de decisão em condição de risco no ramo imobiliário, uma vez que é incerto estabelecer um limite para essa negociação. As decisões ocupando um espaço central nas organizações, tornam-se mais complexas em condições de incerteza. Isto implica que para atender a demanda pelo sucesso e qualidade das decisões, deve-se estabelecer um processo decisório que terá como elementos centrais, os cenários destas decisões, as alternativas e seus impactos. Logo, definiu-se o seguinte objetivo geral: avaliar técnicas inteligentes e desenvolver uma Arquitetura Híbrida Inteligente (AHI) para classificação de liquidez imobiliária urbana em leilões, apoiando o processo de tomada de decisão com Matriz de Probabilidade e Impacto Dupla (MPID). Para atingir este objetivo foi realizada uma série de experimentos aplicando técnicas inteligentes a uma base de dados reais num site de Leilões, contendo imóveis arrematados e não arrematados, no intervalo de anos de 2016 a 2020, coletados de forma aleatória. A avaliação de técnicas inteligentes para mineração dos dados como *Randon Forest* (RF), Árvore de Decisão e Rede Neural Artificial *Multilayer Perceptron* (MLP), determinou as técnicas mais promissoras e mais aderentes aos dados de imóveis coletados, na atuação conjunta com AHI. A principal característica da AHI é a capacidade de prever valores de descontos, tempo de exposição, número de lances e a classificação do arremate. Logo, o modelo proposto é capaz de prever e classificar a liquidez dos imóveis de leilão através de enriquecimento da base de dados, diminuindo o viés da decisão para a classificação de liquidez imobiliária em leilões. A sinergia da AHI com a MPID, possibilitou mapear as ameaças e também as oportunidades nesse setor. Foi criado nesse trabalho um novo conceito denominado borda de lances, que determina a convergência de lances reais para um imóvel. A avaliação do conhecimento extraído é útil e pode ser aplicado nos setores de leilão e bancário. A solução desenvolvida alcançou *Score* de 75%, treinando a técnica RF da AHI com o número de árvores padrão da biblioteca “randomForest” com 500 árvores.

Palavras Chave: Arquitetura híbrida, Imóveis, Leilão, Decisão, Riscos.

ABSTRACT

The role of liquidity in the real estate market has attracted attention in the financial literature because of its strong impact on the economy and the sectors it covers. The liquidity of a property is an indicator of the speed or the degree of ease with which properties are traded, traded and converted into monetary value. Much of the information on these properties is available in large internet databases. If on the one hand, access to real estate data is not a problem, extracting knowledge from these databases is. Knowledge Discovery in Data Bases (KDD) systems are applied as a solution for the extraction of knowledge in decision making in a risky condition in the real estate business, since it is uncertain to establish a limit for this negotiation. Decisions occupying a central space in organizations become more complex under conditions of uncertainty. This implies that to meet the demand for success and quality of decisions, a decision-making process must be established that will have as central elements, the scenarios of these decisions, the alternatives and their impacts. Therefore, the following general objective was defined: to evaluate intelligent techniques and develop an Intelligent Hybrid Architecture (AHI) for classifying urban real estate liquidity in auctions, supporting the decision making process with a Double Impact and Probability Matrix (MPID). To achieve this goal, a series of experiments were conducted applying intelligent techniques to an actual database on an Auction site, containing auctioned and non-armed properties, in the years 2016 to 2020, collected randomly. The evaluation of intelligent techniques for data mining such as Randon Forest (RF), Decision Tree and Multilayer Perceptron Neural Network (MLP), determined the most promising techniques and most adherent to the real estate data collected, in joint action with AHI. The main characteristic of AHI is its ability to predict discount values, exposure time, number of bids and the classification of the end product. Therefore, the proposed model is capable of predicting and classifying the liquidity of auction properties through the enrichment of the database, reducing the decision bias for the classification of real estate liquidity in auctions. The synergy between AHI and MPID made it possible to map the threats and also the opportunities in this sector. A new concept called bidding edge was created in this work, which determines the convergence of real bids for a property. The evaluation of the extracted knowledge is useful and can be applied in the auction and banking sectors. The solution developed reached 75% Score, training the AHI RF technique with the number of standard trees in the "randomForest" library with 500 trees.

Keywords: Hybrid architecture, Real Estate, Auction, Decision, Risks.

ÍNDICE DE ILUSTRAÇÕES

FIGURAS

GRÁFICOS

TABELAS

QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AM - Aprendizado de Máquinas

ANS - Aprendizado Não Supervisionado

AS - Aprendizado Supervisionado

AHI - Arquitetura Híbrida inteligente

DM - Mineração de Dados (*Data Mining*)

DOM - Dias no mercado ou exposição do imóvel (*Days On Market*)

DT - Árvore de decisão (*Decision Tree*)

IA - Inteligência Artificial

IBGE - Instituto Brasileiro de Geografia e Estatística

MLP - *Perceptron* multicamada (*Multilayer Perceptron*)

MPI - Matriz de Probabilidade e Impacto

MPID - Matriz de Probabilidade e Impacto Dupla

MR - Matriz de Risco

RF - Florestas aleatórias (*Random Forest*)

ROC - Representação da curva gráfica do desempenho da técnica de classificação (*Receiving Operating Characteristic*)

RSL - Revisão Sistemática da Literatura

RROC - Representação da curva gráfica do desempenho da técnica de regressão (*Regression Receiving Operating Characteristic*)

TOM - Tempo no mercado ou exposição do imóvel (*Time On Market*)

WEKA - *Waikato Environment for Knowledge Analysis*

WIPO - *World Intellectual Property Organization*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo introdutório tem finalidade de contextualizar e situar as proposições de resolução do problema que serão desenvolvidos nesta pesquisa. Com este escopo serão dispostos conceitos abordados sobre o assunto, por meio dos seguintes tópicos: Contextualização do tema, Identificação de lacunas de pesquisa, Delimitação do tema, Problema de pesquisa, Objetivos, Justificativa da pesquisa, Modelo conceitual da pesquisa e Organização geral da dissertação.

1.1 CONTEXTUALIZAÇÃO DO TEMA

O setor imobiliário estimula o consumo e favorece investimento das empresas que por consequência, impulsionam a atividade econômica e a geração de empregos no âmbito da construção civil. O financiamento imobiliário cresceu 8,2% em maio de 2020 (ABECIP, 2020). O financiamento imobiliário no Brasil com recursos da poupança Sistema Brasileiro de Poupança e Empréstimo (SBPE) atingiu 7,13 bilhões de reais em maio de 2020, alta de 8,2% ante mesmo mês do ano passado. Nos primeiros cinco meses do ano, os empréstimos destinados à aquisição e à construção de imóveis avançaram 23,2%, atingindo 34,1 bilhões de reais (ABECIP, 2020).

Por outro lado, a inadimplência desses empreendimentos faz o setor bancário acionar seus seguros. Deste modo, muitos imóveis acabam sendo penhorados, ou seja, é feito um pedido através de ação judicial, em que o credor cobra formalmente uma dívida. Se o devedor não tem valor monetário para quitar o débito em atraso, a justiça apreende os bens do devedor, realiza uma avaliação de mercado dos bens que irão a leilão para saldar as dívidas.

Logo, a avaliação patrimonial de imóveis urbanos é praticada em vários setores importantes na economia como: bancos, securitizadoras, seguradoras, cooperativas de crédito, leilões, contabilidade, pessoas físicas e jurídicas, dada a complexidade das forças que determinam e modificam seu comportamento e direção a todo momento, (DEL GIUDICE *et al.,* 2019).

A avaliação patrimonial de ativos imobilizados contribui para operações e práticas de mercado nesse segmento como venda, compra, locação, fusão, cisão, incorporação, financiamentos, *Home Equity* (empréstimo com garantia imobiliária), leilão, seguros e finalidades legais tais como partilha de bens, indenizações, desapropriações, servidões, venda de ações, tributos federais, estaduais e municipais. Portanto, trata-se de um serviço que se refere a um processo metodológico que envolve técnicas baseadas na norma NBR 14653 (ABNT, 2019), aplicadas nas atividades mencionadas, entre outras (TAJANI *et al.,* 2018).

Um dos setores que faz uso intenso da avaliação patrimonial de ativos imobilizados é o setor de leilões imobiliários. O leilão é responsável pela condução de todos os procedimentos legais e é totalmente isento de interesses financeiros, nesse tipo de transação, uma vez que não está comprando ou vendendo produtos, mas se coloca a serviço do proprietário do ativo a ser leiloado e dos potenciais compradores para facilitar a negociação por meio de ofertas conhecidas como lances (CASTRO, 2003).

No setor de leilões imobiliários, quando um imóvel é negociado ou arrematado com sucesso, este denomina-se como liquidado. A liquidação de ativos em leilões acontece de acordo com a demanda de mercado, regulamentos específicos e caracteriza-se pela exposição desses ativos de forma rápida, acessível e legal. Por esse motivo, essa dinâmica de negociação é muito empregada em órgãos públicos e em vários países, por exemplo, os tribunais de justiça. Esse tipo de negociação ocorre em uma situação em que o comprador se encontra interessado, porém, não obrigado a comprar e o vendedor encontra-se obrigado a vender (GAN, 2013).

Portanto, a liquidação pode ser diferente para cada imóvel refletindo o interesse dos compradores em função das suas características específicas, ou seja, alguns imóveis podem ser negociados com mais facilidade do que outros. Desse modo, surge uma questão importante que é mensurar a liquidez de um imóvel, através dos atributos próprios como área, quantidade de vagas, localização, valor de primeira praça; bem como atributos exógenos Produto Interno Bruto (PIB), Índice Geral de Preços de Mercado (IGPM), Taxa Selic e Índice de Preços ao Consumidor (IPCA). Todos esses atributos descritos, influenciam o processo de liquidações dos imóveis.

A liquidez refere-se ao número de dias que uma propriedade permanece à disposição no mercado ativo, conceituado pelo termo *Days on Market* (DOM), sendo uma métrica importante da liquidez do mercado imobiliário. O DOM é um indicador útil para os potenciais compradores avaliarem a liquidez de um imóvel, no entanto, é muito desafiador medir o DOM, pois há uma variedade de fatores que podem impactar o DOM em uma propriedade (HENGSHU *et al.,* 2016).

No conceito de métricas de liquidez, Cheng *et al.* (2008) e Xin He *et al.* (2017) descrevem uma fórmula para descobrir a relação teórica entre preços de imóveis e seu tempo exposto no mercado, no mesmo conceito DOM, no entanto, com nomenclatura diferente, *Time On Market* (TOM).

Observando a natureza desses indicadores, verifica-se que eles são úteis no sentido de mensurar a incerteza na liquidação de um imóvel. Ao avaliar essas métricas sob a ótica do clássico conceito de Knigth (1921) de que o risco é uma incerteza que pode ser mensurada, pode-se então afirmar que essas métricas permitem transformar a incerteza da liquidação em risco.

Consequentemente, a liquidação de um imóvel envolve um processo decisório, no qual o decisor tem que levar em conta uma série de fatores que implicam em obter informações sobre os imóveis de seu interesse e avaliar o retorno, através de informações que o decisor irá buscar, seja em sistemas específicos de apoio a decisão ou em informações que já possui para diminuir a incerteza na tomada de decisões, caracterizando assim uma decisão em condição de risco (MCGREAL *et al.*, 2016).

Os sistemas de apoio à decisão, permitem levar em consideração os principais fatores que determinam a “atratividade” dos investimentos imobiliários em contextos urbanos competitivos, a exemplo, no contexto de leilões. Porém, para que esses sistemas forneçam o devido apoio à decisão é necessário extrair conhecimento de dados, muitas vezes disponíveis em bases de dados do mercado imobiliário urbano. Logo, uma forma de gerenciar o risco, visando mitigá-lo, seria contar com o suporte de sistemas de informação de apoio à decisão (DEL GIUDICE *et al.,* 2019).

Por outro lado, nas bases de dados do mercado imobiliário urbano existe uma grande quantidade de dados, dos quais vários padrões são difíceis de descobrir pelos métodos convencionais, suportados por planilhas de cálculo. Desta forma, os sistemas de descoberta de conhecimento *Knowledge Discovery in Data Bases* (KDD) surgem como uma solução para a extração do conhecimento em bases de dados. Mesmo assim, esse tipo de atividade ainda é muito difícil devido à grande quantidade de dados que devem ser processados para a tomada de decisões nas organizações (SASSI, 2006).

A descoberta de conhecimento no processo de KDD pode ser feita empregando Aprendizado de Máquina (AM), que é o campo de estudo que vem crescendo e que dá aos computadores a habilidade de aprender sem serem explicitamente programados. Com a aplicação de técnicas inteligentes oriundas do AM no apoio às decisões, é possível realizar previsões, cálculo da probabilidade de eventos, reconhecimento de padrões e classificar dados (MITCHELL, 1997).

Por outro lado, as técnicas inteligentes podem ser combinadas para obtenção de resultados mais acurados tornando modelos inteligentes mais promissores, combinando mais de uma técnica. Segundo Goldschmidt e Passos (2005), técnicas podem ser combinadas para gerar as chamadas Arquitetura Híbridas Inteligentes (AHI), que podem classificar dados. Nesse sentido, os riscos podem ser classificados e posteriormente mapeados.

Dentre as ferramentas empregadas para mapear riscos, têm-se as Matrizes de Probabilidade e Impacto (MPIs) proposta por Cox, (2008), que se mostram muito convenientes, mas que, por outro lado, mapeiam apenas ameaças. Para uma análise de risco mais completa, é necessário contemplar ameaças e oportunidades.

Hillson (2002) propõe, portanto, a estruturação de uma Matriz de Probabilidade e Impacto Duplas (MPDI), variando de muito baixo a muito alta aplicada à probabilidade e impactos dos eventos de risco. Porém, a escala de impactos comportaria valores positivos e negativos, representando oportunidades e ameaças, respectivamente.

1.2 IDENTIFICAÇÃO DE LACUNAS DE PESQUISA

De modo geral a literatura consultada sobre a liquidez imobiliária tem sido desenvolvida no sentido de avaliar quais os mecanismos que estão envolvidos no processo de liquidação de um imóvel e indicadores para mensurar liquidez imobiliária, que permite uma maior compreensão das transações do mercado imobiliário e que possa refletir conhecimento necessário para aprimorar políticas públicas, maior eficiência econômica do setor de leilões entre outros do setor econômico.

Por outro lado, ainda existem oportunidades para estender as fronteiras desse conhecimento, sendo um tema importante que demanda a proposição de técnicas mais sofisticadas para análise dos dados na tomada de decisão. Dentre as quais se verificou emprego promissor de técnicas inteligentes oriundas do AM (HENGSHU *et al.,* 2016).

Portanto, este trabalho possui três aspectos relevantes: o primeiro é a avaliação de técnicas inteligentes para seleção e aplicação na AHI para classificação de liquidez de imóveis, o segundo é que tais técnicas possam ser empregadas para um processo decisório no mapeamento de ameaças e também de oportunidades, tema que muitas vezes não é abordado nas pesquisas sobre modelos de tomada de decisão em imóveis urbanos, o terceiro é a proposta de desenvolver-se um estudo sobre a liquidez no setor de leilões, tema ainda pouco abordado em trabalhos acadêmicos, o que configura uma lacuna e um tema relevante e que merece ser estudado.

Tais lacunas são exploradas no Capítulo 2 onde é descrito todo o processo de Revisão Sistemática da Literatura (RSL), que além de identificar os autores seminais neste campo de pesquisa, possibilita estabelecer as lacunas dessa pesquisa acadêmica.

1.3 PROBLEMA DE PESQUISA

1.3.1 Situação problema

A situação problema a seguir, consiste em demostrar a dinâmica de operação dos pilares da temática considerada, explicando como se relacionam e suportam esse trabalho acadêmico.

A classificação de liquidez imobiliária, seja no mercado convencional de imóveis, seja no mercado de leilões de imóveis, é desafiadora por conta de todas as variáveis que afetam e modificam a todo momento sua direção. Logo, existe a necessidade de avaliar esses imóveis a fim de posicioná-lo e classifica-lo frente a esses mercados.

O serviço de avaliação patrimonial é respaldado por legislações e normas técnicas que regulamentam essa atividade e subsidia práticas e procedimentos. Segundo Del Giudice *et al.* (2019), esse serviço é um dos setores mais importantes na economia, tendo em vista os setores envolvidos em seu campo de atuação.

A avaliação de imóveis possui diversas finalidades: venda, negociação de ativos no mercado de ações, locação e sublocação (TAJANI *et.al.,* 2018). Entretanto, no mercado de leilões, essas negociações possuem uma característica específica de venda em tempo menor que o convencional praticado pelo mercado imobiliário tradicional. Por outro lado, segundo Cheng *et al.,* 2008 e Hengshu *et al.,* 2016, o TOM mensura a liquidez desses imóveis no mercado tradicional.

Boa parte desses imóveis estão disponíveis em grandes bases de dados pela internet. Se por um lado, o acesso a dados de imóveis não é um problema, extrair conhecimento dessas bases é. Nesse sentido, Singh *et al.* (2020), propuseram e afirmam que é possível ter previsibilidade de comportamentos imobiliários fazendo o uso de grandes bases de dados.

Logo, os sistemas de descoberta de conhecimento *Knowledge Discovery in Data Bases* (KDD) são aplicados como uma solução para que a extração desse conhecimento possibilitando a tomada de decisão em condição de risco no ramo imobiliário, uma vez que é incerto estabelecer um limite para essa negociação (FAYYAD *et al.,* 1996).

Esses conhecimentos serão consolidados na forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo de KDD, o que é uma maneira genérica de obter a compreensão e interpretação dos resultados (BIGUS, 1996).

No KDD as técnicas de Aprendizado de Máquina (AM), oriundas da Inteligência Artificial (IA), são de fundamental importância para detectar padrões e permitir a criação de modelos preditivos que venha a gerar conhecimento que possam servir de apoio a tomada de decisão. Esse processo é conhecido como mineração de dados, pois trata-se de um processo que possibilita transformar dados em conhecimento.

A mineração de dados é caracterizada pela existência de algoritmos e dadas as tarefas especificadas, o algoritmo será capaz de extrair eficientemente conhecimentos implícitos e úteis do banco de dados. Portanto, a mineração de dados é considerada a etapa mais importante no processo de KDD (SASSI, 2006).

Por outro lado, mesmo utilizando um KDD, tal atividade pode continuar sendo extremamente difícil devido à grande quantidade de dados que deve ser processada. Ainda mais que as decisões, ocupando um espaço central nas organizações, tornam-se mais complexas em condições de incerteza e risco para os negócios (DEL GIUDICE *et al.,* 2019).

Portanto, esse trabalho verifica que existe muitas transações imobiliárias ocorrendo e que há uma grande quantidade de informações disponíveis, porém não extruturadas, que necessitam ser processadas para geração de conhecimento no apoio as decisões no setor de leilões e até mesmo bancário.

1.3.2 Proposições de resolução do problema

As proposições de pesquisa estão relacionadas à predição consistente do arremate imobiliário urbano em leilões através de AHI, que possa apoiar as decisões no setor de leilão.

Deste modo, as Matrizes de Probabilidade e Impacto Duplas (MPID), podem ser empregadas no apoio ao processo de tomada de decisão, mapeando em ameaças e oportunidades as classificações feitas pela AHI.

Abaixo, são apresentadas 2 proposições para esse trabalho de pesquisa:

1. É possível classificar a liquidez imobiliária urbana a partir de aplicação de AHI;
2. A MPID pode apoiar o processo de tomada de decisão no setor de leilões imobiliários.

Essas proposições permitem estabelecer uma resposta à questão de pesquisa: **como classificar liquidez imobiliária urbana em leilões através de AHI, apoiando as decisões com MPID?**

Tais proposições também permitem estabelecer objetivos que direcionam a definição dos procedimentos metodológicos que visam realizar investigações para atingir os objetivos desse trabalho acadêmico.

1.4 OBJETIVOS

1.4.1 Objetivo geral

Para responder à questão de pesquisa proposta, define-se o seguinte objetivo geral: **avaliar técnicas inteligentes e desenvolver uma AHI para classificação de liquidez imobiliária urbana em leilões, apoiando o processo de tomada de decisão com MPID.**

1.4.2 Objetivos específicos

I. Realizar revisão sistemática da literatura sobre o tema abordado, apoiada por técnicas bibliométricas para identificação das lacunas de pesquisa;
II. Desenvolver um inventário para coleta de dados que possibilitem o emprego da AHI para classificação da liquidez de imóveis urbanos em leilão, com auxílio de planilhas do Excel e da linguagem de programação Python;
III. Avaliar e selecionar técnicas inteligentes que melhor se ajustem aos dados coletados através da plataforma experimental de mineração no *software* WEKA;
IV. Aplicar as técnicas na AHI com o emprego da linguagem de programação R;
V. Mapear os resultados das classificações em Matrizes de Probabilidade Impacto Duplas, identificando as ameaças e oportunidades no apoio a tomada de decisão no setor de leilões.

1.5 JUSTIFICATIVA DA PESQUISA

Com o início do *boom* imobiliário, foram grandes aumentos nos empréstimos *Home Equity*, onde o empréstimo acontece com garantia de imóvel. Em caso de inadimplência, os bancos penhoram os bens, ou seja, ocorre a apreensão dos bens do devedor, por mandado judicial, para pagamento da dívida (HE *et al.,* 2015).

Cheng *et al.* (2013) desenvolveram uma técnica para a quantificação do risco de liquidez, inferindo que esse tipo de risco não é uma preocupação somente de credores, mas como de todo o mercado imobiliário local. Sendo assim, os autores levantam que há uma escassez de métricas empregadas para fim de identificação dos principais impulsionadores das medidas de risco de liquidez.

Para as instituições financeiras não precisarem acionar suas garantias através da recuperação judicial em leilões, cada vez mais esse setor necessita de métricas mais robustas na classificação imobiliária. Em outras palavras, essas métricas evitam a consolidação e futura penhora em leilões de ativos imobiliários vinculados a uma condição de garantia a um credor (GIANNOTTI *et al.,* 2011).

O mercado imobiliário está e sempre foi vulnerável a fatores que modificam a todo momento, sua direção e velocidade. Algumas variáveis já estudadas por diversos autores na literatura, conforme exemplos visto no Quadro 1 abaixo, moldam e impactam a liquidez dos imóveis submetidos à venda.

Quadro 1 - Variáveis endógenas e exógenas

Variáveis	**Autor**	**Ano**
Vagas	Turnbull e Zahirovic-Herbert	2011
Casas ocupadas	Turnbull e Zahirovic-Herbert	2012
Taxa de desconto sobre a propriedade	Bian *et al.*	2015
Taxa de juros	Kok *et al.*	2018

Fonte: Autor.

No setor de leilões, Gan (2013) explica que o mecanismo de compra nesse segmento, onde o vendedor possui maior urgência em liquidar o ativo imobiliário, propicia uma oportunidade para quem compra, pois não se encontra compelido a adquirir. Nesse sentido, o setor de leilões mostra-se como uma modalidade mais dinâmica em comparação com a liquidação de imóveis no mercado convencional.

Em vista dos argumentos apresentados, entende-se que o setor de leilões é o meio por onde as instituições financeiras recorrem para liquidar os ativos imobilizados recuperados, para a proteção do patrimônio institucional. Entretanto, esse estudo pode estender suas fronteiras para outros seguimentos que lidam com imóveis, como imobiliárias, seguradoras, cooperativas de crédito e securitizadoras, não só pelo fato de a insolvência prejudicar o sistema de crédito local, mas por prejudicar o comportamento mercadológico imobiliário local (MANSUR *et al.,* 2017).

Isto implica que os empréstimos hipotecários realizados por instituições financeiras de crédito podem se beneficiar com técnicas de classificação de liquidez, mitigando futuros impactos sobre a demanda habitacional imobiliária, evitando que esses ativos sejam penhorados por consequência das retomadas bancárias. Por outro lado o setor de leilões também pode se beneficiar, já que uma classificação equivocada fará o leiloeiro arcar com custas como: avaliação patrimonial, custas de cartório, custas administrativas, custas bancárias, publicidade e publicação em edital.

Sendo assim, essa pesquisa é justificada por existirem poucos trabalhos envolvendo o tema proposto, bem como pela contribuição, possibilitando uma maior amplitude de alcance e visibilidade do emprego dos métodos estudados nos setores bancários, securitários, imobiliário, de cooperativas de crédito, e principalmente no setor de leilões, área ainda pouco abordada em trabalhos acadêmicos.

Considerando esses pontos, o desenvolvimento de AHI capaz de classificar com precisão a liquidez imobiliária, torna a tomada de decisão mais assertiva e consistente, possibilitando a transformação da informação em conhecimento para o tomador de decisão, alinhando sua ampla difusão com a capacidade de automatizar procedimentos complexos. A temática apresenta relevância, atualidade e importância com as áreas de pesquisa, na qual este trabalho se desenvolve, justificando sua pertinência.

1.6 DELIMITAÇÃO DO TEMA

A proposta principal do trabalho de pesquisa consiste em avaliar e selecionar técnicas inteligentes capazes de mensurar liquidez imobiliária urbana em leilões, através de uma arquitetura híbrida, mapeando ameaças e oportunidades, somente em imóveis urbanos em situação de leilão.

Portanto, não configura como escopo, classificar liquidez de bens que não sejam imobilizados e urbanos como: semoventes, veículos, máquinas, equipamentos agrícolas, aeronaves e ativos de natureza intangível.

Durante a realização do trabalho, os imóveis não foram vistoriados fisicamente pelos autores da pesquisa, mas sim, por peritos nomeados por diversas varas cíveis, que faz a gestão e distribuição dos ativos a serem penhorados.

1.7 EIXOS TEÓRICOS DA PESQUISA

O modelo conceitual desta pesquisa acadêmica se baseia na intersecção multiespectral de três eixos teóricos básicos que são: Mercado Imobiliário, Leilões e Aprendizado de Máquinas, que oferecem conceitos e métodos para a solução do problema de pesquisa.

Entende-se que a intersecção dos eixos teóricos Mercado Imobiliário e Leilões, concentra-se a parte de riscos, conceito importante para o problema de pesquisa.

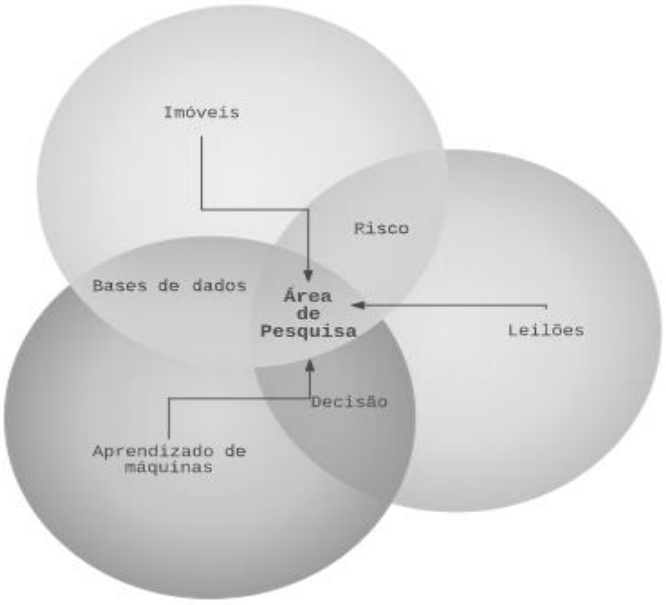
As grandes bases de dados imobiliárias no setor de leilões, pertencem a uma parte dos conjuntos, que intersecciona os eixos teóricos Mercado Imobiliário e Aprendizado de Máquinas. Nesse sentido, abre-se espaço para a extração de conhecimento para aplicação nas decisões.

Por outro lado, as decisões pertencem a uma parte dos conjuntos dado pela intersecção do eixo teórico Leilão com o eixo Aprendizado de Máquinas.

Por fim, o eixo central da pesquisa onde interseccionam todos os 3 eixos teóricos (Mercado imobiliário, Leilões e Aprendizado de Máquinas), estabelece o núcleo de conhecimento proposto por essa dissertação, que possibilitará determinar as técnicas inteligentes capazes de resolver o problema de pesquisa.

Tal estrutura multiespectral, é apresentada na Figura 1 abaixo:

Figura 1 - Eixos teóricos da pesquisa



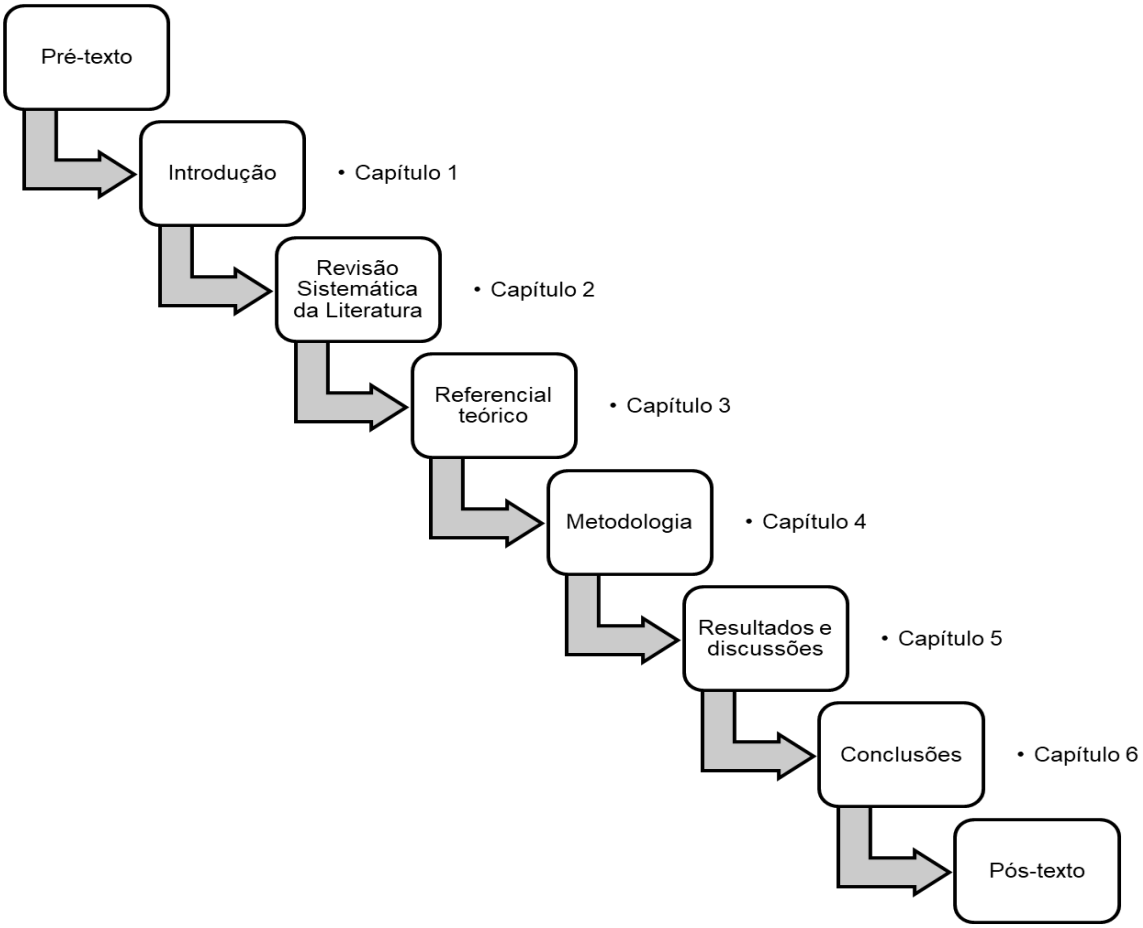
Fonte: Autor.

1.8 ORGANIZAÇÃO DA DISSERTAÇÃO

Com a finalidade de relatar a pesquisa desenvolvida, esta dissertação divide-se em 6 capítulos, dos quais o primeiro é esta Introdução. A Revisão Sistemática da Literatura é apresentada no segundo capítulo. No terceiro capítulo desenvolveu-se um referencial teórico, que teve como objetivo estabelecer os fundamentos teóricos necessários para apoio acadêmico. No quarto capítulo, são descritos os procedimentos metodológicos empregados na pesquisa, entre eles as técnicas de coleta de dados e demais informações necessárias à adequada compreensão dos resultados parciais no quinto capítulo. O quinto capítulo, apresentam-se os resultados parciais da coleta de dados, para desenvolvimento e teste dos modelos propostos e discutem-se os resultados obtidos até o momento. No sexto capítulo, são apresentadas as conclusões parciais deste estudo, sendo indicadas as suas limitações, ações para a continuidade da pesquisa e um cronograma de trabalho executivo, bem como as contribuições desta dissertação para a academia. No pós-texto desta dissertação encontram-se as Referências, os Apêndices e o Anexo I.

A organização desse trabalho acadêmico é mostrada resumidamente na Figura 2 a seguir:

Figura 2 - Estrutura da dissertação



Fonte: Autor.

2 REVISÃO SISTEMÁTICA DA LITERATURA

Nesse capítulo é apresentada a Revisão Sistemática da Literatura (RSL), com objetivo de identificar e relatar pesquisas que suportam as proposições dessa dissertação, bem como, identificar pesquisas relevantes disponíveis na literatura (KITCHENHAM; CHARTERS, 2007).

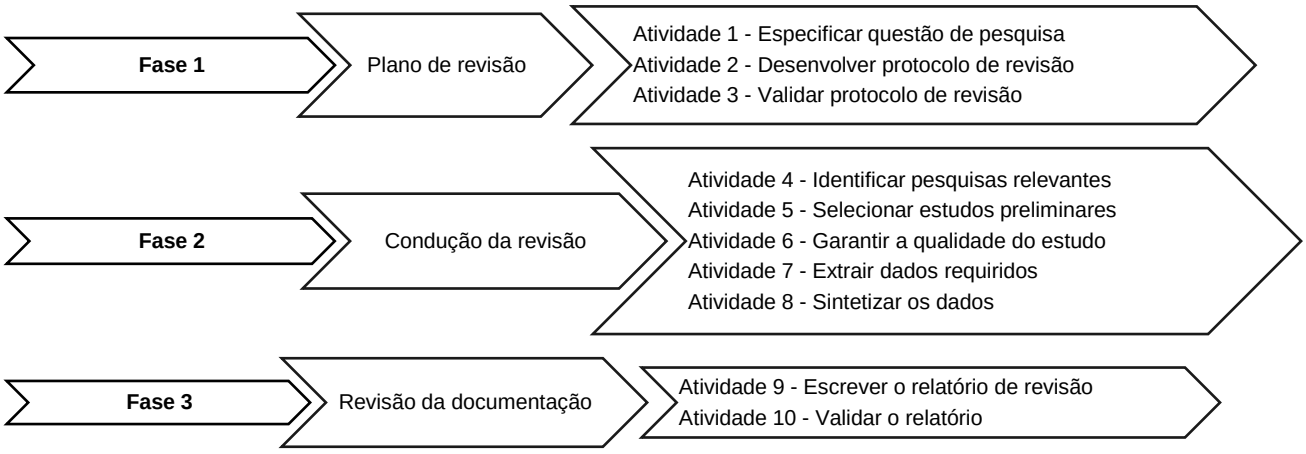
Para isso, a pesquisa baseou-se nos seguintes constructos: imóveis, liquidez, leilões, aprendizado de máquinas e habitação. Esses constructos foram utilizados como palavras-chave para a realização da RSL. Foram levantados dados em 3 bases diferentes, *SCOPUS*, *Web of Science* e IEEE *Xplore.*

O uso dos constructos imóveis, liquidez e leilões, representam os dois eixos teóricos dessa pesquisa acadêmica, Imóveis e Leilão. Esse trabalho propõe desenvolver uma AHI no apoio a decisão no setor de leilões, portanto, o uso do constructo AM tem o objetivo de extrair o conhecimento de bases de dados imobiliárias urbanas em leilão e compor o terceiro eixo teórico, pois . Por fim, o constructo habitação, restringe a categoria no âmbito imobiliário, porém a ativos tangíveis imobilizados.

No Apêndice A, é apresentado o estudo bibliométrico que apoiou a RSL.

A Figura 3 apresenta o modelo da RSL proposta pelos autores Kitchenham e Charters (2007).

Figura 3 - Resumo da Revisão Sistemática da Literatura



Fonte: Adaptado de KITCHENHAM e CHARTERS (2007).

A Figura 3 apresenta que a SLR possui 3 fases em 10 atividades necessárias para que a pesquisa seja realizada e documentada. A vantagem da realização de uma SLR em uma pesquisa é que ela permite que o estudo tenha maior imparcialidade ao reduzir o viés dos autores (KITCHENHAM; CHARTERS, 2007). Nos próximos itens será aplicada a SLR, passo a passo, até se obter os resultados desejados.

2.1 FASE 1 - PLANO DE REVISÃO

De acordo com Kitchenham e Charters (2007) no plano de revisão, a etapa inicial ou a atividade 1 consiste em especificar questão ou questões de pesquisa. A RSL exige a especificação das questões de pesquisa e do método relacionando-as ao objetivo da realização da SLR. As questões de pesquisa são elementos-chave na condução desse método e, para tal, devem ser cuidadosamente selecionadas.

Segundo os mesmos autores, o mais importante nas revisões sistemáticas é conseguir fazer a pergunta correta, que deve ser significativa e principalmente importante não só para a área acadêmica, mas para a prática. A pergunta deve levar mudanças não somente no âmbito acadêmico, mas para as práticas atuais.

Toda via, algumas revisões sistemáticas podem fazer perguntas que são de interesse primordial para os pesquisadores e não tanto para os profissionais. Tais revisões fazem perguntas, que visam identificar e determinar atividades de pesquisa, que pode se encaixar no corpo atual de conhecimento (KITCHENHAM; CHARTERS, 2007).

- QP1: como classificar liquidez imobiliária urbana em leilões através de AHI, apoiando as decisões com MPID?

Após a definição da questão de pesquisa, a SLR propõe a atividade 2, que é desenvolver um protocolo de revisão, onde segundo Kitchenham e Charters (2007) deve conter os métodos que serão usados para realizar uma revisão sistemática específica. Um protocolo pré-definido é necessário para reduzir a possibilidade de viés do pesquisador. Dessa forma, são eleitos alguns critérios de inclusão e critérios de exclusão que ajudam o pesquisador a avaliar os trabalhos e filtrá-los de modo que permaneçam somente os que fazem sentido para as questões de pesquisas propostas.

Foram definidos os critérios de inclusão (CI) e exclusão (CE) apresentados no Quadro 2, vale ressaltar que aqui não foi utilizado o operador lógico “ou” para expandir as buscas visto que o operador “e”, limita ainda mais as buscas realizadas.

Quadro 2 - Protocolo de revisão

Critérios de Inclusão	CI1	Adoção de trabalhos com conceitos aderentes ao tema pesquisado como: liquidez imobiliária, leilão, tempo de exposição, preço, abordagem computacional inteligente, instituições bancárias e habitação.
Critérios de Exclusão	CE1	Remoção de duplicatas por DOI e títulos no Excel, somente artigos, por período compreendido no intervalo de 2005 a 2020 e por categorias.
Critérios de Exclusão	CE2	Remoção por temas como: Fundos de investimento, crédito, meio ambiente, economia, mercado de vinhos, saúde, fusões, crise de mercados, empreendedorismo, políticas urbanas e falência de empresas.

Fonte: Autor.

Na **atividade 3 da SLR**: “**Validar o protocolo de revisão**, Kitchenham e Charters (2007) afirmam que as questões básicas podem ser adaptadas de forma a auxiliar a avaliação do protocolo de revisão.

Para a validação do protocolo foram levantados os seguintes pontos que devem ser verificados e validados:

- Os resultados da pesquisa são apropriadamente derivados das questões de pesquisa.
- Os dados a serem extraídos abordarão adequadamente as questões de pesquisa.
- Os procedimentos de análise serão realizados de forma a conseguir responder adequadamente às questões de pesquisa.

Os experimentos computacionais estão voltados a aplicação e desenvolvimentos de uma AHI com dados reais e inéditos coletados do mercado no segmento de leilões de modo a responder à questão de pesquisa proposta. Em virtude dos argumentos apresentados e dos procedimentos apresentados, é possível validar o presente protocolo de revisão.

2.2 FASE 2 - CONDUÇÃO DA REVISÃO

A condução da revisão ocorre a partir do que foi definido anteriormente no protocolo de revisão. Na atividade 4: “Identificação das pesquisas relevantes”, conforme Kitchenham e Charters (2007) é realizada uma busca preliminar em bibliotecas e bases de dados com a intenção de refinamento iterativo dessa pesquisa através da seleção de *strings* de busca.

A revisão sistemática da literatura foi realizada com base em um conjunto de palavras-chave, relacionadas em *strings* conforme Quadro 1:

a) String 1: “*Real estate*” para “Imóveis”;
b) String 2: “*Liquidity*” para “liquidação”;
c) String 3: “*Auction*” para “Leilão”;
d) String 4: “*Machine learning*” para “Aprendizado de Máquina”;
e) String 5: “*Housing*” para “Habitação”.

A partir dos constructos levantados, foram realizadas 6 pesquisas distintas nas bases de dados escolhidas, conforme o Quadro 3 abaixo.

Quadro 3 - Relação resultados por consulta nas bases de dados sem filtros

Consulta	***String***	***Web of Science***	***Scopus***	**IEEE *xplore***
1	1 e 2	371	350	6
2	1, 2 e 3	3	1	0
3	1, 2 e 4	2	2	0
4	2, 4 e 5	2	1	0
5	2 e 5	461	333	3
6	2, 3 e 5	5	1	0
Total	**1.541**	**844**	**688**	**9**

Fonte: Autor.

As buscas para compor essa dissertação foram iniciadas em 18 de setembro de 2019, tendo como última busca dia 1 de novembro de 2020.

Importante ressaltar que, quando se efetua a pesquisa com a combinação das *strigns* com a *string* 3 referente a leilão, retornam poucos trabalhos acadêmicos. Por outro lado, a bibliometria contida no Apêndice A, retornou apenas 1 trabalho na área da computação quando se estuda e procura liquidação imobiliária.

Em relação à Atividade 5: “Seleção de estudos preliminares” foi realizada a escolha das pesquisas nas bases de dados a partir dos critérios definidos no protocolo de revisão e relacionados à questão de pesquisa. A pesquisa foi realizada com base nos constructos ou palavras chaves listadas.

Para tanto, foi obtido um universo total de artigos de 1.193 por conta das 348 duplicatas removidas, mostrando que a base da *Scopus* contribuiu pouco com artigos novos, bem como a IEEE *Xplore*. A concentração maior está na base da WoS.

É de suma importância que a condução da SLR seja documentada da melhor forma possível para que se torne transparente e replicável e que os leitores possam avaliar o nível de detalhamento da pesquisa (KITCHENHAM; CHARTERS, 2007).

A maior parte das consultas realizadas levaram em consideração o constructo “Imóveis”, considerando-se este constructo como pilar central da pesquisa, justificando a correlação com os demais constructos.

Em relação à Atividade 6 “Garantir a qualidade do estudo”, de acordo com KITCHENHAM e CHARTES (2007), a qualidade do estudo é a fase na qual os critérios adotados no protocolo de revisão são aplicados aos artigos selecionados pela fase de seleção de estudos preliminares. Para isso considerou-se os mesmos critérios definidos no trabalho de Salvetti (2019), e que se enquadraram nesse estudo:

I. O estudo está baseado na pesquisa?
II. Os dados foram coletados conforme os objetivos do estudo?
III. A análise dos dados foi suficientemente rigorosa?
IV. O estudo é relevante considerando práticas ou pesquisas?

Segundo Salvetti (2019) para que o estudo seja ainda mais rigoroso, quanto à qualidade científica pode ser acrescida aos mesmos três outras questões que cobrem e que foram utilizadas no crivo da seleção dos artigos resultante da pesquisa:

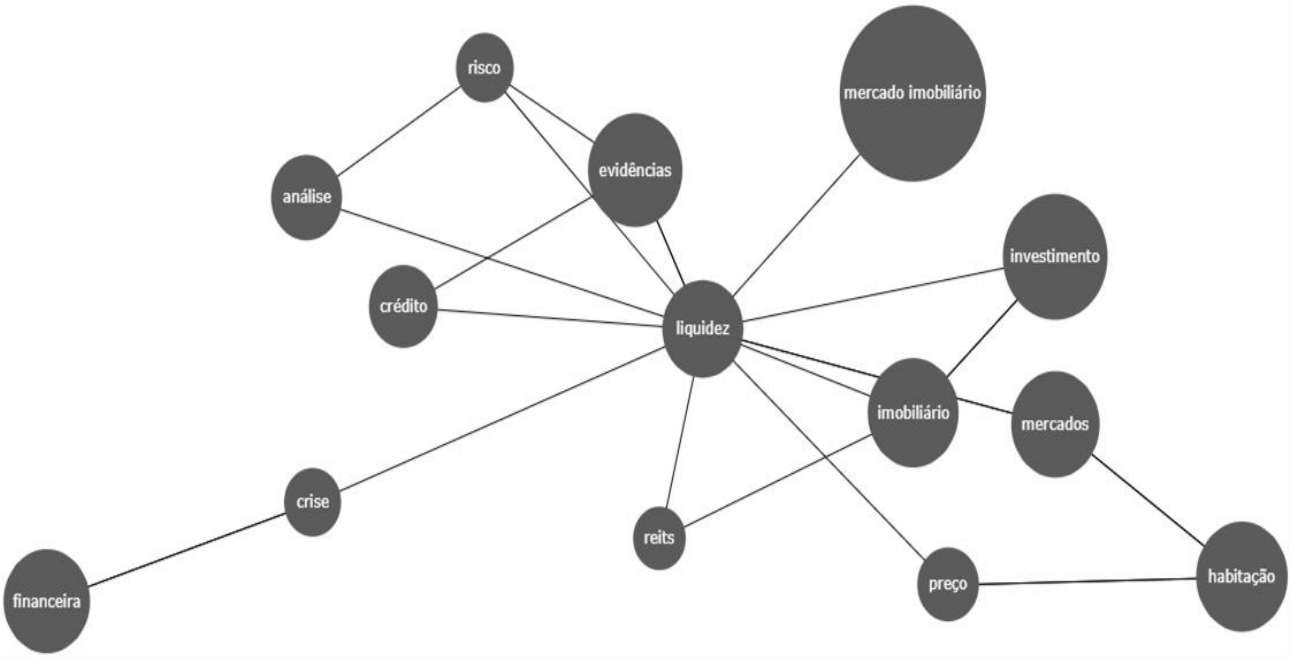
I. Rigor: Uma abordagem completa foi aplicada aos métodos de investigação no estudo?
II. Credibilidade: Os resultados são apresentados de forma significativa?
III. Relevância: Os resultados são uteis para a indústria de software e a comunidade científica?

A Atividade 7: “Extrair dados requeridos”, se deu a partir do registro das informações relevantes nos estudos aprovados pela fase de qualidade do estudo e que foram selecionados para uma análise mais profunda. Para essa dissertação, utilizou-se uma planilha eletrônica para a elaboração de um modelo para obter as informações pertinentes aos estudos. Essa forma foi útil para sintetizar os conhecimentos em um único ponto, conforme pode ser observado no apêndice A.

Segundo Kitchenham e Charters (2007), após a extração, ocorre a atividade 8: “Sintetizar os dados”, que é uma síntese qualitativa, quantitativa e descritiva dos resultados que foram encontrados de forma a se ter uma caracterização inicial dessa seleção. A sintetização foi realizada de forma descritiva (síntese qualitativa).

Os artigos eliminados pelos critérios de exclusão ligados a liquidez, e não relacionados com imóveis, tratam de mercado de ações imobiliárias, rendimentos de mercado, *Reits* (ativos no mercado de ações), crise, carteiras imobiliárias, decisões políticas, estabilidade macroeconômicas, transações imobiliárias, recessão de mercado, política monetária, seguros de mercado e falência como mostra a Figura 4 abaixo feita no *software* Sobek que mostra conceitos em formas de grafos.

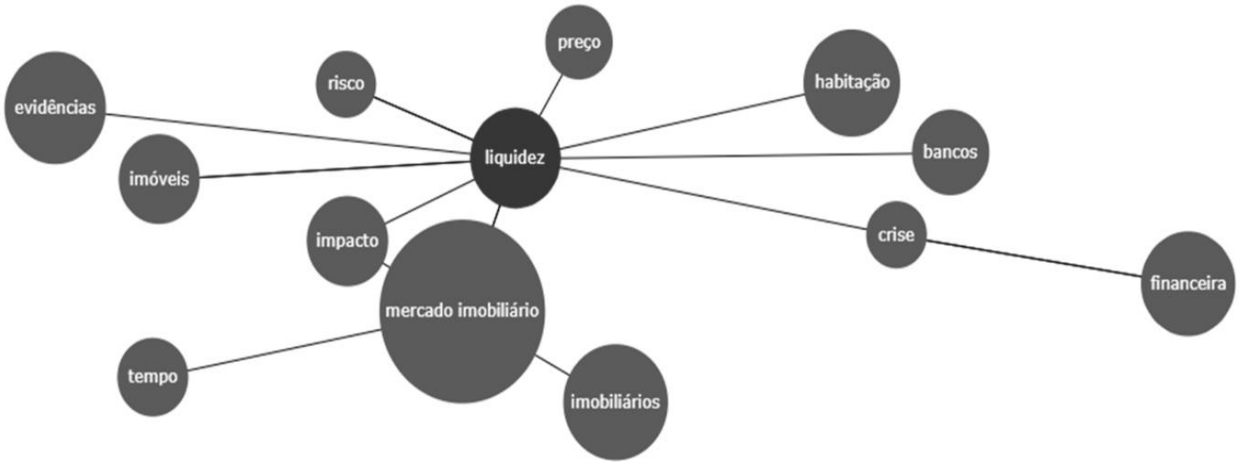
Figura 4 - Áreas abordadas e eliminadas da RSL



Fonte: Autor.

Os trabalhos selecionados pelos critérios de aderência ligados a liquidez, relacionam-se com: imóveis, risco, bancos, exposição no mercado, estatística convencional, indicadores de liquidez e de mercado, como mostra a Figura 5 abaixo.

Figura 5 - Áreas abordadas e selecioandas da RSL



Fonte: Autor.

A Revisão Sistemática da Literatura (RSL), sobre liquidez imobiliária, mostra que a abordagem conjunta de ameaças e oportunidades é um tema que configura uma oportunidade de pesquisa. Uma vez que tem sido pouco explorado na literatura, sobretudo com aplicação de técnicas inteligentes em uma AHI para imóveis urbanos situados em condições de leilão, onde a velocidade da venda difere daquela praticada pelo mercado de imóveis convencional.

2.3 FASE 3 - REVISÃO DA DOCUMENTAÇÃO

Segundo Kitchenham e Charters (2007), a Atividade 9: “Escrever o relatório da revisão”, significa integrar os diferentes estudos que incluam resultados e conclusões em que os pesquisadores que os realizaram usaram conceitos ou termos sutilmente diferentes. Nessa atividade todos os estudos de grande importância já foram identificados nas etapas anteriores e aqui são sintetizados, visando responder as questões bibliométricas e as questões de pesquisa propostas.

A seguir são descritas as respostas para as perguntas bibliométricas apresentadas na Atividade 1 dessa SLR.

●QP1: como classificar liquidez imobiliária urbana em leilões através de AHI, apoiando as decisões com MPID?

Verificando-se o retorno das buscas realizadas observa-se que não se tem ainda nenhum artigo científico que consiga responder a essa questão de pesquisa. O conceito DOM, empregado no trabalho de Hengshu *et. al.* (2016), é uma métrica importante da liquidez do mercado imobiliário.

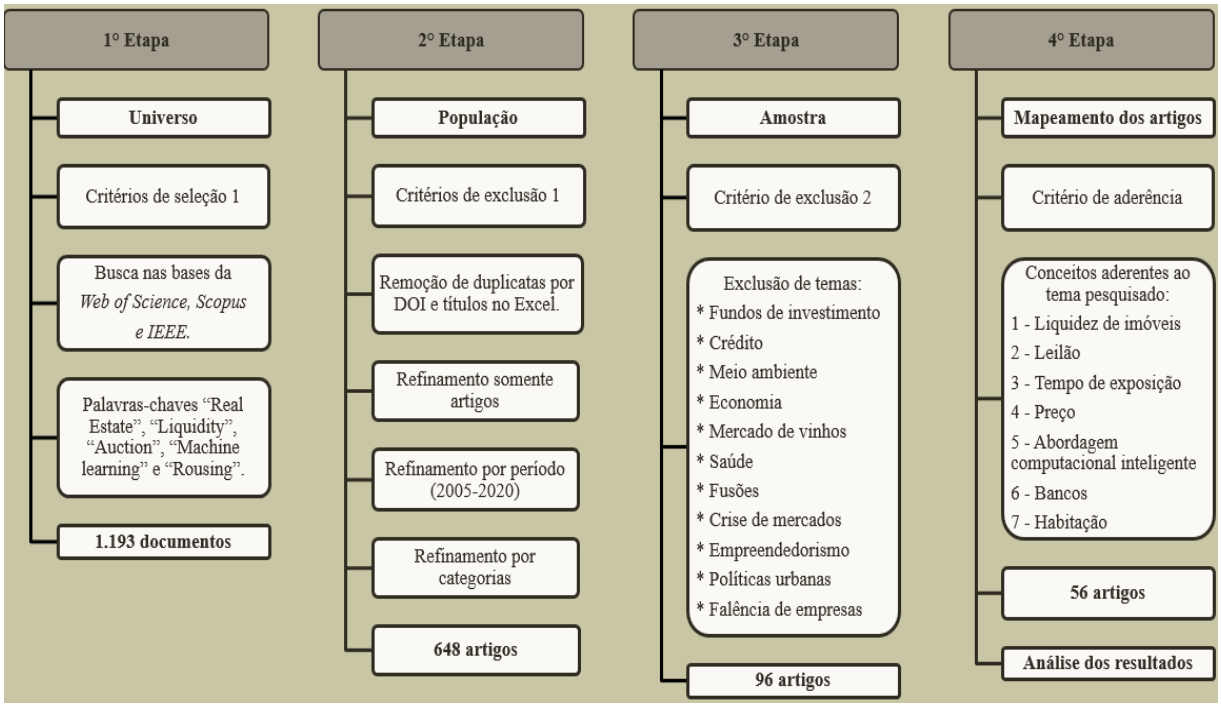
Porém no mercado tradicional. O DOM é um indicador útil para os potenciais compradores avaliarem a liquidez de um imóvel, no entanto, é muito desafiador medir o DOM, pois há uma variedade de fatores que podem impactar o DOM em uma propriedade. O DOM mensura o tempo de exposição desses imóveis no mercado de imóveis tradicional.

Tendo como base a afirmação nos trabalhos dos autores e também uma grande quantidade de artigos retornados nas bases, entendeu-se que é um tema válido para esse trabalho e para sua continuação.

Segundo Kitchenham e Charters (2007), a atividade 10: “Validar o relatório”, quando se tratar de artigos de periódicos, esses serão revisados por pares como uma questão natural. Especialistas revisam teses de doutorado como parte do processo de exame.

A Figura 6 abaixo, sintetiza e ilustra o *design* aplicado a essa RSL em 4 etapas de revisão dos artigos encontrados.

Figura 6 - Design da RSL



Fonte: Autor.

A primeira etapa consistiu na obtenção do universo de trabalhos acadêmicos que abordassem a temática liquidez em imóveis. Detalhados um pouco mais a frente, a segunda e terceira etapas, iniciaram-se no processo de exclusão de trabalhos não aderentes ao tema dessa dissertação. Por fim, na quarta etapa, também detalhada mais à frente, foram selecionados os trabalhos com maior aderência ao tema abordado nesse trabalho. Reduzindo a amostra de 96 artigos para 56 artigos.

Portanto, o procedimento acima mostra que se trata de um tema relevante que merece atenção da comunidade científica, por retornar poucos trabalhos com essa temática. Por outro lado, foram encontrados autores importantes na RSL, que por sua vez, compõe os eixos teóricos abordados no modelo conceitual apresentado e que fornecem suporte direto aos principais eixos dessa pesquisa acadêmica.

Klemperer (1999) teoriza pela primeira vez o fenômeno leilão, onde um enorme volume de recursos é transacionado por meio desse mecanismo. Entre eles, destacando-se os imóveis.

Gan (2013) analisou que esse fenômeno depende da oferta e demanda de mercado, regidas por regulamentos específicos, expostos de maneira simples, rápida, acessível e legal.

Tang e Ren (2008), conceituaram liquidez imobiliária como a capacidade de converter imóveis em dinheiro. Através do conceito de liquidez, Cheng *et al.* (2008) e Xin He *et al.* (2017) descreveram a fórmula para descobrir a relação teórica entre preços de imóveis e seu tempo exposto no mercado, *Time On Market* (TOM), utilizando métricas estatísticas tradicionais de risco e retorno imobiliário.

Hengshu *et al.* (2016) abordam o conceito DOM e liquidez de imóveis. No entanto, no mercado convencional ou tradicional, o conceito DOM é um importante indicador de liquidez de imóveis.

Em contrapartida, Goldschmidt e Passos (2005), introduzem o conceito da combinação de técnicas inteligentes para gerar as chamadas arquiteturas híbridas.

Logo, verifica-se que muitos autores abordam liquidação no mercado de imóveis tradicional. Entretanto, nenhum autor abordou liquidação no mercado de leilões, sobretudo, empregando AHI para automatizar o processo, diminuindo o viés e dando suporte as decisões.

3 REFERENCIAL TEÓRICO

Com base nos objetivos propostos para essa dissertação, o referencial teórico tem finalidade de orientar e apoiar esta pesquisa. Com esta premissa serão dispostos conceitos abordados na literatura acadêmica sobre o assunto. Inicialmente tem-se no subcapítulo 3.1 Mercado Imobiliário no setor de leilões, 3.2 Descoberta de conhecimento em bases de dados, 3.3 Tarefas do KDD, 3.4 Conceito de Decisão, Risco e Incerteza, 3.5 Matrizes de Probabilidade e Impacto, 3.6 Matrizes de Probabilidade e Impacto Duplas, 3.7 Quadro com os principais conceitos do Referencial Teórico.

3.1 MERCADO IMOBILIÁRIO NO SETOR LEILÕES

A (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS - ABNT, 2020) define e estabelece o conceito de imóvel e liquidação nos itens 1 e 2 abaixo:

1. Imóvel segundo (ABNT, 2020, p.4), é um “bem constituído de terreno e eventuais benfeitorias a ele incorporadas. Pode ser classificado como urbano ou rural, em função da sua localização, uso ou vocação”;

2. Liquidação forçada segundo (ABNT, 2020, p.4), trata-se de uma “Condição relativa à hipótese de uma venda compulsória ou em prazo menor que o médio de absorção pelo mercado”.

O setor imobiliário estimula o consumo e favorece investimento das empresas que por consequência, impulsionam a atividade econômica e a geração de empregos no âmbito da construção civil. Um levantamento do Sindicato da Construção Civil do Estado de São Paulo (SINDUSCON-SP) mostra que as vendas de imóveis residenciais seguem em alta em maio de 2020, apesar da pandemia acometida pelo Covid-19 (MARKO, 2020).

Deste modo, o financiamento imobiliário cresce 8,2% em maio. O financiamento imobiliário no Brasil com recursos da poupança SBPE (Sistema Brasileiro de Poupança e Empréstimo) atingiu 7,13 bilhões de reais em maio de 2020, alta de 8,2% ante mesmo mês do ano passado. Nos primeiros cinco meses do ano, os empréstimos destinados à aquisição e à construção de imóveis avançaram 23,2%, atingindo 34,1 bilhões de reais (ABECIP, 2020).

Por outro lado, foi visto que a inadimplência desses empreendimentos faz o setor bancário acionar seus seguros. Deste modo, muitos imóveis acabam sendo penhorados, ou seja, é feito um pedido através de ação judicial, em que o credor cobra formalmente uma dívida. Se o devedor não tem dinheiro em espécie suficiente para pagar o valor atrasado, a justiça pode apreender bens que irão a leilão para saldar os débitos.

Nesse sentido, a mecânica do setor de leilões é diferente do mercado de imóveis tradicional e envolve uma série de riscos às decisões. Entre eles, o valor de referência a ser ofertado, que facilite a liquidação ou arremate em tempo inferior praticado pelo mercado de imóveis convencional. E, por outro lado, o comprador também está sujeito a decisões em condições de risco, uma vez que é incerto o que outros interessados irão oferecer. Sendo assim, é difícil estabelecer um limite para essa negociação.

O fenômeno liquidez observado é apoiado pela Teoria de Leilões (Klemperer, 1999), onde um enorme volume de recursos é transacionado por meio de leilões. Entre eles, destacando-se os imóveis. Nesse sentido, observa-se que muitos bancos empregam esse método para saldar as dívidas e recuperar o patrimônio. Em vários países, empregam-se leilões para privatizar estatais, vender concessões, direitos de exploração de petróleo, etc. No Brasil, não é diferente.

Segundo a Almeida (2020), o Banco do Brasil e o Banco Santander colocaram à venda em maio de 2020, 1.464 imóveis com até 70% de desconto até o final do mês. Os imóveis foram divididos entre casas (48%), apartamentos (40%), entre outros e os valores variam de R$ 30.000,00 a até pouco mais de R$ 9.000.000,00.

A região do País com mais imóveis para venda foi a sudeste, onde o desconto máximo foi de 60%. Os imóveis estavam 100% quitados e sem nenhum tipo de dívida para quem compra. Todos imóveis foram divulgados também no website do Tribunal de Justiça do Estado São Paulo.

Para melhor entender o mercado de leilões, um pequeno grupo de conceitos básicos se faz importante nesse momento.

I. Leiloeiro - Parte responsável pela condução de todos os procedimentos legais do leilão. Geralmente um terceiro, totalmente isento de interesses financeiros nesse tipo de transação, que se coloca a serviço do proprietário do bem a ser leiloado e dos potenciais compradores (CASTRO, 2003).

II. Arrematante - Parte que oferece lances em um leilão. Em geral, trata-se do potencial comprador (CASTRO, 2003).

III. Leilão Inglês - Também conhecido de leilão aberto, oral ou de lances ascendentes. Nesse tipo de leilão, o arrematante deve dispor um lance maior do que o atual (KLEMPERER, 1999).

IV. Leilão Holandês - Funciona de modo inverso ao leilão inglês. No entanto, o leiloeiro fixa um preço inicial alto e progressivamente diminui o valor a fim de que se apresente um lance (KLEMPERER, 1999).

V. Leilões eletrônicos - Os chamados leilões *on-line* ou leilões eletrônicos tem sido popularizado pela Internet nos últimos tempos por oferecer bens com preços mais baixos do que o praticado pelo mercado convencional (CASTRO, 2003).

Por outro lado, no setor de leilões, a venda e a compra de ativos patrimoniais dependem da oferta e demanda de mercado, regidas por regulamentos específicos, expostos de maneira simples, rápida, acessível e legal. Essa modalidade de negociação é empregada por órgãos públicos em vários países do mundo. Porém, o comprador tem liberdade de escolha e o vendedor obrigação de vender (GAN, 2013).

O setor de leilões possui diversos sistemas comerciais de liquidação ou arremate imobiliário através dos diferentes tipos de leilão. No entanto, o uso de softwares para automatizar tarefas e processos em um leilão é relativamente recente e conta com um número reduzido de sistemas disponíveis no mercado (SANDHOLM, 2000).

Lopes (2016) descreveu e implementou uma plataforma de gestão de processos de insolvência em leilões, no apoio e expansão a dispositivos móveis. Visando essa lacuna no mercado, o modelo desenvolvido permite a gestão, registo, inventariação dos respetivos bens e controle de vendas.

Através do conceito de métricas de liquidez, Cheng *et al.* (2008) e Xin He *et al.* (2017) descrevem a fórmula para descobrir a relação teórica entre preços de imóveis e seu tempo exposto no mercado, pelo meio do mesmo conceito DOM, no entanto, com nomenclatura diferente, *Time On Market* (TOM), utilizando métricas estatísticas tradicionais de risco e retorno imobiliário.

Logo existe a necessidade de análise desses dados de liquidez imobiliária, disponíveis nas bases de dados do mercado de leilões. Sistemas de apoio à decisão que permitam considerar os principais fatores que determinam a “atratividade” dos investimentos imobiliários em contextos urbanos competitivos, por exemplo, no contexto de leilões (DEL GIUDICE *et al.,* 2019).

3.2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

O mecanismo de liquidação de bens imóveis envolve processo onde os tomadores de decisão devem considerar diversos fatores, para redução das incertezas nesse setor. Esse processo pode ser realizado por meio de um sistema de apoio à decisão. Porém, para que esses sistemas ofereçam suporte adequado à decisão, é necessário analisar dados reais, disponíveis nas bases de dados do mercado de imóveis (MCGREAL *et al.,* 2016).

Entretanto, as bases de dados do mercado imobiliário possuem grandes volumes de dados, dentre os quais existem diversos padrões difíceis de descobrir através de métodos tradicionais como planilhas de cálculo e relatórios informativos operacionais. Desta forma, os sistemas de descoberta de conhecimento *Knowledge Discovery in Data Bases* (KDD) surgem como uma solução para a extração do conhecimento. Mesmo utilizando um KDD, tal atividade pode continuar sendo extremamente difícil devido à grande quantidade de dados que deve ser processada para a tomada de decisão (SASSI, 2006).

Considerando que as bases de dados costumam ser grandes, Da Silva *et al.* (2017) estabelece que o conhecimento pode ficar oculto. Portanto, é necessário realizar uma tarefa de busca detalhada, denominada tarefa de "mineração", associada a um processo analítico, sistemático e, até onde possível, automatizado.

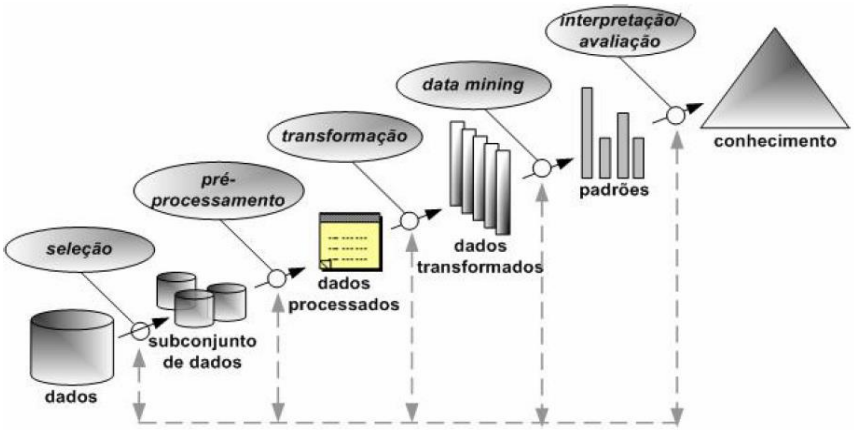
Data Mining (DM) pode ser definido como um processo automático de análise e exploração de grandes bancos de dados para descobrir padrões relevantes que aparecem nos dados e que são importantes para apoiar as decisões. Em geral, esse processo pertence a uma subárea do conhecimento, lecionado em cursos de informática ou engenharia, e utiliza conceitos de IA, AM, estatística e bancos de dados (DA SILVA *et al*., 2017).

Embora os termos KDD e DM sejam altamente relacionados, eles não podem ser entendidos como sinônimos. Outros termos muito populares são confundidos com o termo "mineração de dados", como ETL (*Extraction, Transform and Loading*), DW (*Data Warehouse*), OLAP (*Online Analytical Processing*) e Estatística Descritiva. Na verdade, a mineração de dados deve ser entendida como uma etapa do processo de descoberta do conhecimento no banco de dados, e este pode conter os demais termos (DA SILVA *et al.,* 2017).

Portanto, o objetivo de descobrir o conhecimento em um conjunto de dados é encontrar os padrões neles contidos e apresentá-los de forma a promover a absorção de conhecimento adquirido.

A Figura 7 a seguir, mostra as fases do processo do KDD, na aquisição de conhecimento.

Figura 7 - Fases do processo de KDD



Fonte: Fayyad *et al.* (1996).

3.2.1 Seleção dos dados

Na etapa de seleção dos dados, após determinar o objetivo para o processo de extração de conhecimento, busca-se dados e seus atributos variáveis relevantes de um ou mais banco de dados específicos, alinhados a estratégia de captação e de mineração, de modo estruturado ou não.

3.2.2 Pré-processamento dos dados

Grandes bancos de dados são extremamente suscetíveis a ruídos, valores ausentes e inconsistências. Dados limpos e compreensíveis são o requisito básico para uma mineração de dados bem-sucedida. Portanto, o pré-processamento de dados garante a qualidade dos dados selecionados e coletados (SASSI, 2006).

O pré-processamento inclui verificar a consistência das informações, corrigir possíveis erros, remover valores vazios, redundantes, valores discrepantes ou *outliers*. Esses dados geralmente ocorrem devido a erros humanos, ou porque a informação não está disponível no momento do levantamento dos dados, gerando informações eventualmente contraditórias. (LIU *et al.,* 2002).

Em termos de desempenho do algoritmo minerador, outra técnica de pré-processamento muito disseminada no âmbito acadêmico é a redução de dados. A execução desta etapa corrige o banco de dados, eliminando consultas desnecessárias que serão executadas pelo algoritmo de mineração, prejudicando assim seu bom desempenho (BATISTA, 2003).

3.2.3 Transformação dos dados

O principal objetivo da transformação dos dados ou estágio de codificação de dados é converter o conjunto de dados original em um conjunto padronizado de uso. Portanto, esta tarefa requer habilidades nesse processo e experiência de analistas de dados (SASSI, 2006).

Os benefícios da transformação consistem em melhorar a compreensão do conhecimento descoberto, reduzir o tempo de processamento do algoritmo de mineração, assim ajudando, o algoritmo a tomar uma melhor decisão. A desvantagem é que a medição da quantidade de conhecimento encontrada é reduzida e os detalhes relevantes sobre as informações extraídas são perdidas.

3.2.4 *Data Mining* (Mineração dos dados)

Mineração de dados é definida em termos de esforços para descoberta de padrões em bases de dados. A partir dos padrões descobertos, têm-se condições de gerar conhecimento útil para um processo de tomada de decisão. Portanto, a mineração de dados é considerada a etapa mais importante no processo de KDD (DA SILVA *et al.,* 2017).

Segundo Berry (2004), *Data Mining* é a exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados para descobrir modelos e regras significativas.

3.2.4.1 Tarefas do KDD

O KDD possui vários métodos de interpretação de dados, chamados tarefas. As tarefas mais comuns são a associação de dados, classificação, regressão e agrupamento. Logo, as técnicas de AM, oriundas da Inteligência Artificial (IA) são de fundamental importância na detecção padrões e permitir a criação de modelos preditivos que venha a gerar conhecimento no apoio a tomada de decisão (FAYYAD *et al.,* 1996).

Fayyad *et al.* (1996) afirmam que a tarefa de mineração de dados é dividida em tarefas preditivas e descritivas. Tarefas preditivas usam os valores dos atributos descritivos para predizer valores futuros ou desconhecidos de outros atributos de interesse. Por outro lado, as tarefas descritivas visam encontrar padrões que descrevam os dados de uma forma que os humanos possam interpretar.

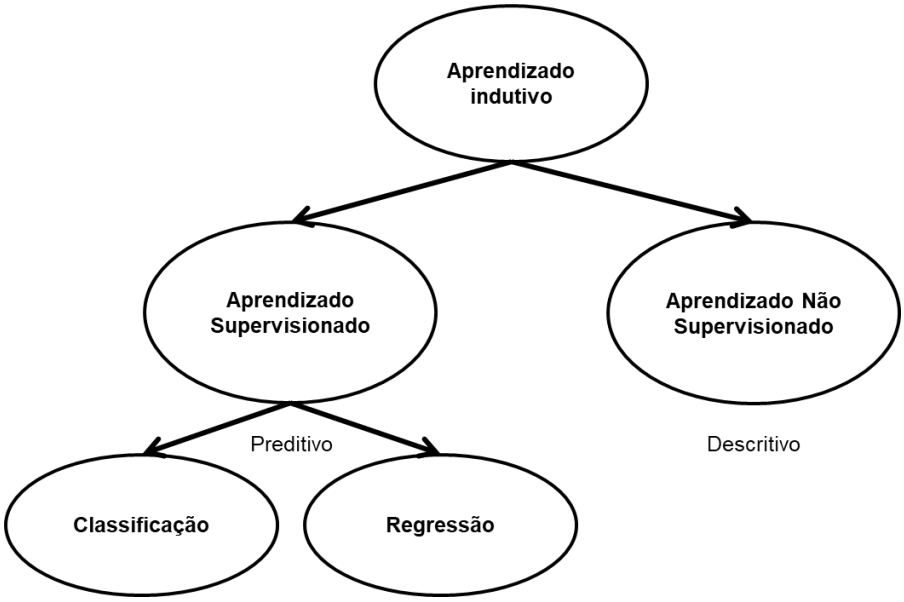
Deste modo, o AM pode ser usado para gerar dados ausentes. Dessa forma, esses sistemas aprendem a como combinar dados de entrada para gerar previsões úteis sobre dados não vistos. Por outro lado, os modelos de regressão predizem valores numéricos, enquanto os modelos de classificação predizem valores nominais (DA SILVA *et al.,* 2017).

Por outro lado, as técnicas de AM empregam um princípio de inferência denominado indução, no qual é possível obter conclusões genéricas a partir de um conjunto particular de exemplos (LORENA; CARVALHO, 2007).

Em outras palavras, a inferência indutiva, também conhecida como viés indutivo, busca a hipótese que melhor se ajusta aos dados de treinamento. O viés define como as hipóteses são pesquisadas no espaço de hipóteses. Uma vantagem de visualizar sistemas de inferência indutiva em termos de viés indutivo é que ele fornece um meio não-processual de caracterizar sua política para generalizar além dos dados observados (MITCHELL, 1997).

Classificação e Regressão são respectivamente, exemplos de dois tipos de aprendizado preditivo conforme Figura 8 abaixo:

Figura 8 - Hierarquia do Aprendizado de Máquinas



Fonte: Adaptado pelo autor de MONARD e BARANAUSKAS (2003).

O reconhecimento de padrões tem como objetivo classificar os dados com base em conhecimentos prévios (preliminares ou dedutivos) ou informações estatísticas extraídas de padrões. Existem dois tipos de reconhecimento, um é o treinamento supervisionado, que usa o conjunto de treinamento para classificar e organizar os dados obtidos de acordo com as categorias existentes e rotuladas; o outro é não supervisionado, que utiliza dados no reconhecimento de categorias em vez de separar os dados com base nas categorias existentes e rotuladas (DA SILVA *et al.,* 2017).

Logo após definir que tipo de tarefa de aprendizagem será usada, supervisionada ou não supervisionada, é necessário avaliar e selecionar o algoritmo, capaz de resolver o problema em questão com a performance desejada.

No aprendizado supervisionado, a técnica inteligente emprega um conjunto de dados, divididos em treino e teste, com instâncias e atributos variáveis, afim de predizer o valor de saída para novos exemplos. Quando um problema possui rótulo com um valor discreto, este é denominado classificação e, quando os rótulos são contínuos, é chamado de regressão (DA SILVA *et al.*, 2017).

O motivo para essa ressalva é que modelos preditivos, a depender de como são gerados, podem levar à manifestação de um fenômeno bastante conhecido, o sobreajuste (do inglês *overfitting*). Portanto, pode-se entender que é um termo usado em estatística para descrever quando o modelo estatístico está em boa concordância com o conjunto de dados observado anteriormente (DA SILVA *et al.,* 2017).

Portanto, o Aprendizado de Máquinas (AM) é uma área da IA que lida com problemas de aprendizado computacional a fim de adquirir conhecimento de forma automática, generalizando dados novos e extração de novos conhecimentos (MONARD; BARANAUSKAS, 2003).

Monard e Baranauskas (2003) definem que o aprendizado indutivo pode ser dividido em Aprendizado Supervisionado (AS) e Aprendizado Não Supervisionado (ANS). O AS é utilizado para classificação dos exemplos em classes predefinidas como resolver problemas preditivos. O ANS é utilizado para agrupamento, para problemas de cunho descritivo.

3.2.4.2 Árvore de decisão (*Decision Tree* - DT)

O professor Quinlan da Universidade de Sydney cunhou pela primeira vez a técnica inteligente árvores de decisão. Sua contribuição foi reconhecida por conta do desenvolvimento dos algoritmos ID3 em 1983 e C4.5 em 1993, que ainda são amplamente usados para gerar árvores de decisão (DA SILVA *et al.*, 2008).

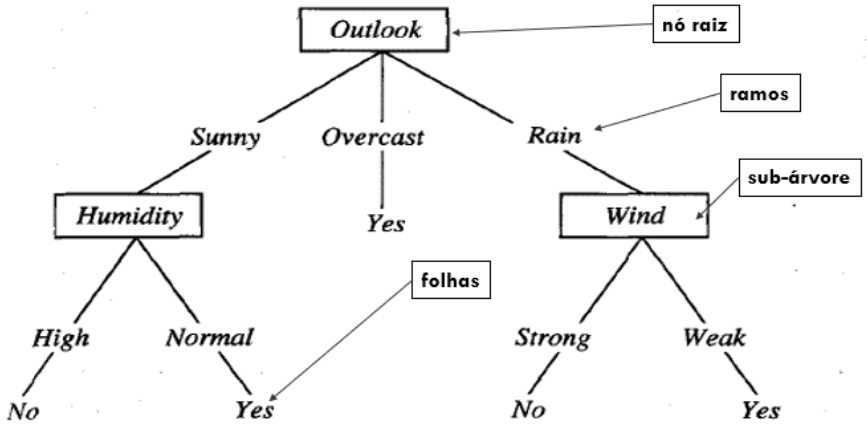
De acordo com o Mitchell (1997), árvore de decisão é um método para aproximar funções-alvo com valores discretos, em que a função aprendida é representada por uma árvore de decisão; emprega conjuntos de regras se-então (*if-then*); utiliza inferência indutiva; é resistente a ruídos nos dados, a falta de dados, utiliza um viés indutivo, ou seja, o método guloso de preferência por árvores menores, que priorizam hipóteses mais simples de ajuste aos dados (Navalha de Occam).

Árvore de decisão é considerado um modelo de classificação e regressão, representado através de uma função-alvo com valores discretos, que se utiliza de treinamento supervisionado para predizer dados futuros. Durante este processo de treinamento, os parâmetros serão inseridos e produzidos. Como resultado, obtém-se vários pequenos problemas, e esses pequenos problemas são resolvidos recursivamente (GAMA, 2004).

No contexto da resolução da tarefa de classificação, uma árvore de decisão representa o modelo capaz de guiar a tomada de decisão sobre a determinação da classe à qual um exemplar pertence (DA SILVA *et al*., 2017).

Logo, árvores de decisão classificam instâncias ordenando-as na árvore de cima para baixo, a partir da raiz até alguma folha, conforme Figura 9; classificando as instâncias inicialmente pelo nó raiz, testando o atributo especificado por este nó; cada nó da árvore especifica o teste de algum atributo da instância e cada ramo partindo de um nó corresponde a um dos valores possíveis dos atributos. Este processo é repetido para a sub-árvore originada no novo nó (MITCHELL, 1997).

Figura 9 - Representação da árvore de decisão



Fonte: Adpatado pelo autor de Mitchell (1997).

O método de decisão em árvore pode ser estendido para a função de aprendizagem de mais de duas opções com valores de saída em cada ramo. Os dados de treinamento podem conter erros e valores ausentes, sendo possível realizar testes estatísticos em cada atributo realiza busca no espaço de hipótese, fornece uma hipótese única, não possui *backtracking* (recuo/volta atrás), robustez a ruídos nos dados e o bias indutivo tem preferência por árvores menores (MITCHELL, 1997).

Portanto, as medidas de seleção de atributos mais frequentemente usadas na indução da árvore de decisão são o critério de Ganho da Informação (QUINLAN; CAMERON-JONES, 1993), o Índice de Gini e entropia (BREIMAN, 1999).

O índice de GINI mede o grau de impureza dos dados. Logo, pode ser usado para medir a pureza de um nó. Quando este índice é igual a zero, o nó é puro. Entretanto, quando se aproxima do valor um, o nó é impuro (AZEVEDO, 2018).

O Índice de GINI é definido pela Equação (1) a seguir:

$$Gini = \sum_{k \neq k'} \hat{p}_{mk}\, \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\,(1 - \hat{p}_{mk}) \tag{1}$$

A entropia caracteriza a falta de homogeneidade dos dados de entrada em relação à sua classificação (AZEVEDO, 2018). Desta forma, a entropia máxima é igual a 1, quando o conjunto de dados é heterogêneo.

A entropia também é utilizada como média de impureza e é dada pela Equação (2) abaixo:

$$S = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{2}$$

Simplicidade para compreensão e interpretação dos dados sem necessidade de pré processamento, boa capacidade de lidar com dados numéricos quanto categóricos, possibilidade de validar um modelo através de testes estatísticos, bom desempenho em grandes conjuntos de dados em um tempo curto, a árvore de decisão traz consigo inferências para aplicação em múltiplas áreas, como na área médica, de avaliação de crédito nos setores financeiros (MITCHELL, 1997). Porém, o aprendizado de árvores de decisão pode criar árvores muito complexas que não generalizam bem os dados.

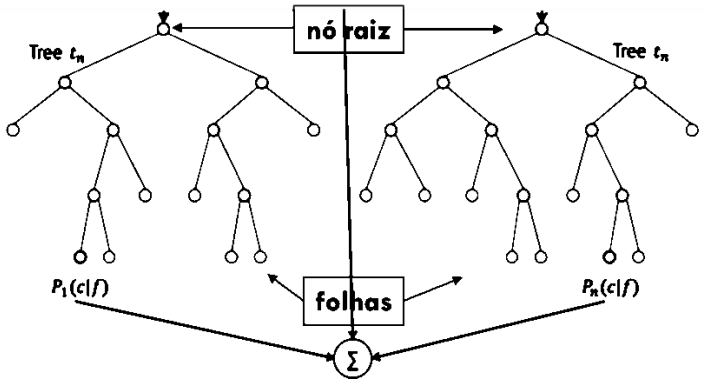
3.2.4.3 Florestas aleatórias (*Random Forest* - RF)

Segundo Breiman (1999), o classificador *Random Forest* (RF) é um método de aprendizagem desenvolvido para classificação, regressão e outras tarefas, realizado através da construção de um número ilimitado de árvores de decisão durante o treinamento e geração da classe com o método ensacamento denominado *Bagging*, que é a combinação dos modelos de aprendizado que majoram o resultado geral.

Breiman (1999) sugere que à medida que o número de árvores aumenta, o erro de generalização sempre converge mesmo sem podar a árvore e o *overfitting*, sendo assim, não se torna um problema. Portanto, na maioria dos casos, o RF evita *overfitting* porque pode lidar com subconjuntos aleatórios de recursos e constrói árvores menores a partir desses subconjuntos. Após o treinamento, mescla-se as sub-árvores. Este método diminuirá a velocidade de cálculo, dependendo de quantas árvores RF construirá.

Os resultados do método de *Bagging*, segundo Singh *et al.*(2020), é produto da melhor combinação de modelos de aprendizado de máquina e melhora dos resultados gerais, conforme mostrado na Figura 10 a seguir:

Figura 10 - Método *Bagging* no classificador *Randon Forest*



Fonte: Adaptado pelo autor de SINGH *et al.* (2020).

O modelo combina, portanto, um número finito de modelos para obter uma melhor performance preditiva. Sendo assim, o RF é um método de aprendizado de conjunto, cujo objetivo é treinar várias árvores de decisão, obtidas a partir de amostras do dataset para fazer predições aproveitando os resultados que mais aparecem em caso de um problema de classificação, ou a média dos valores obtidos em caso de regressão (SINGH *et al.,* 2020).

Ao criar várias árvores e treiná-las todas elas no mesmo dataset, suas predições serão idênticas. Portanto, outro método é dividir o conjunto de dados em várias partes, uma para cada árvore. Dessa forma, cada árvore terá seu próprio conjunto de dados que gerará diferentes previsões. O problema com esse método é que cada árvore quase não tem dados de treinamento. Esse processo é denominado *bootstrapping*, e o novo exemplo é denominado *bootstrap sample* (GISLASON, 2006).

Floresta aleatória se dá pelo fato de ser treinada em diferentes conjuntos de dados, e os recursos são selecionados aleatoriamente, ou seja, os atributos variáveis de cada instância, esses recursos serão incluídos no conjunto de dados, isso é chamado de aleatoriedade de recursos (SINGH *et al.,* 2020).

O objetivo é garantir que a correlação entre as árvores seja baixa. Isso significa que duas árvores de decisão selecionadas aleatoriamente não podem se descrever por meio de um relacionamento linear. Portanto, ao utilizar o *bootstrapping*, o fator é reduzido devido à alta sensibilidade da árvore de decisão, ou seja, uma pequena alteração no conjunto de dados causará uma grande alteração no modelo gerado. Logo as características são selecionadas aleatoriamente, explicando a diversidade das árvores (GISLASON, 2006).

Portanto, os modelos com baixa correlação têm maior probabilidade de não cometer os mesmos erros. A diversificação pode garantir que os erros sejam compensados ou sobrepostos pela correspondência, fornecendo assim um modelo mais poderoso em termos de desempenho.

Como cada árvore usa apenas uma parte dos atributos de entrada em uma floresta aleatória, o algoritmo é consideravelmente mais leve do que o ensacamento convencional com um classificador de tipo de árvore comparável (GENUER, 2010).

Para estimar a precisão do conjunto de teste, as amostras restantes do conjunto de treinamento que não estão no *bootstrap* para uma árvore específica de cada árvore executam validação cruzada. Da mesma forma, a importância das variáveis pode ser estimada trocando aleatoriamente todos os valores das variáveis na amostra fora da troca de cada classificador. Nesse sentido, o método de erro *out-of-bag* é muito alto, indicando a importância das variáveis, porque pode medir os erros de predição de florestas aleatórias, árvores de decisão aprimoradas e outros modelos de aprendizado de máquina que usam *bootstrap* agregado a amostras de dados de subamostras usadas para treinamento (GISLASON, 2006).

Ao limitar o número de variáveis usadas para divisão, a complexidade computacional do algoritmo pode ser reduzida e a correlação entre as árvores também pode ser reduzida. Em última análise, a árvore em RF não será podada, o que reduz ainda mais a quantidade de cálculo. Como resultado, o algoritmo de RF pode processar dados de alta dimensão e usar um grande número de árvores na coleção.

Isso se soma ao fato de que uma seleção aleatória de variáveis para uma partição é projetada para minimizar a correlação entre as árvores no conjunto (GISLASON, 2006).

Poucos trabalhos acadêmicos abordam dois conceitos em termos de ajuste desse modelo. De modo simplificado, o erro devido à falta de complexidade do modelo, correspondente ao *underfitting* (Van Der Aalst *et al.,* 2010), e a variância é o erro devido em excesso de complexidade do modelo, correspondente ao *overfitting* (GENUER, 2010). Existe ainda um erro intrínseco, ou seja, aquele que vem dos dados (MITCHELL, 1997).

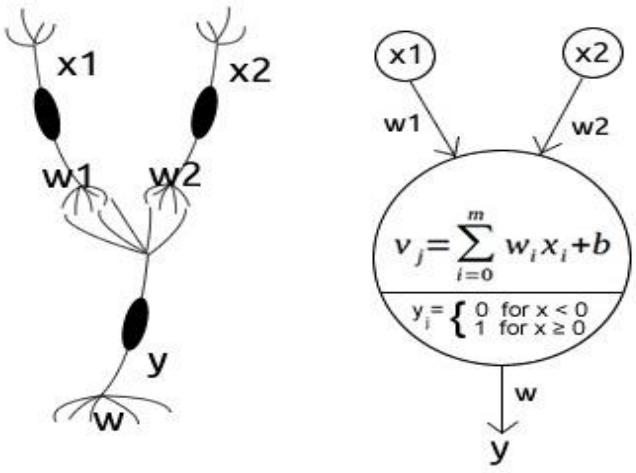
3.2.4.4 Rede Neural Artificial (Multilayer Perceptron - MLP)

Rede neural artificial (RNA) é uma abstração de redes neurais biológicas, originada dos trabalhos de Hebb (1949) e Rosenblatt (1958). Esta simulação é realizada através de um modelo baseado na fisiologia básica dos neurônios biológicos. Pode ser definida como uma estrutura complexa interligada por elementos de processamento que possuem a capacidade de processar dados e converter em conhecimento (GARDNER; DORLING, 1998).

Seu primeiro conceito foi introduzido em 1943, mas ganhou popularidade algumas décadas depois com a introdução de algoritmos de treinamento como o *backpropagation*, que permite a realização de um treinamento posterior para aperfeiçoar os resultados do modelo. Desde que foi desenvolvida, essa técnica vem sendo amplamente utilizada e validada por diversas áreas de pesquisa que pretendem antecipar cenários e acontecimentos, dando suporte à tomada de decisão. (GRÜBLER, 2018).

A relação entre as redes artificiais e biológica é que ambas possuem axônio e dendrito e comunicam-se por sinapses. A representação dessa relação é exibida na Figura 11 a seguir, onde a letra x representa os sinais recebidos e a força sináptica recebida é simbolizada por w. Ambas as redes possuem a capacidade de ajustar a amplitude das sinapses em uma série de camadas interligadas (GRÜBLER, 2018).

Figura 11 - Relação entre uma rede neural biológica e artificial *Perceptron*



Fonte: GRÜBLER (2018).

O *Perceptron* propõe classificar as entradas xi (ou estímulos) em duas classes através de um hiperplano, que funciona como fronteira dos resultados. No caso simples de um espaço em duas dimensões, o hiperplano fica reduzido a uma reta, cuja equação é representada na Equação (3):

$$\sum_{i=0}^{n} x_i + w_0 = 0 \tag{3}$$

A ativação do neurônio artificial é realizada através da função de ativação $\boldsymbol{\varphi}$, que ativa ou não a saída dependendo da soma ponderada de suas entradas.

Enfim, a aprendizagem do *Perceptron* se dá através dos ajustes dos pesos sinápticos. O valor do peso sináptico W(t+1) no instante t+1, será determinado em função do seu valor na iteração anterior wt, conforme na Equação (4) a seguir:

$$w_i^{t+1} = w_i^t + \Delta w_i^t \tag{4}$$

A atualização dos pesos depende do algoritmo de aprendizado utilizado, mas geralmente procura-se a minimização do erro ε_i entre os valores previstos pela rede e as saídas yi desejadas, conforme Equação (5):

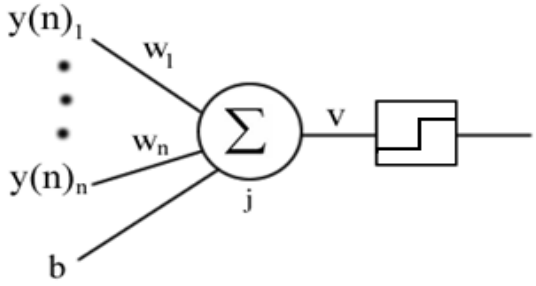
$$\varepsilon_i = \sum \varphi(w_i x)_i - y_i \quad (5)$$

Desta forma, o aprendizado em uma RNA é definido por Haykin (2010), como o ajuste iterativo dos pesos sinápticos, visando a minimização dos erros.

A arquitetura de uma RNA varia com o problema no qual se pretende sua aplicação, e é definida entre outros fatores, como número de camadas, número de nós em cada camada, pelo tipo de conexão entre os nós (feedforward ou feedback) e por sua topologia (KOVÁCS, 2006).

Na Figura 12 abaixo, o neurônio artificial é um *Perceptron* que recebe diversos valores de entradas y(n). Essas entradas multiplicam-se pelo peso da sinapse w e, no final, somam-se formando um conjunto de entrada ξ = ∑ w * y(n). Esse resultado passa por uma função de ativação linear, que será explicada detalhadamente mais à frente, transmitindo a saída v. Quando o valor ξ exceder o limite da função de ativação, o neurônio será ativado e retornará um valor (GARDNER; DORLING, 1998).

Figura 12 - Visão geral do funcionamento de uma Rede Neural Artificial

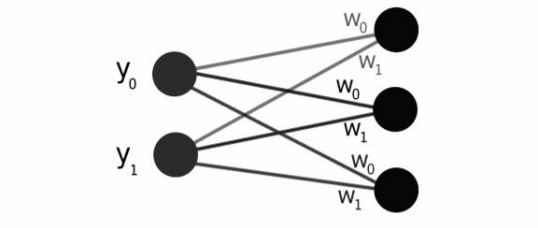


Fonte: GRÜBLER (2018).

Com o intuito de lidar com os problemas não linearmente separáveis, foram adicionadas camadas de neurônio ocultas no modelo de Rosenblatt, formando então a Rede Neural Artificial *Multilayer Perceptron* (MLP).

Essa nova topologia funciona como uma rede *feedforward* (rede progressiva, a saída de um neurônio se conecta com outro neurônio da próxima camada, no sentido esquerda/direita), formada por um conjunto de neurônios denominados “nós”, como demonstrado na Figura 13. A rede possui uma camada de entrada (sem função computacional), uma ou mais camadas ocultas e uma camada de saída. A complexidade da rede MLP se dá pela quantidade de camadas ocultas que houver e a quantidade de neurônios que essas camadas possuírem.

Figura 13 - Sinapses de cada neurônio



Fonte: GRÜBLER (2018).

Cada neurônio recebe todos os valores das entradas, representadas pelo símbolo y, que são multiplicadas pelos pesos sinápticos simbolizados pelo w e somadas entre si junto com uma constante chamada de polarização ou bias, representada pelo símbolo *b* na Equação (6) abaixo:

$$v_j = \sum_{i=0}^{m} w_i y_i + b \tag{6}$$

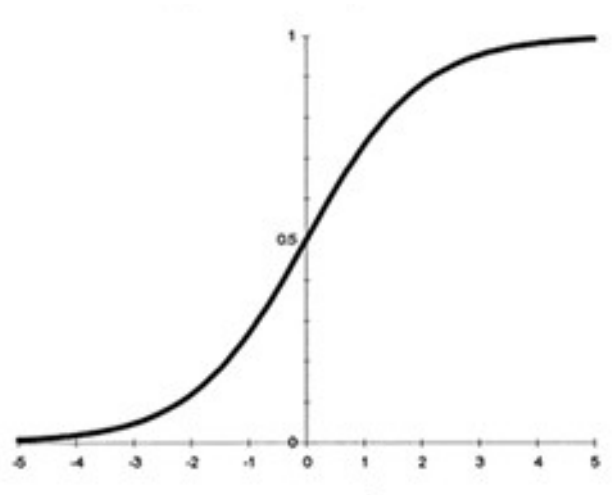
Essa constante possui o papel de centralizar a curva da função de ativação em um valor conveniente. Caso seja positivo, o movimento do gráfico é realizado para a esquerda, diminuindo o valor do eixo x. Porém, caso seja negativo, o movimento do gráfico é feito para a direita, aumentando o valor do eixo x.

A soma ponderada apresentada na Equação (7) abaixo, gera o potencial de ativação que é utilizado para determinar seu valor e propagar para outros neurônios da próxima camada. O objetivo é limitar a amplitude de saída do neurônio, ou seja, o valor obtido no somatório é normalizado através de um intervalo fechado, como [0,1], podendo ser interpretado também como a probabilidade do resultado.

$$A = \frac{1}{1 + e^{-x}} \tag{7}$$

Já em funções não lineares, a técnica logística sigmoidal é uma das mais populares em redes MLP e sua fórmula pode ser visualizada na Figura 14 a seguir:

Figura 14 - Função logística sigmoidal - Não linear



Fonte: GRÜBLER (2018).

Quando uma rede neural artificial é inicializada, os pesos sinápticos recebem valores aleatórios que, quando multiplicados pelos valores recebidos, algumas vezes da camada de entrada e outras de neurônios da camada anterior, não atingem os valores desejados no momento do treinamento. Para corrigir os pesos sinápticos, uma das técnicas populares é a retropropagação, conhecido também como *backpropagation*, que corrige os valores dos pesos pela diferença entre o valor obtido e o esperado (recebido no treinamento) pelo algoritmo.

Em 1986, David Rumelhart e seus colegas introduziram o algoritmo *backpropagation*, possibilitando o treinamento das redes neurais com diversas camadas através da retropropagação. O processo de correção ocorre em dois passos. No primeiro passo, chamado de *feedforward*, é introduzido um valor na camada de entrada e outro na camada de saída. O resultado percorre as camadas internas até que a resposta seja reproduzida na camada de saída. Já no segundo passo, ocorre o aprendizado da rede, comparando o valor obtido com o valor desejado, através da Equação (8), que subtrai o valor esperado (rr) do valor obtido (ro) do neurônio j e, caso o resultado não esteja no padrão aceitável, a rede calcula o erro e propaga a correção para as demais camadas internas até a entrada, ajustando os seus pesos sinápticos.

$$e_j = ro_j - rr_j \tag{8}$$

A técnica *backpropagation* propaga os sinais de erro na direção oposta ao *feedforward*, camada a camada, computando os gradientes locais de cada neurônio, visualizado na imagem acima. Esse processo permite que sejam executadas correções nos pesos sinápticos através da Equação (9). Nessa equação, o símbolo Δwji representa o novo valor do peso sináptico i do neurônio j, já y simboliza o sinal de entrada da célula nervosa j, pertencente ao neurônio i, η é a razão de aprendizado (um valor definido no momento da configuração da rede) e o δ é o gradiente local.

$$\Delta w_{ji} = \eta \delta_j y_i \tag{9}$$

Para utilizar a Equação (9) de ajuste dos pesos e propagar as correções entre as sinapses, é necessário calcular o gradiente local. Nos neurônios da camada de saída, o gradiente local é o sinal de erro, multiplicado pela derivada de sua função de ativação, simbolizada por φ(v), conforme a Equação (10). Após a correção na camada de saída, utiliza-se a Equação (11) para obter o gradiente local dos neurônios predecessores, escondidos na estrutura da rede. Esse processo de correção é recursivo e termina apenas quando o algoritmo de treinamento chega à primeira camada, a de entrada.

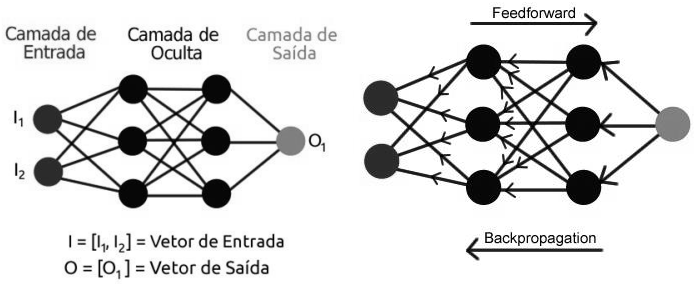
$$\delta_j = \varphi_j(v_j) e_j \tag{10}$$

$$\delta_j = \varphi_j(v_j) \sum_k \delta_k w_{kj} \tag{11}$$

Na equação (11), que visa obter os gradientes locais de neurônios internos da rede, a derivada da função de ativação (função logística sigmoidal, tangente hiperbólica , ou outras) representada pelo símbolo φ(v), é multiplicada pelo somatório de outro gradiente local, simbolizado pelo δ, obtido através da retropropagação advinda dos neurônios posteriores ao j, sendo, nesse exemplo, o neurônio k vezes os pesos sinápticos w entre os neurônios j e k.

O princípio básico do algoritmo MLP é o cálculo dos erros nas camadas intermediárias realizado por retroalimentação, possibilitando desta forma o ajuste dos pesos proporcionalmente a cada um dos valores das conexões entre camadas, conforme Figura 15 abaixo:

Figura 15 - Modelo de uma Rede Neural Artificial MLP



Fonte: Adapatado pelo autor de GRÜBLER (2018).

Para Haykin (2010), um dos algoritmos de treinamento mais usados nas MLP é o de retropropagação do erro (*error backpropagation*) cujo funcionamento se dá da seguinte maneira:

- Apresenta-se um padrão à camada de entrada da rede,
- O padrão é processado camada por camada de forma recursiva até que a camada de saída forneça a resposta processada, fMLP, conforme a Equação (12) abaixo:

$$f_{MLP} = \varphi\left(\sum_{1}^{n} v_i \cdot \varphi\left(\sum_{1}^{m} w_{li} x_l + b_{10}\right) + b_0\right) \quad (12)$$

Na qual, vl e wlj são pesos sinápticos; b10 e b0 são os biases; e φ a função de ativação, comumente especificada como sendo a função sigmóide.

Perceptrons multicamadas têm a capacidade de aprender por meio do treinamento. O treinamento requer um conjunto de dados de treinamento, que consiste em uma série de vetores de entrada e saída associados. Durante o treinamento, o perceptron multicamadas é repetidamente apresentado com os dados de treinamento e os pesos na rede são ajustados até que ocorra o mapeamento de entrada / saída desejado.

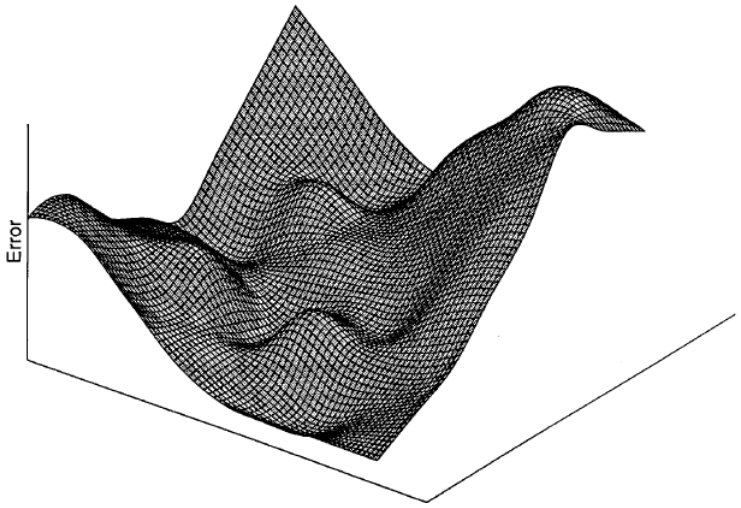
Existem muitos algoritmos que podem ser usados para treinar um *perceptron* multicamadas. Uma vez treinado adequadamente, o *perceptron* multicamadas pode generalizar para novos dados de entrada não vistos. O objetivo é encontrar a combinação de pesos que resulta no menor erro. Na prática, não é possível plotar tal superfície devido à grande quantidade de pesos. O que é necessário é um método para encontrar o ponto mínimo da superfície de erro.

Uma técnica possível é usar um procedimento conhecido como gradiente descendente. Os benefícios de cada abordagem são discutidos em Battiti (1992). Na prática, milhares de iterações de treinamento serão necessários antes que o erro de rede atinja um nível satisfatório.

A taxa de aprendizado determina o tamanho do passo dado durante o processo de aprendizado de descida gradiente descendente. Se for muito grande, o erro de rede mudará consideravelmente devido a grandes mudanças de peso, com a possibilidade de pular os mínimos globais. Por outro lado, se a taxa de aprendizagem for muito baixa, o treinamento levará muito tempo. O termo momentum é usado para auxiliar o processo de descida do gradiente, case fique preso em um mínimo local (GARTNER; DORLING, 1998).

Ao adicionar uma proporção da mudança de peso anterior à mudança de peso atual (que será muito pequena em um mínimo local), é possível que os pesos possam escapar do mínimo local. A Figura 16 abaixo, ilustra a superfície de erro para o MLP.

Figura 16 - Superfície de erro para um MLP com 2 pesos



Fonte: GARTNER e DORLING (1998).

3.2.4.5 Arquitetura Híbrida Inteligente (AHI)

De acordo com a pesquisa de Goldschmidt e Passos (2005), as técnicas inteligentes podem ser combinadas para produzir as chamadas arquiteturas híbridas. A grande vantagem desse sistema se deve à sinergia obtida pela combinação de duas ou mais técnicas. Esta sinergia se reflete na obtenção de um sistema mais poderoso (em termos de interpretação, aprendizado, estimativa de parâmetros, generalização, etc.) e menos obstáculos.

Existem três formas básicas de se associarem duas técnicas para a construção de uma arquitetura híbrida (SOUZA 1999).

Híbrida Sequencial: nesta forma, uma técnica atua como entrada de outra técnica. A arquitetura híbrida proposta no trabalho do Sassi (2006) é sequencial, porque usa o *Rough Sets* como pré-processador da rede SOM.

Híbrida auxiliar: esta forma poderia ser exemplificada do seguinte modo: uma RNA invoca um algoritmo Genético para a otimização de seus pesos ou de sua estrutura. Neste caso, tem-se um maior grau de hibridização em comparação com o híbrido sequencial.

Híbrida incorporada: nesta forma praticamente não há separação entre as duas técnicas. Pode-se dizer que a primeira técnica possui a segunda técnica e vice-versa. Poderia ser exemplificado por um sistema neuro-*fuzzy* híbrido em que um sistema de inferência *fuzzy* é implementado segundo a estrutura de uma rede neural artificial.

A arquitetura híbrida proposta nesse trabalho acadêmico visa estabelecer uma conexão sinérgica de técnicas inteligentes de regressão e classificação para predizer o arremate imobiliário em leilões com objetivo de apoiar as decisões nesse setor.

3.2.5 Interpretação e avaliação do conhecimento

Após a etapa de mineração de dados, é necessário interpretar o conhecimento descoberto ou processá-lo. Sendo assim, o objetivo principal desta etapa é melhorar a compreensão do conhecimento descoberto pelo algoritmo de mineração e verificá-lo medindo a qualidade da solução e a percepção do analista de dados.

Esses conhecimentos serão consolidados em forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo de KDD. Uma maneira genérica de obter a compreensão e interpretação dos resultados (BIGUS, 1996).

De acordo com Goldschmidt e Passos (2005), a tecnologia de visualização estimula a percepção e inteligência humana e aumenta a capacidade de compreender e associar novos padrões.

3.3 CONCEITO DE DECISÃO, RISCO E INCERTEZA

O mecanismo do Leilão envolve uma série de riscos às decisões, entre elas, qual o valor de referência a ser oferecido, por exemplo, um valor de desconto para facilitar liquidação do imóvel. Por outro lado, o comprador também está sujeito a decisões em condições de risco, uma vez que é incerto estabelecer um limite para essa negociação, tendo em vista, o interesse de outros compradores como uma espécie de concorrência.

Para melhor entender os conceitos acima abordados, uma breve descrição dos conceitos se faz importante nesse momento.

3.3.1 Decisão

A escolha é o resultado do processo de tomada de decisão, e o comportamento de escolha ocorre de acordo com as preferências por parte do tomador de decisão. Sendo assim, o tomador de decisão fará uma escolha com base nos resultados percebidos e geralmente escolherá o melhor benefício. Esse fenômeno, foi estudado por Bernoulli, que esboçou essa visão subjetiva pela primeira vez em um modelo chamado Utilidade (NAPOLITANO, 2014).

Por outro lado, o estabelecimento de um pensamento comum consiste em considerar o ponto de vista de cada um, para que as decisões tomadas nas organizações tenham um nível de qualidade superior. Logo a tomada de decisão nas organizações vai exigir cada vez mais trabalhos em equipe e maior participação das pessoas (ANGELONI, 1992).

O processo decisório passa, portanto, do nível individual para o nível de equipe. Por outro lado, na tomada de decisões, especialmente na tomada de decisões em equipe, não se pode ignorar o papel da tecnologia (ANGELONI, 2003).

A Utilidade como método de mensuração de risco é um tema clássico e amplamente utilizado na pesquisa em ciências sociais aplicadas, como economia e administração, principalmente porque se baseia no argumento de dois elementos básicos: fatos objetivos e visões subjetivas dos benefícios (BERSTEIN, 1997).

A liquidação de um imóvel envolve o processo de tomada de decisão. Nesse processo, o tomador de decisão deve considerar uma série de fatores, o que significa obter informações sobre o imóvel que lhe interessa, em suma, avaliar os benefícios por meio das informações objetivas e as visões subjetivas, buscando ao máximo, reduzir a incerteza na tomada de decisão (MCGREAL *et al.,* 2016).

Portanto, sistemas de apoio à decisão podem considerar os principais fatores que determinam a “atratividade” do investimento imobiliário em um ambiente urbano altamente competitivo (DEL GIUDICE *et al.,* 2019).

3.3.2 Risco e Incerteza

Através do contexto de decisões, o conceito do risco foi uma das grandes preocupações de estudiosos seminais como Knigth (1921) e Keynes (1921), ambos ligados à economia. Segundo eles, risco é uma incerteza mensurável. Assim, o risco de que um evento ocorra é dado por uma distribuição de suas probabilidades (JERÓNIMO, 2006).

Ainda segundo Jerónimo (2006), ao procurar estimar a probabilidade dos fenômenos, estas metodologias quantitativas visam fornecer aos decisores subsídios promovendo, portanto, uma decisão apoiada na ciência.

O risco está associado à prevenção, ao passo que a incerteza está articulada à precaução (Godard *et al.,* 2002).

Os conceitos baseados em risco podem levar a processos de mitigação, negociação e aceitação de riscos, enquanto métodos que enfatizam a incerteza no sentido de ignorância e incerteza podem promover direções prudentes e rejeitar certas decisões e escolhas (JERÓNIMO, 2006).

Para Knigth (1921) a mensuração do risco pode ser realizada através de ferramentas de probabilidade, o que nem sempre é possível porque há informações sobre a decisão a ser tomada ou como calcular a probabilidade de consequência.

Nesse sentido, medir o risco significa ter um certo grau de conhecimento e informação sobre eventos incertos, o que ajudará a prevê-lo. As escolhas observacionais possibilitam o comportamento dos tomadores de decisão diante de eventos incertos, o que reflete sua prontidão para agir (Ramsey, 1931).

Tal observação desencadeia não apenas a discussão do risco, mas também considera a conexão do risco como uma característica inerente a todos, pois, os tomadores de decisão irão considerar as escolhas, a partir de pensamentos ou inferências sobre ideias, cuja noção geral foi definida por Keynes (1921) como uma relação $\alpha=a/h$, onde *a* são conclusões e *h* são premissas, desde que exista um conhecimento de *h* que justifique uma crença racional em *a*, este pode ser medido por um grau de crença α.

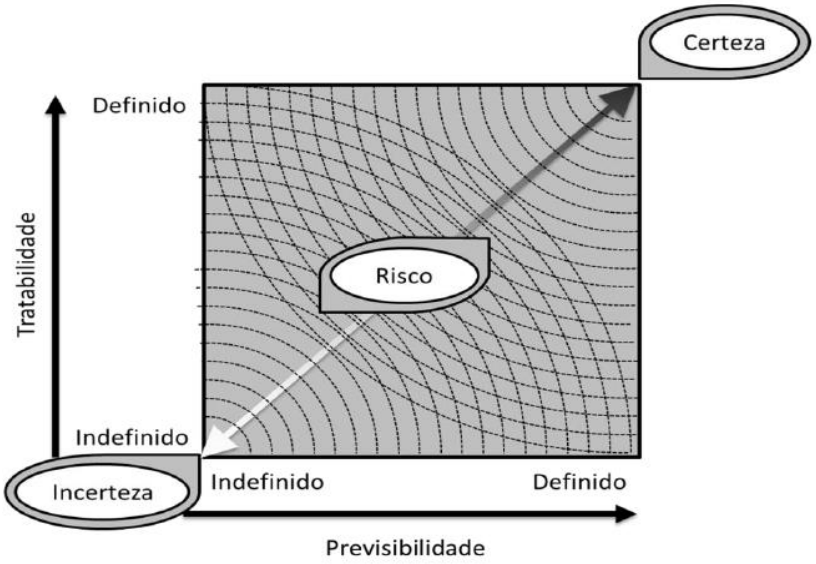
Assim quando se diz que um evento é provável, ou certo é estabelecido um grau de crença para a ocorrência deste evento. Portanto, ao lidar com riscos, eventos que podem afetar seu sucesso serão determinados (Kerzner, 2011). De acordo com as proposições de Ramsey e Keynes, as probabilidades podem ser atribuídas ou não com base no grau de crença.

Para Wideman (1992) a incerteza é constituída de uma falta de conhecimento sobre eventos futuros, sendo que estes podem ser favoráveis ou desfavoráveis. Deste modo, ao liquidar um imóvel no leilão, o leiloeiro pode dispor ou não de informações e conhecimento que diminuam as incertezas com esse setor.

Por outro lado, o risco envolve a compreensão das variáveis que afetam o resultado das decisões, informações sobre seu possível estado e sua relação causal, possibilitando assim, definir formas de lidar com os riscos e como lidar com as diferentes incertezas (PICH *et al.,* 2002).

A Figura 17 mostra como a classificação proposta por Wideman (1992) que se baseia na previsibilidade quando combinada com a noção de tratabilidade definida por Pich *et al.* (2002), onde é possível combinar ambos os conceitos representados em um plano definido pela previsibilidade de ocorrência de um risco e pela possibilidade de seu tratamento.

Figura 17 - Grau de risco



Fonte: Napolitano (2014).

Deste ponto de vista, risco pode ser tanto prejudicial (ameaça) quanto benéfico (oportunidade) em um projeto. Por outro lado, incerteza é a falta de informação ou de conhecimento sobre o resultado de uma ação, decisão ou evento.

Uma vez compreendida a conceituação da incerteza, do risco e da certeza é necessário entender sua incidência ou aplicação prática na classificação de ativos imobilizados urbanos em leilão.

3.4 MATRIZES DE PROBABILIDADE E IMPACTO

Matrizes de Probabilidade e Impacto (MPI), é uma ferramenta, cujo mecanismo é empregado para análise de riscos, no entanto, seu uso tem foco apenas das ameaças (COX, 2008 e PMI, 2017).

De modo geral as MPIs refletem uma expectativa, sobre um ou mais eventos incertos, sobre os quais pode-se estimar uma probabilidade e um impacto, o que se configura num risco, de acordo com a definição de Knigth (1921).

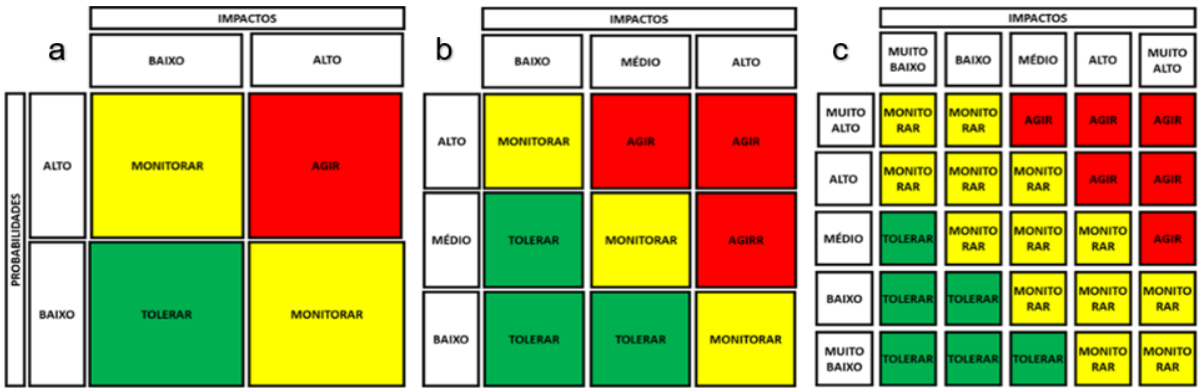
A matriz de risco combina a gravidade e a frequência das consequências que ocorrem em um determinado cenário de risco. Portanto, para estabelecer uma matriz de risco, é necessário classificar e dimensionar a gravidade das consequências, a frequência das consequências, o índice de risco, acumular conhecimento das regras baseadas no risco e editar a matriz de risco graficamente (MARKOWSKI; MANNAN, 2008).

Portanto, a classificação de gravidade e frequência depende do tipo de atividade ou da particularidade do processo envolvido.

As MPIs possibilitam a visualização do risco na forma de quadrantes de um plano que combinam as estimativas de probabilidade e impacto, com uma decisão a tomar. Com base em uma certa combinação de probabilidades ou influências, as operações podem ser predefinidas e tratadas como a terceira dimensão, na maioria das vezes em um esquema de cores de vermelho, amarelo e verde (COX, 2008).

Assim uma MPI é uma ferramenta que além de padronizar a decisão sobre os riscos, também permite a visualização e analisar conjuntos de riscos para auxiliar a tomada de decisões. Na Figura 18 são apresentados três exemplos de MPI citadas por Cox (2008), variando apenas, o número de linhas e colunas.

Figura 18 - Exemplos de MPI



Fonte: Adaptado de Cox (2008).

Nos exemplos contidos na Figura 18 acima, com o eixo das ordenadas para probabilidade e das abscissas para impacto, o risco é classificado e mapeado como sendo verde para risco baixo tolerável, amarelo para risco médio monitorável e vermelho para risco alto que requer atenção ou mesmo, ação por parte do decisor.

Portanto, o plano de impacto probabilístico que a matriz de risco oferece, colore as áreas de acordo com a prioridade exigida por cada combinação de dimensões de risco. A cor vermelha é empregada para os riscos de prioridade mais alta e processada com sendo uma urgência mais alta, o amarelo indica riscos de prioridade média a serem monitorados e a área verde contém riscos de baixa prioridade (HILLSON, 2009).

Cox (2008) sugere que para fins de avaliação de uma MPI, que se considere o eixo de impactos com uma escala normalizada, ou seja, o valor máximo de um impacto equivale a um e o mínimo a zero permitindo que diferentes formatos de matrizes sejam visualizados de uma forma mais clara.

3.5 MATRIZES DE PROBABILIDADE E IMPACTO DUPLAS

Um tema pouco explorado na literatura revisada, é o das Matrizes de Probabilidade e Impacto Dupla (MPDI), onde o risco é avaliado sob duas perspectivas distintas e descriminadas como oportunidades e ameaças, através do mesmo instrumento de visualização.

De modo geral, de acordo com a definição de Knigth (1921), a MPI reflete as expectativas de um ou mais eventos incertos, podendo estimar a probabilidade e o impacto do evento e configurá-lo como um risco.

As MPIs permitem visualizar os riscos na forma de quadrantes de planejamento, que combinam estimativas de probabilidade e impacto com as decisões. De acordo com a probabilidade dada ou combinação de influência, a ação pode ser pré-definida e considerada em quadrantes escalares em verde, amarelo e vermelho (COX, 2008).

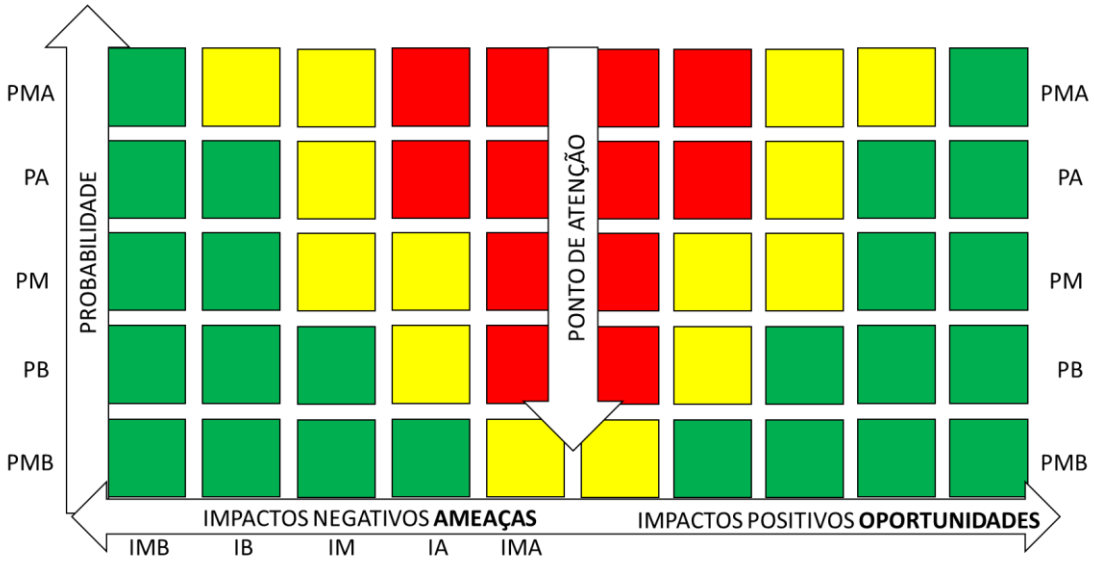
Posteriormente, surge um conceito similar às escalas de probabilidades, bem como a aplicação de estimativas subjetivas em impactos dos projetos, proposto por Hillson (2002), para o uso da ferramenta denominada Matriz de Probabilidades e Impactos Dupla (MPID) para a estimativa de riscos, que sejam estes ameaças ou oportunidades.

Hillson (2002) propõe a estruturação de uma matriz de probabilidade e impacto, similar as escalas empregadas nas MPIs, variando de muito baixo a muito alta aplicada à probabilidade e impactos dos eventos de risco. Porém, a escala de impactos comportaria valores positivos e negativos, representando oportunidades e ameaças, respectivamente.

Portanto, Hillson (2002) recomenda o uso do MPID como ferramenta de análise qualitativa de risco e, após a determinação de cada variável, o nível de atenção e tratamento será fornecido com base nos resultados de influência coordenada e probabilidade de abcissa.

A MPID proposta por Hillson (2002) é apresentada na Figura 19 a seguir.

Figura 19 - MPID proposta por Hillson (2002)



Fonte: Napolitano (2018).

A MPID é uma ferramenta que permite que ameaças e oportunidades sejam avaliadas e priorizadas usando um processo similar ao que é empregado nas MPIs, porém, com a diferença de abranger e mapear as ameaças e oportunidades, no apoio a tomada de decisão, de acordo com PMI (2017).

No PMBoK (2017) são feitas recomendações para o emprego de MPIDs apresentando riscos no interior dos quadrantes que é apresentado na Figura 20.

Figura 20 - MPID do PMBoK

Fonte: Napolitano (2018).

3.6 QUADRO COM CONCEITOS DO REFERENCIAL TEÓRICO

A seguir é apresentado o Quadro 4 que sumariza os principais conceitos e definições apresentados no referencial teórico, considerados como pilares teóricos para aplicação na presente pesquisa de dissertação.

Quadro 4 - Conceitos e Definições abordados no Referencial Teórico

Categorias	Conceito/Definição	Autor(es)
Leilões	Teoria de Leilões onde um enorme volume de recursos é transacionado por meio de leilões. Entre eles, destacando-se os imóveis.	Klemperer (1999)
	O mecanismo do setor de leilões envolve uma série de riscos às decisões, entre elas qual o valor de referência a ser ofertado nesse mercado que facilite a liquidação ou arremate, em um tempo inferior praticado pelo mercado de imóveis convencional.	Gan, 2013 e MCGREAL *et al.* (2016)
	O setor de leilões possui diversos sistemas comerciais de liquidação ou arremate imobiliário dentre os diferentes tipos de leilão.	Sandholm (2000)
	Define que leiloeiro é a parte responsável pela condução de todos os procedimentos legais do leilão. Geralmente um terceiro, totalmente isento de interesses financeiros nesse tipo de transação.	Castro (2003)
	Automação de processos no setor de leilões.	Lopes (2016)

Itens	Conceito/Definição	Autor(es)
Mercado imobiliário	Descrevem a fórmula para descobrir a relação teórica entre preços de imóveis e seu tempo exposto no mercado, dentro do mesmo conceito DOM, no entanto com nomenclatura diferente, *Time On Market* (TOM), utilizando métricas estatísticas tradicionais de risco e retorno imobiliário.	Cheng *et al.* (2008); Xin He *et al.* (2017) e Hengshu *et al*., (2016)
KDD	Os sistemas de descoberta de conhecimento *Knowledge Discovery in Data Bases* (KDD) surgem como uma solução para a extração do conhecimento.	Sassi (2006)
	Estabelece que é necessário realizar um trabalho de busca detalhado, denominado trabalho de "mineração", associado a um processo analítico, sistemático e, até onde possível, automatizado.	Silva (2016)
	O pré-processamento e transformação de dados garantem a qualidade e desempenho na extração do conhecimento.	Sassi (2006)
	Data Mining é a exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados para descobrir modelos e regras significativas.	Berry (2004)
	O KDD possui vários métodos de interpretação de dados, chamados tarefas. As tarefas mais comuns são associação de dados, classificação de dados, agrupamento de dados e visualização de dados.	Fayyad *et al.* (1996)
Aprendizado de máquinas	Definem que Inteligência Artificial é o estudo de como fazer os computadores realizarem tarefas as quais, até o momento, o ser humano faz melhor.	RICH e KNIGHT (1993)
	O AM também é adepto da previsão, como calcular a probabilidade de um evento ou prever o resultado.	Da Silva *et al. (*2017)
	As técnicas de aprendizado de máquina empregam um princípio de inferência denominado indução, no qual é possível obter conclusões genéricas a partir de um conjunto particular de exemplos.	Lorena e Carvalho, (2007)
	Aprendizado supervisionado, técnica inteligente que emprega um conjunto de dados, divididos em treino e teste, com instâncias e atributos variáveis, afim de predizer o valor de saída para novos exemplos com rótulos.	Da Silva *et al. (*2017) e HAN *et al.* (2011)
	Árvore de decisão é considerado um modelo estático representado através de uma função-alvo com valores discretos, que se utiliza de treinamento supervisionado para verificar a predizer dados futuros.	Quinlan e Cameron-Jones (1993) e Gama (2004)
	O classificador floresta aleatória ou floresta de decisão aleatória é um método de aprendizagem definido para classificação, regressão e outras tarefas, realizado através da construção de um número ilimitado de árvores de decisão durante o treinamento e geração da classe.	Breiman (1999)
	Rede neural artificial (RNA) é uma abstração de redes neurais biológicas. Adicionado camadas de neurônio ocultas no modelo de Rosenblatt, formado então a RNA Multilayer Perceptron (MLP).	Hebb (1949) e Rosenblatt (1958)
	As tecnologias podem ser combinadas para produzir as chamadas arquiteturas híbridas.	Goldschmidt e Passos (2005)

Itens	Conceito/Definição	Autor(es)
Decisões	A escolha é o resultado do processo de tomada de decisão, e o comportamento de escolha ocorre de acordo com as preferências por parte do tomador de decisão.	Knigth (1921) e Keynes (1921)
	A Utilidade como método de mensuração de risco é um tema clássico e amplamente utilizado na pesquisa em ciências sociais aplicadas.	Berstein (1997)
	A liquidação de um imóvel envolve o processo de tomada de decisão.	MCGREAL *et al.* (2016)
	Sistemas de apoio à decisão podem considerar os principais fatores que determinam a “atratividade” do investimento imobiliário em um ambiente urbano altamente competitivo.	DEL GIUDICE *et al.* (2019)
Risco e incerteza	Risco e Incerteza. Risco é risco é uma incerteza mensurável. A a incerteza é constituída de uma falta de conhecimento sobre eventos futuros.	Knigth (1921); Keynes (1921) e Wideman (1992)
	Matrizes de Probabilidade e Impacto (MPIs), é uma ferramenta prática, comumente empregada para análise de riscos em projetos, no entanto, seu uso tem foco na gestão apenas das ameaças.	COX (2008) e PMI (2017)
	Matrizes de Probabilidade e Impacto (MPI), é uma ferramenta prática, comumente empregada para análise de riscos em projetos, no entanto, seu uso tem foco na gestão apenas das ameaças.	COX (2008) e PMI (2017)
	Matriz de Probabilidade e Impacto Dupla (MPID), é uma ferramenta para a estimativa de riscos, que sejam estes ameaças ou oportunidades.	Hillson (2002)

Fonte: Elaborado pelo autor com base nos autores citados no referencial teórico.

4 METODOLOGIA DE PESQUISA

Uma vez determinados o problema de pesquisa, objetivos gerais e específicos, a revisão sistemática da literatura e o referencial teórico que apoiam a solução do problema de pesquisa, é definida a metodologia da pesquisa, bem como os procedimentos que serão empregados.

Inicialmente tem-se no subcapítulo 4.1 caracterização da pesquisa, 4.2 coleta dos dados de entrada, 4.3 apresentação do hardware e software empregados nos experimentos e 4.4 condução dos experimentos computacionais.

4.1 CARACTERIZAÇÃO DA PESQUISA

A essência ou a natureza dessa pesquisa científica é a aplicação, pois objetiva gerar conhecimentos para aplicação prática dirigida à solução de problemas específicos (PRODANOV, DE FREITAS, 2013).

Já o método de pesquisa científica empregado é considerado quantitativo, pois utiliza recursos e técnicas estatísticas para tentar converter em números o conhecimento gerado pelos pesquisadores (PRODANOV; DE FREITAS, 2013).

O propósito dessa pesquisa científica é explicativo, pois, o pesquisador tenta explicar as causas das coisas e suas causas registrando, analisando, classificando e explicando os fenômenos observados. Visa determinar os fatores que determinam o fenômeno ou contribuem para a ocorrência do fenômeno; “aprofundar a compreensão da realidade, pois explica as causas e as causas das coisas.” (GIL, 2008, p. 28). A maioria das pesquisas explicativas usa métodos experimentais que permitem a manipulação e o controle de variáveis para determinar qual variável independente determina a causa da variável dependente ou do fenômeno em estudo.

Quanto ao método da pesquisa científica, caracteriza-se sendo indutivo, porque generaliza, isto é, parte de algo particular para uma questão mais ampla, mais geral (LAKATOS; MARCONI, 2007).

Portanto, as técnicas e procedimentos utilizados na pesquisa científica são de natureza experimental, porque incluem, sujeitar o objeto de pesquisa a certas variáveis sob condições controladas e conhecidas pelo pesquisador para observar os efeitos das variáveis sobre o objeto (GIL, 2008).

Esse experimento se divide em dois momentos distintos. O primeiro momento, faz-se a avaliação e seleção das técnicas inteligentes que melhor se ajustem aos dados de imóveis em leilão coletados. Em outras palavras, foram selecionadas as técnicas com melhor acuracidade nas classificações do arremate desses imóveis. No segundo momento, foram selecionadas e empregadas duas técnicas inteligentes no desenvolvimento de uma AHI que seja capaz de classificar a liquidez de imóveis em leilão, com suporte de MPID no apoio as decisões.

4.2 COLETA DOS DADOS DE ENTRADA

Um aspecto importante a determinar na pesquisa é estabelecer um instrumento de mensuração baseado em construtos que possibilitem confirmar as propostas formuladas nesse trabalho acadêmico.

A partir do exposto e dos eixos teóricos apresentados nesse trabalho acadêmico, a Quadro 5 a seguir, usada para a coleta dos dados de entrada, apresenta as fontes utilizadas para elaboração e estruturação de um inventário, apresentando objetivos, e os tipos de evidências coletadas, conforme estipulado previamente nos objetivos específicos.

Quadro 5 - Inventário para os dados de entrada

Item	Evidência	Tipo	Objetivo
E01	Entrevista com especialista em imóveis	Entrevista	Entender os tipos de atributos importantes para análise de mercado.
E02	Entrevista com especialista em leilões	Entrevista	Entender o processo de transação de imóveis nesse setor.
D01	Levantamento de websites que possuem informações consistentes em grande quantidade para a análise	Documento	Mapear ou selecionar ambientes adequados para a coleta de dados.
E03	Entrevista com o responsável do departamento jurídico	Entrevista	Levantar as custas que envolve o processo de arrematação.
D02	Coleta dos dados de leilões encerrados	Documento	Verificar ativos patrimoniais imobiliários que foram e que não foram liquidados.

Fonte: Autor.

Nesse inventário, os cinco objetivos realizados possuem uma sequência cronológica em relação a ordem de coleta das informações bem como dos dados que delinearam esse experimento. Por outro lado, as partes interessadas possuem vasta experiência de mercado, o que pode ser considerado com conhecimento prévio especialista. Esse aspecto foi fundamental nessa etapa do processo.

A seguir será apresentada de modo simplificado, as questões feitas de modo não estruturado com os especialistas, para construção do inventário.

I. ENTREVISTA 1 - ESPECIALISTA EM IMÓVEIS

- O que é valor de mercado?
- O que é valor de liquidação?
- O que é liquidez de imóveis?
- O quê ou quais as variáveis afetam a liquidez de um imóvel no mercado tradicional? Exemplo: tempo de exposição, preço, localização, acesso, idade, vagas, desconto à vista, padrão de acabamento, tipo de imóvel, etc.

II. ENTREVISTA 2 - ESPECIALISTA EM LEILÕES DE IMÓVEIS

- O quê ou quais as variáveis afetam a liquidez de um imóvel no mercado de leilões?
- Qual o meio para coletar dados de imóveis urbanos em leilão para construção da base de dados nesse setor? E-mail, site de tribunal de justiça, site do próprio leiloeiro, etc.

III. ENTREVISTA 3 - ESPECIALISTA JURÍDICO EM LEILÕES

- Quais são as custas que normalmente, são submetidas a um imóvel em leilão?

Na etapa subsequente, coletou-se 1.620 imóveis no intervalo de período de 2017 até 2020, sendo 522 imóveis arrematados e 1.098 imóveis não arrematados.

A base coletada possui 3 classes de imóveis urbanos: residenciais, comerciais e industriais em 8 tipos de imóveis: apartamentos, casas, galpões, terrenos, glebas, lojas, salas comerciais e prédios comerciais.

Apesar do Brasil possuir 5.570 municípios, a amostra de 1.620 imóveis está contida em uma amplitude de 188 municípios em 21 Estados diferentes.

Na sequência, foi realizado um dicionário de atributos, conforme Quadro 6 a seguir, com as variáveis e suas informações, com finalidade de sumarizar os dados imobiliários coletados. Sem essa etapa, posteriormente dificulta o trabalho de transformar esses dados a serem minerados pelas técnicas inteligentes.

Quadro 6 - Dicionário dos atributos variáveis dos dados de entrada

N°	Variável	Classe	Tipo	Elementos	Dependência
1	classe	Nominal	Categórica	3	Independente
2	municipio	Nominal	Categórica	188	Independente
3	uf	Nominal	Categórica	21	Independente
4	tipo	Nominal	Categórica	8	Independente
5	situacao	Nominal	Categórica	2	Independente
6	area	Numérica	Contínua	-	Independente
7	vpp	Numérica	Contínua	-	Independente
8	vup	Numérica	Contínua	-	Independente
9	vm	Numérica	Contínua	-	Independente
10	lances	Numérica	Contínua	-	Dependente
11	exposicao	Numérica	Contínua	-	Dependente
12	selic	Numérica	Contínua	-	Independente
13	igpm	Numérica	Contínua	-	Independente
14	ipca	Numérica	Contínua	-	Independente
15	incc	Numérica	Contínua	-	Independente
16	pib	Numérica	Contínua	-	Independente
17	idade	Numérica	Contínua	-	Independente
18	condominio	Nominal	Categórica	2	Independente
19	esquina	Nominal	Categórica	2	Independente
20	vagas	Numérica	Contínua	-	Independente
21	abaixo	Numérica	Contínua	-	Dependente
22	arrematado	Nominal	Categórica	2	Atributo-alvo

Fonte: Autor.

O dicionário de atributos, divide-se pelo grau de dependência das variáveis estudadas.

Quanto aos 7 atributos variáveis independentes categóricos, o atributo “**classe**” divide-se em “1 - residencial”, “2 - comercial” e “3 - industrial”; o atributo “**municipio**” e “**UF**” possui uma variação de 188 municípios para 18 Estados conforme numeração específica do Instituto Brasileiro de Geografia e Estatística (IBGE); o atributo “**tipo**” é divido em “1 - apartamento”, “2 - casa”, “3 - galpao”, “4 - gleba”, “5 - loja”, “6 - lote”, “7 - sala comercial” e “8 - terrenos”; o atributo “**situacao**” indica se o imóvel encontra-se livre ou ocupado, dividindo-se em “1 - ocupado” ou “ 2 - livre”; os atributos “**condominio**” e “**esquina**” dividem-se em “1 - sim” e “2 - não”.

Quanto aos 11 atributos variáveis independentes contínuos, o atributo “**area**” relaciona-se com a área pertinente a cada imóvel, “**vpp**” refere-se ao valor de primeira praça, “**vup**” referente ao valor de última praça, “**vm**” referente ao valor de mercado, “**selic**” referente a taxa básica de juros, “**igpm**” referente ao Índice Geral de Preços do Mercado que trata da movimentação da economia, “**ipca**” referente ao Índice de preços no consumidor que é usado para observar tendências de inflação, “**incc**” referente ao Índice Nacional da Construção Civil, “**pib**” referente ao Produto Interno Bruto que representa a soma de todos os bens e serviços finais produzidos numa determinada região, “**idade**” referente a idade aparente ou real dos imóveis e “**vagas**” referente a quantidade de vagas que cada imóvel possui.

Quanto aos 3 atributos variáveis dependentes, porém contínuos, temos os atributos “**abaixo**” para o percentual de desconto aplicados nos imóveis, “**exposicao**” para o tempo de exposição desse imóvel no leilão e “**lances**” para a quantidade de lances que um imóvel possa ter em uma praça.

O atributo-alvo desse problema é prever as classes futuras do atributo variável categórico “**arrematado**”, que diz sobre a liquidação de imóveis em leilões indicando “1 - sim” e “2 - não”.

Nota-se que os atributos variáveis não possuem acentuação e são escritos todos em ordem minúscula para facilitar a leitura da linguagem de programação a ser empregada para extração do conhecimento.

Através da linguagem de programação Python foi gerada uma visão geral e simplificada da base de dados para finalidade de visualização da transformação feita. O Quadro 7 contém a base de dados já processada, transformada e pronta para ser minerada pelas técnicas inteligentes.

Quadro 7 - Sinopse da base de dados para visualização

	classe	municipio	uf	tipo	situacao	area	vpp	vup	vm	lances	...	igpm	ipca	incc	pib	idade	condominio
0	1	3507605	35	6	1	143.27	98637.24	49318.62	7.397793e+04	13	...	82786.0000	45754	41029	1.1	6	2
1	1	3550308	35	2	1	172.00	887633.33	624698.59	8.121082e+08	0	...	82786.0000	45754	41029	1.1	36	2
2	1	3548708	35	1	1	55686.00	223893.83	157572.00	2.048436e+05	0	...	82786.0000	45754	41029	1.1	21	1
3	1	3548708	35	1	1	583287.00	211017.54	106078.51	1.591178e+08	1	...	82786.0000	45754	41029	1.1	31	1
4	1	3534401	35	2	1	292.75	464380.11	233443.88	3.501658e+05	2	...	82786.0000	45754	41029	1.1	31	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1615	1	3530607	35	2	1	513.64	575949.97	411250.00	5.305125e+05	0	...	-0.8777	28039	42539	1.3	19	2
1616	2	3550308	35	7	1	344.40	828234.12	496940.47	7.006861e+09	0	...	-0.8777	28039	42539	1.3	41	2
1617	2	3550308	35	7	1	506.00	6730749.02	5048061.76	6.360558e+10	0	...	-0.8777	28039	42539	1.3	35	2
1618	1	3513009	35	2	2	87.89	244896.43	146937.86	2.057130e+08	11	...	-0.8777	28039	42539	1.3	17	1
1619	1	3550308	35	2	1	175.00	276908.13	166460.55	2.330448e+05	1	...	68389.0000	40049	41539	1.0	38	2

Fonte: Autor.

4.3 HARDWARE E SOFTWARE EMPREGADOS NOS EXPERIMENTOS

Por meio de um computador Lenovo, modelo ideapad s145, configurado com sistema operacional Windows 10®, processador INTEL® Core™ i7-8565U da 8th geração, 1 TB de HD, SSD NVMe M.2 500 GB, CPU @ 1.99 GHz, memória RAM OPTANE™ de 8 GB, placa de vídeo modelo NVIDIA® GEFORCE® MX110 com 2 GB, os experimentos computacionais foram realizados empregando as Linguagens de programação Python e R por meio do JUPYTER, Power BI, Excel, e do *software* WEKA.

Na linguagem de programação Python, foram empregadas, para visualização de dados, as 5 bibliotecas a seguir:

1. Pandas, para manipulação e análise de dados;
2. Numpy, para funções matemáticas;
3. Matplotlib, para visualizações de dados;
4. Seaborn, para estatística e visualizações;
5. Sklearn, para estatística e também visualizações.

Na linguagem de programação R, foram empregadas, para a extração do conhecimento da base de dados, as seguintes 20 bibliotecas a seguir:

1. “FactoMineR”, para análise exploratória multivariada de dados e Mineração de Dados;
2. “factoextra”, para extração e visualização dos resultados de análises de dados multivariadas;
3. “cluster”, para plotar, validar, prever novos dados e estimar o número ótimo de clusters. O pacote aproveita “RcppArmadillo” para acelerar as partes computacionalmente intensivas das funções;
4. “csv”, para ler, escrever, formatar e salvar arquivos do Excel;
5. “dplyr”, para trabalhar com objetos semelhantes a quadros de dados, tanto na memória quanto fora dela;
6. “Factoshiny”, para análise fatorial, desenho de gráficos de forma interativa graças ao “FactoMineR” e um aplicativo *Shiny*, ou seja, é um sistema para desenvolvimento de aplicações web;
7. “xlsx”, para ler, escrever, formatar e salvar arquivos do Excel;

8. “arules”, para representar, manipular e analisar dados e padrões de transações (conjuntos de itens frequentes e regras de associação);
9. “caTools” que contém várias funções utilitárias básicas, incluindo: funções estatísticas de janela de movimento (rolagem, execução), leitura e gravação para arquivos binários GIF e ENVI, cálculo de AUC, classificador *LogitBoost*, entre outras;
10. “rpart”, para particionamento recursivo de árvores de classificação e regressão.
11. “rpart.plot”, para modelos de plotagem de modelos da biblioteca “rpart”;
12. “caret”, para funções de classificação e regressão;
13. “e1071”, para análise de classe latente, agrupamento difuso, máquinas de vetores de suporte, cálculo de caminho mais curto, agrupamento em pacotes, classificador Bayes ingênuo, entre outros;
14. “ROCR”, para visualização do desempenho dos classificadores;
15. “pROC”, para exibir e analisar curvas ROC;
16. “PRROC”, para exibir e analisar curvas ROC para dados ponderados e não ponderados;
17. “corrplot”, para visualização de matrizes de correlação.
18. “Hmisc”, que contém muitas funções úteis para análise de dados, gráficos de alto nível, operações de utilitários, funções para calcular o tamanho e potência da amostra, importar e anotar conjuntos de dados, imputar valores ausentes, criação de tabela avançada, agrupamento de variáveis, manipulação de sequência de caracteres, conversão de objetos R para LaTeX e código html e variáveis de recodificação.
19. “PerformanceAnalytics”, tem como objetivo de auxiliar os profissionais e pesquisadores na utilização das pesquisas mais recentes na análise de fluxos de retorno não normais. Em geral, ele é mais testado em dados de retorno (ao invés de preço) em uma escala regular, mas a maioria das funções funcionará com dados de retorno irregulares também, e um número crescente de funções funcionará com dados de P&L ou preços onde possível.
20. “randomForest”, para classificação e regressão com base em uma floresta de árvores usando entradas aleatórias, com base em (BREIMAN, 1999).

O Microsoft Power BI e o Excel foram empregados apenas para visualização e exposição dos dados.

O software WEKA, foi empregado para avaliação das técnicas inteligentes.

4.4 CONDUÇÃO DOS EXPERIMENTOS COMPUTACIONAIS

A condução dos experimentos computacionais foi dividida em três etapas para classificação de liquidez imobiliária em leilão, com foco no posterior mapeamento dos riscos, e nos subtópicos a seguir:

a) Seleção e avaliação das técnicas inteligentes;
b) Protocolo de condução dos experimentos computacionais;
c) Métricas de avaliação das técnicas inteligentes.

4.4.1 Seleção e avalição das técnicas inteligentes

Nesse ponto, necessita-se selecionar técnicas supervisionadas para as tarefas de classificação e regressão dos dados de entrada. Portanto, foi empregado o *software* WEKA para avaliar tais técnicas.

O WEKA é uma coleção de algoritmos de AM e ferramentas de pré-processamento, projetado para experimentar métodos existentes em conjuntos de dados. Ele fornece amplo suporte para todo o processo de mineração experimental de dados, incluindo a preparação dos dados de entrada, avaliação de esquemas de aprendizagem, visualização dos dados de entrada e os resultados da aprendizagem (FRANK *et al.,* 2016).

Segundo a universidade que o desenvolveu, o *software* WEKA é escrito em Java e distribuído sob os termos da *General Public License* (GNU). O WEKA roda em quase todas as plataformas e foi testado nos sistemas operacionais Linux, Windows e Macintosh (FRANK *et al.,* 2016).

Ayu *et al.* (2012) empregaram o software WEKA em sua pesquisa para avaliarem e compararem o desempenho de sete categorias diferentes de algoritmos de classificação com finalidade de reconhecimento de atividades em tempo real do uso de celulares.

No trabalho de DASH (2013) foi analisado três conjuntos de dados para avaliar qual categoria de classificador deram melhores resultados. Sendo assim, o autor utilizou o *software* WEKA como ferramenta de mineração de dados para esse propósito.

Com base no exposto, são listadas as principais características das técnicas inteligentes a serem empregadas no *software* WEKA no Quadro 8 abaixo:

Quadro 8 - Categorias das Técnicas Inteligentes no WEKA

Técnica inteligente	**Categorias**	**Descritivo**
Decision Tree - DT	Árvores	Algoritmos baseados em árvores de decisão que permitem definir limite de confiança em uma folha, escolher a poda com erro reduzido, determinar o tamanho do conjunto de poda, suprimir o aumento da sub-árvore produzindo um algoritmo mais eficiente, e aplicar a suavização de Laplace para probabilidades previstas.
Random Forest - RF		
Multilayer Perceptron - MLP	Função	Algoritmos que se enquadram na categoria de funções incluem um grupo variado de classificadores que podem ser escritos como equações matemáticas de uma forma razoavelmente natural.

Fonte: Autor com base em (FRANK *et al.,* 2016).

Os experimentos foram organizados em uma estrutura que visa facilitar a organização da fase experimental, bem como da apresentação dos resultados na sequência, sob o ponto de vista operacional.

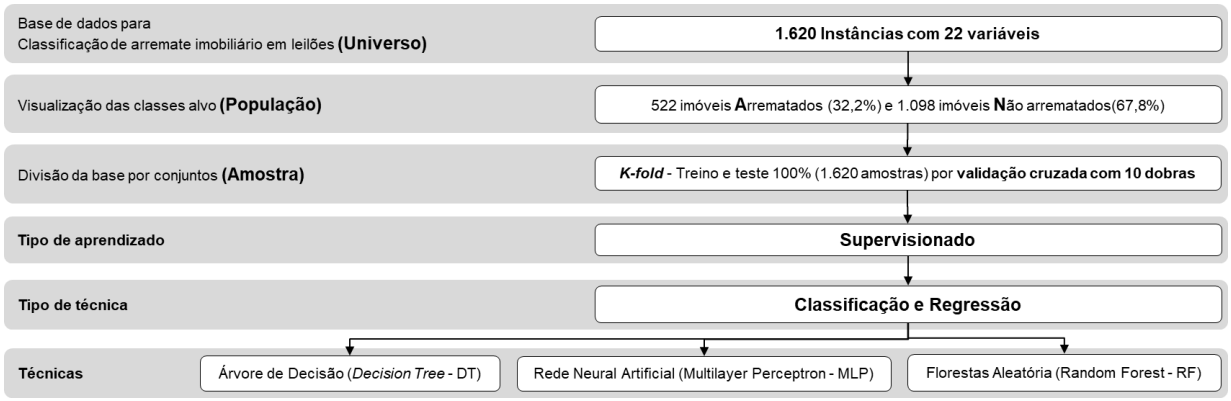
Para avaliar a capacidade de generalizar o modelo em relação ao conjunto de dados de entrada, são aplicadas técnicas de validação cruzada. Essa técnica é amplamente utilizada em problemas em que o objetivo da modelagem é a previsão. Em seguida, estima-se a acurácia do modelo na prática, ou seja, seu desempenho em um novo conjunto de dados.

Para verificação de ajuste do modelo aos dados, diversas formas de se realizar validação cruzada dos dados foram sugeridas na literatura, sendo que às duas mais utilizadas são *Holdout* e o *K-fold* (KOHAVI *et al.,* 1995).

O método denominado k-fold por validação cruzada consiste em dividir o conjunto de dados total em k subconjuntos mutuamente exclusivos do mesmo tamanho e, em seguida, usar um subconjunto para teste e os k-1 restantes são usados para estimar parâmetros e calcular a precisão do modelo. Este processo é executado de forma cíclica, testando alternadamente o subconjunto k vezes (DA SILVA *et al.*, 2017).

Para a avaliação das técnicas inteligentes de classificação, em uma primeira etapa, foram estruturadas validações *K-fold* em 10 partes conforme Figura 21 a seguir:

Figura 21 - Estrutura operacional dos experimentos para *K-fold* em 10 dobras



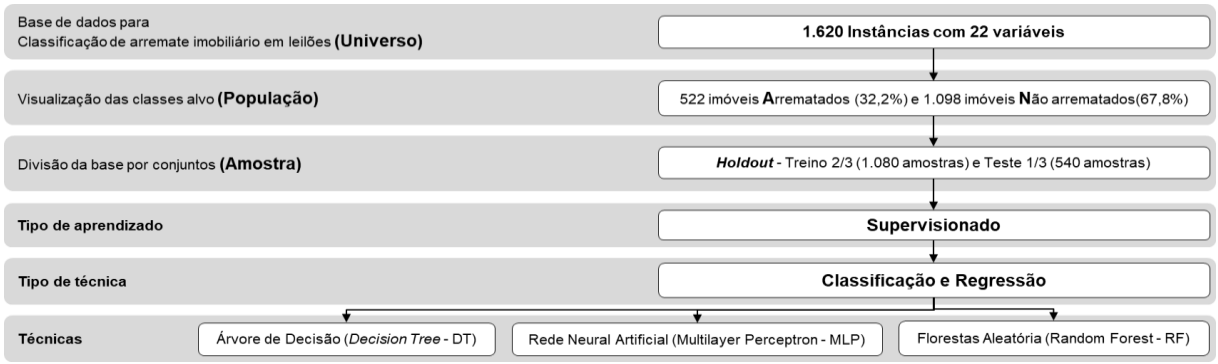
Fonte: Autor.

O método denominado *K-fold* por validação cruzada, aplicado nesse primeiro momento, dividiu em 10 subconjuntos mutuamente exclusivos do mesmo tamanho. Em seguida, usando um subconjunto para teste e os k-1 restantes foram usados para estimar parâmetros e calcular a precisão do modelo. Este processo foi executado de forma cíclica, testando o subconjunto 10 vezes com as técnicas DT, MLP e RF.

Outra forma de avaliar a capacidade de generalizar o modelo em relação ao conjunto de dados de entrada, o método *Holdout*, pressupõe a criação de dois subconjuntos de dados a partir do conjunto de dados disponível para uso na indução do modelo (AZEVÊDO, 2018).

Portanto, em uma segunda etapa, ainda com intuito de avaliar as técnicas inteligentes de classificação e regressão, foram estruturadas validações do tipo *Holdout*, sendo 2/3 para treino e 1/3 para teste do modelo conforme Figura 22 a seguir:

Figura 22 - Estrutura operacional dos experiementos para *Holdout*



Fonte: Autor.

Nos patamares subsequentes, os dados de entradas, as técnicas inteligentes bem como as validações cruzadas empregadas nesta pesquisa, necessitam de uma forma retilínea e cronológica de uma condução experimental através de um protocolo a ser elaborado com base na fase mineração, cujo objetivo é de extração de conhecimentos desta base de dados coletada.

4.4.2 Protocolo de condução dos experimentos computacionais

A Arquitetura Híbrida Inteligente (AHI) baseou-se em dois grupos de dados de entrada. Na fase de seleção dos dados, o primeiro conjunto de dados com 1.600 imóveis para treino e teste, e outro com 20 imóveis remanescentes com finalidade de teste final do modelo híbrido proposto com base no modelo de extração de conhecimento.

Para garantir a qualidade dos dados de entrada selecionados, foram realizadas as tarefas de processamento e transformação de dados. No pré-processamento dos dados de entrada, foram removidas do conjunto, dados com valores ausentes e duplicatas de instâncias. A transformação dos dados de entrada foi realizada a conversão das variáveis passando-as de características nominais para características numéricas a fim de que o algoritmo interprete cada instância, já na fase de mineração onde ocorre esse processo de transformação.

A avaliação do modelo final consistiu em empregar a técnica *Holdout* com 2/3 do conjunto de dados para treino e 1/3 para teste para avaliar o modelo e predizer dados novos do segundo conjunto de dados de entrada.

A fase de interpretação dos dados foi realizada via *desktop*. Portanto, mesmo havendo a possibilidade de implantação de *webservices*, a título experimental, foi realizado desta forma, pois, o objetivo em primeira instância é desenvolver o modelo.

4.4.3 Métricas de avaliação das técnicas inteligentes

4.4.3.1 Matriz de confusão

A matriz de confusão, ou do inglês (*confusion matrix*), é uma matriz utilizada para descrever o número de previsões corretas e incorretas feitas por um classificador treinado (AZEVEDO, 2018).

No Quadro 9 abaixo são relacionadas as previsões positivas e negativas relacionando valores oriundos de previsões comparados com valores reais.

Quadro 9 - Representação da matriz de confusão e suas medidas de avaliação

<table>
<tr><td colspan="2" rowspan="2">-</td><td colspan="2">Valor predito</td><td rowspan="2">-</td></tr>
<tr><td>Positivo</td><td>Negativo</td></tr>
<tr><td rowspan="4">Valor real</td><td rowspan="2">Positivo (P)</td><td rowspan="2">Verdadeiro Positivo (VP)</td><td rowspan="2">Falso Negativo (FN)</td><td>TVP = $\frac{VP}{P} = \frac{VP}{FN+VP}$</td></tr>
<tr><td>TFN = $\frac{FN}{P} = \frac{FN}{FN+VP}$</td></tr>
<tr><td rowspan="2">Negativo (N)</td><td rowspan="2">Falso Positivo (FP)</td><td rowspan="2">Verdadeiro Negativo (VN)</td><td>TFP = $\frac{FP}{N} = \frac{FP}{FP+VN}$</td></tr>
<tr><td>TVN = $\frac{VN}{N} = \frac{VN}{FN+VP}$</td></tr>
<tr><td colspan="2">-</td><td>VPP = $\frac{VP}{VP+FP}$</td><td>VPN = $\frac{VN}{FP+VN}$</td><td>ACC = $\frac{VP+VN}{FP+FN+VP+VN}$</td></tr>
</table>

Fonte: Adaptado pelo autor de Azevedo (2018).

A matriz de confusão, representada no Quadro acima, relaciona as contagens das previsões verdadeiras positivas, verdadeiras negativas, falsas positivas e falsas negativas de um classificador onde:

- VP - Verdadeiro Positivo, representa às instâncias positivas que foram corretamente rotuladas como positivas;
- FN - Falso Negativo, representa às instâncias negativas que foram incorretamente rotuladas como negativas;
- FP - Falso Positivo, representa às instâncias positivas que foram incorretamente rotuladas como positivas;
- VN - Verdadeiro Positivo, representa às instâncias negativas que foram corretamente rotuladas como negativas;
- TVP - Taxa Verdadeiro Positivo ou sensibilidade, que refere-se à razão de resultados corretamente classificados como positivo do modelo, quando comparado com a parcela dos valores definidos como positivos na amostra;
- TFN - Taxa Falso Negativo, que se refere à razão de resultados incorretamente classificados como falso no resultado do modelo, quando comparado com a parcela dos valores negativos na amostra;
- TFP - Taxa Falso Positivo ou especificidade, que se refere à razão de resultados incorretamente classificados como falso no resultado do modelo, quando comparado com a parcela dos valores definidos como positivos na amostra;

- TVN - Taxa Verdadeiro Negativo ou especificidade, que se refere à razão de resultados corretamente classificados como negativos no resultado do modelo, quando comparado com a parcela dos valores definidos como negativos na amostra;
- ACC - Acurácia, que se refere à quantidade classificada como Positivos e Negativos corretamente.

Logo observa-se que VP e VN indicam o quanto o classificador está acertando nas predições, enquanto FP e FN indicam o quanto o classificador está errando nas predições.

4.4.3.2 Curva *Receiver Operating Characteristic* (ROC) e *Area Under the Curve* (AUC)

O termo *Receiver Operating Characteristic* (ROC) derivou de testes da capacidade dos operadores de radar da Segunda Guerra Mundial de determinar se o aviso na tela do radar tratava-se de um objeto do tipo de sinal ou possível ruído (FAN *et al.,* 2006).

Por outro lado, a determinação de um valor de corte “ideal” é quase sempre uma troca entre sensibilidade (verdadeiros positivos) e especificidade (verdadeiros negativos). Como ambos mudam com cada valor de “corte”, torna-se difícil para o leitor imaginar qual corte é ideal. Logo a curva ROC oferece uma ilustração gráfica desses conflitos em cada corte para qualquer teste de diagnóstico que usa uma variável contínua (FAN *et al.,* 2006).

Portanto, a análise ROC, em ciência de dados, é um método gráfico para avaliação de modelos de classificação em ML. É especialmente útil em áreas onde há uma grande proporção de incompatibilidade entre categorias ou quando diferentes custos / benefícios devem ser considerados para diferentes erros / sucessos de classificação. Portanto, a fim de simplificar a análise ROC, AUC (“*area under the curve*”) entende-se que nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da ROC, calculando a “área sob a curva” (ALBANO; NAPOLITANO, 2020).

A métrica AUC é constante em escala porque funciona com precisão de classificação em vez de valor absoluto. Além disso, independentemente do limite de classificação, ele também pode medir a qualidade das previsões do modelo. O intervalo do valor de AUC é de 0,0 a 1,0 e o limite entre os níveis são de 0,5. Em outras palavras, se esse limite for ultrapassado, o algoritmo será classificado em uma categoria e depois em outra categoria. Segundo Prati *et al.* (2008), quanto maior o AUC, melhor.

4.4.3.3 Precisão

Essa métrica pode ser definida como a porcentagem de exemplos de classificações de classe Positivo que o modelo fez, quantas estão corretas. A precisão pode ser usada em uma situação em que os Falsos Positivos são considerados mais prejudiciais que os Falsos Negativos (POWERS, 2011).

A seguir é apresentada a fórm. (12) a seguir:

$$Precisão = \frac{VP}{VP + FP} \tag{3}$$

4.4.3.4 *Recall* | Revocação | Sensibilidade

Essa métrica representa a porcentagem de exemplos classificados como positivos por um modelo de classificação que são verdadeiros positivos. O *recall* pode ser usado em uma situação em que os Falsos Negativos são considerados mais prejudiciais que os Falsos Positivos (POWERS, 2011).

A seguir é apresentada a fórm. (13) a seguir:

$$Recall = \frac{VP}{VP + FN} \tag{4}$$

4.4.3.5 *F-Measure*

A *F-Measure* é uma métrica que representa a combinação entre Precisão e *Recall*. Sua representação é definida pela média harmônica das duas métricas, Precisão e *Recall*, e é usada como um score de desempenho agregado. Média que está muito mais próxima dos menores valores do que uma média aritmética simples. Ou seja, quando se tem um *F-Measure* baixo, é um indicativo de que ou a precisão ou o recall está baixo (POWERS, 2011).

A seguir é apresentada a fórm. (14) a seguir:

$$F - Measure = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad (5)$$

4.4.3.6 *Matthews Correlation Coefficient* (MCC)

O coeficiente de correlação Matthews ou *Matthews Correlation Coefficient* (MCC) é empregado na aprendizagem de máquina como uma medida da qualidade de duas classes binárias introduzido por Brian W. Matthews em 1975. A métrica leva em consideração verdadeiros e falsos positivos e é geralmente considerado uma medida equilibrada que pode ser usada mesmo se as classes forem de tamanhos muito diferentes (MATTHEWS, 1975).

O MCC é em essência um coeficiente de correlação entre as classificações binárias observadas e previstas; ele retorna um valor entre -1 e +1. Um coeficiente de +1 representa uma predição perfeita, 0 não melhor do que a predição aleatória e -1 indica discordância total entre predição e observação. MCC está intimamente relacionado à estatística para uma tabela de contingência 2 × 2 (MATTHEWS, 1975).

O MCC pode ser calculado diretamente a partir da matriz de confusão usando fórm. (15) a seguir:

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

Se qualquer uma das quatro somas no denominador for zero, o denominador pode ser arbitrariamente definido como um; isso resulta em um coeficiente de correlação de Matthews de zero, que pode ser mostrado como o valor limite correto (MATTHEWS, 1975).

4.4.3.7 Estatística Kappa

O teste de concordância Kappa (K), é conhecido por coeficiente de Kappa, foi proposto por Jacob Cohen em 1960, com a finalidade de medir o grau de concordância entre proporções derivadas de amostras dependentes (FLEISS *et al.,* 2013).

Para Fleiss *et al.* (2013), descrever a intensidade da concordância entre esses juízes, bem como entre os testes de diagnóstico utilizados, uma alternativa é recorrer ao coeficiente Kappa. Esse coeficiente se baseia no número de respostas concordantes, mais precisamente, no número de casos cujo resultado é o mesmo entre os juízes.

O coeficiente Kappa é calculado pela fórm. (16) a seguir:

$$Kappa = \frac{P(O) - P(E)}{1 - P(E)} \tag{7}$$

em que: P(O): proporção observada de concordâncias (soma das respostas concordantes dividida pelo total); P(E): proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes divididas pelo total).

Assim sendo, o Kappa é considerado como uma medida de concordância Inter observador que permite avaliar tanto se a concordância está além do esperado tão-somente pelo acaso, quanto o grau dessa concordância. Essa medida tem como valor máximo o valor unitário, que representa total concordância. Os valores próximos e até mesmo abaixo de zero indicam nenhuma concordância, ou a presença de uma eventual discordância entre os juízes (FLEISS *et al.,* 2013).

4.4.3.8 Erro Absoluto Médio (EAM)

O Erro Absoluto Médio é uma medida de precisão de um modelo de Aprendizado de Máquinas, comumente empregado como uma função de perda para problemas de regressão e na avaliação de modelos, devido à sua interpretação muito intuitiva em termos de erro relativo (ALENCAR *et al.,* 2011).

O EAM é apresentado na fórm. (17) a seguir:

$$EAM = N^{-1} \sum_{i=1}^{N} |P_i - O_i| \tag{8}$$

Em que N representa o número de observações; Pi são valores estimados e Oi são valores estimados pelo método PMFAO 56 (mm d^{-1}).

4.4.3.9 Erro Quadrático Médio (EQMR)

Em estatística, o Erro Quadrático Médio da Raiz é uma medida de precisão de um modelo de Aprendizado de Máquinas, apresentada na fórm. (19), que é empregado para avaliar problemas de regressão, no entanto, com uma penalidade maior em relação ao erro absoluto médio (ALENCAR *et al.,* 2011).

$$EQMR = \sqrt{N^{-1}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{9}$$

4.4.3.10 Erro Absoluto (EA)

Erro absoluto, mostrado na fórm. (19) a seguir, é dado pelo módulo da diferença do valor verdadeiro pelo valor encontrado (HALLAK, 2011).

$$EA = |VV - VE| \tag{10}$$

4.4.3.11 Erro Relativo (ER)

Erro relativo, mostrado na fórm. (20) a seguir, é dado pela razão do erro absoluto pelo valor verdadeiro

$$ER = \frac{EA}{VV} \tag{11}$$

4.4.3.12 Erro Quadrático Relativo da Raiz

O Erro Quadrático Relativo da Raiz (EQRR) raiz quadrada média é uma medida utilizada as diferenças entre os valores previstos por um estimador e os valores observados. O EQRR representa a média quadrática dessas diferenças. Estes desvios são chamados resíduos quando os cálculos são executados através da amostra de dados que foi utilizado para a estimativa e são chamados erros (ou erros de predição) quando calculado para fora da amostra (HALLAK, 2011).

O RMSD de um estimador em relação a um parâmetro estimado é definido como a raiz quadrada do erro quadrático médio na fórm. (21) a seguir:

$$EQRR(\hat{\theta}) = \sqrt{REQM)} \tag{12}$$

Para um estimador , o RMSD é a raiz quadrada da variância, conhecido como o desvio padrão .

4.4.3.13 Tempo de processamento

Na atual era do *Big Data*, a cada dia ocorre o aumento horizontal das bases de dados. Sinal esse observado por Clésio (2020), em termos computacionais que faz com que os algoritmos de mineração de dados tenham que processar um volume de dados muito maior, aumentando a complexidade do processamento que por consequência aumenta custos demandados por essa tarefa.

Dada essa pequena introdução, essa é a razão na qual a redução da dimensionalidade é muito importante para toda tarefa de mineração dos dados e pela economia de recursos financeiros computacionais.

Clésio (2020), expôs 7 técnicas para redução da dimensionalidade, dentre as quais apontou a utilidade do Random Forest como classificador eficaz por conta da redução da dimensionalidade gerada pelo conjunto de árvores em relação a um atributo de destino a fim de descobrir subconjunto com maior poder preditivo. Por outro lado, essa técnica demanda mais tempo de processamento.

Na literatura acadêmica, encontram-se outros autores que abordam essa temática como Machado (2015), que analisou a teoria e as avaliações empíricas do conjunto de técnicas propostas para a redução do custo computacional da classificação em tempo de teste de classificadores a partir dos dados de treinamento. O autor encontrou técnicas que permitem reduzir a quantidade de bits necessária para se realizar a classificação e trocar cada multiplicação em ponto flutuante por um simples deslocamento de bit em inteiro gerando mais economia em termos de custo computacional.

Deste modo, essas operações de baixo custo computacional são importantes por permitir aumentar a velocidade de classificação, ao mesmo tempo que diminui o consumo de energia. Desse modo, Machado (2015), mostrou em sua tese a redução em quase 50% a quantidade de bits necessária para a extração de características na maioria dos experimentos realizados.

Oliveira *et al.* (2007), avaliaram como o custo computacional de referência dos algoritmos conhecidos de estimadores se correlaciona com o tempo médio total de processamento no processo de identificação de etiquetas.

A conversão do aprendizado para uma hipótese depende da complexidade da amostra. Em outras palavras, depende de quantos exemplos de treinamento são necessários para a conversão do aprendizado, ou quanto esforço é despendido antes do aprendizado convergir para uma hipótese promissora (MITCHELL, 1997).

O objetivo, portanto, é caracterizar classes de conceitos-alvo que possam ser aprendidos com segurança em um número razoável de exemplos de treinamento sorteados aleatoriamente e quantidade de esforço computacional necessário para a aprendizagem indutiva (MITCHELL, 1997).

5 ANÁLISE DOS RESULTADOS

Neste capítulo foram expostos os dados coletados de imóveis urbanos em leilão para a realização dos experimentos, foram avaliadas e selecionadas técnicas inteligentes que melhor se ajustam aos dados coletados, foi realizado uma análise da aplicação da AHI para classificação de liquidez dos imóveis em leilões e por último, foi feito o mapeamento dessas classificações na MPID no apoio as decisões que são tomadas nesse setor, verificando por fim se essas análises suportam às proposições desse trabalho acadêmico.

5.1 EXPOSIÇÃO DOS DADOS DE ENTRADA

Sabe-se que a mineração de dados se faz muito útil quando a quantidade de dados disponível é grande e representativa, motivo pelo qual a fase de coleta de dados e a tarefa de amostragem são muito importantes no processo de descoberta de conhecimento.

A seguir, na Figura 23 ilustra a distribuição dos imóveis urbanos em leilões arrematados e não arrematados em todo território nacional.

Figura 23 - Dispersão dos dados imobiliários de leilão

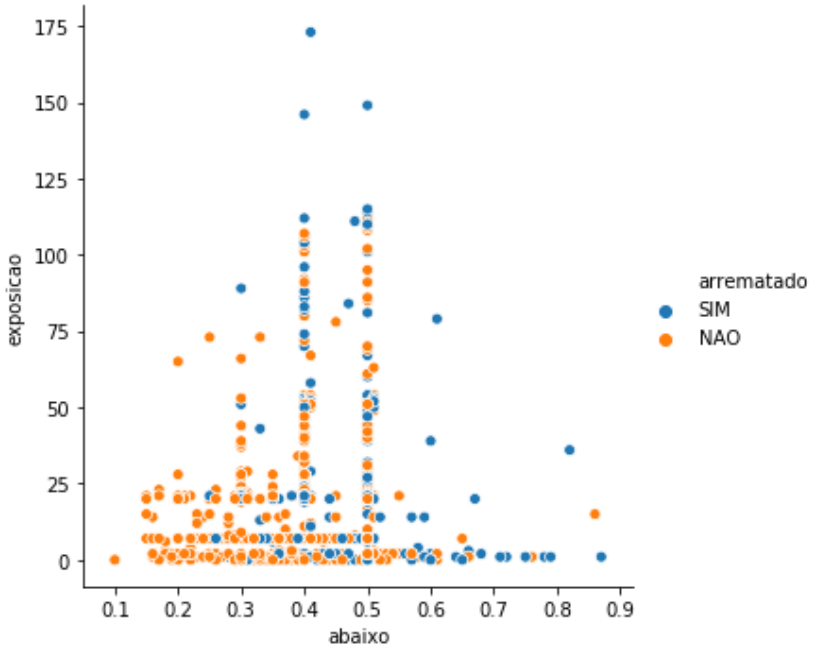
Fonte: Autor.

Apesar da aleatoriedade da coleta dos dados, a região Sudeste, principalmente no município de São Paulo, concentra maior número de imóveis transacionados nesse setor. As cores na legenda, indicam se o imóvel foi arrematado ou não em uma praça de leilões. Essa concentração de imóveis pertence às áreas urbanas de cada Estado.

A taxa de imóveis não liquidados alcança 68,78%, representando 1.098 imóveis não arrematados. Sendo assim, a base compondo 1.620 imóveis com processo de leilão encerrado, possui 522 imóveis arrematados compondo um percentual de 32,22%.

Outra forma de visualização dos dados coletados é através da biblioteca *Seaborn* da linguagem de programação Python. Com essa biblioteca foi gerado um gráfico de distribuição dos imóveis urbanos em leilão, subdivididos nas classes arrematado e não arrematado no eixo das abscissas, e no eixo das ordenadas pelo tempo que esses imóveis ficaram expostos no mercado de leilões de imóveis conforme ilustrado no Gráfico 1 abaixo:

Gráfico 1 - Distribuição dos imóveis em leilões com rótulos



Fonte: Autor.

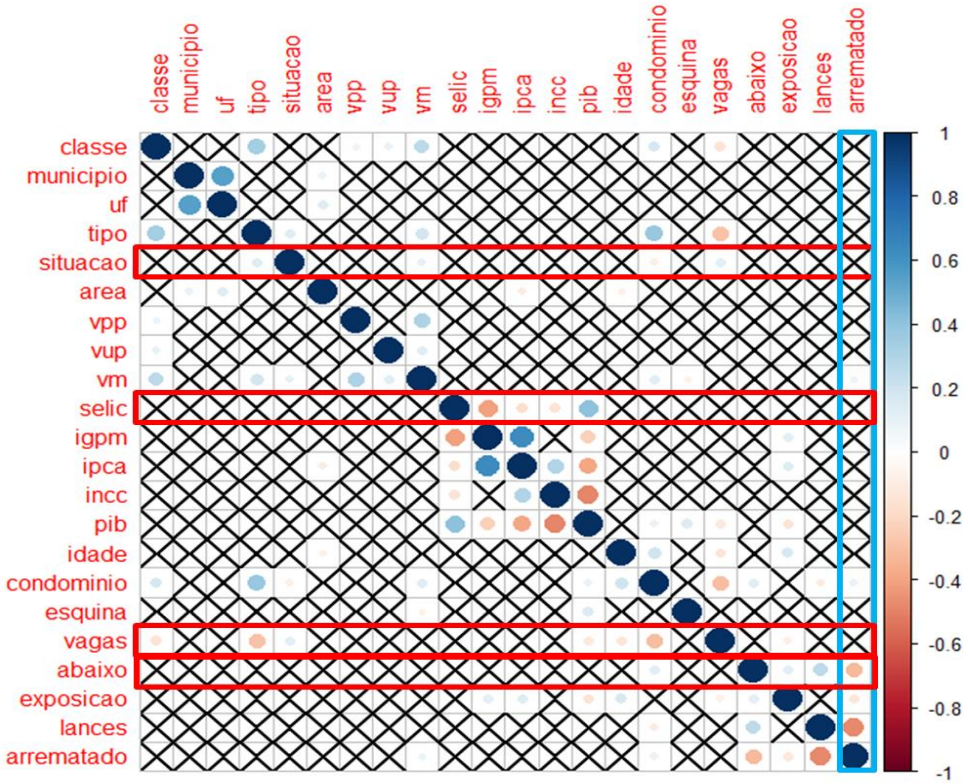
A partir do Gráfico 1, observa-se uma relação de liquidez maior para imóveis com desconto entre 40% e 50%. Ainda assim, a entropia é consideravelmente grande. Ou seja, visualmente, não é possível extrair qualquer tipo de conhecimento. Existem diversos fatores que modificam cada um dos 1.620 imóveis.

Outro modo de visualizar os dados consiste em verificar a correlação estatística dos dados. A correlação ou associação é qualquer relação estatística, causal ou não causal, entre duas variáveis e onde envolva dependência entre elas.

A análise de correlação, segundo Da Silva *et al.* (2017), permite quantificar o quanto uma variável está relacionada com outra, no sentido de determinar a intensidade e a direção dessa relação. Em outras palavras, a correlação indica evidencias estatísticas de uma variável em relação a outra.

Com base nos dados coletados e nos argumentos apresentados foi realizado uma visualização de correlação em linguagem de programação R, empregando a biblioteca “corrplot” no R Studio para verificar o grau de correlação das variáveis selecionadas nesse trabalho e discutidas por outros autores na literatura acadêmica. Abaixo a Figura 24, mostra a correlação dos vinte e dois atributos variáveis da base de dados.

Figura 24 - Matriz de correlação dos atributos variáveis



Fonte: Autor.

A barra lateral da matriz de correlação, indica nas cores azuis e vermelha, o grau de correlação entre os atributos variáveis. Ou seja, azul indica correlação positiva e vermelha correlação negativa.

Os primeiros resultados observados são a correlação das variáveis “arrematado” que refere-se a liquidação ou não desses imóveis com as variáveis “vm” relacionada com o valor de mercado, “condominio” relacionada ao posicionamento dentro ou fora do condomínio, “abaixo” relacionada com o desconto ofertado, “exposicao” relacionada ao tempo no leilão e “lances” que refere-se ao número de lances ofertados em nas respectivas praças de arrematação.

Com relação ao Quadro 1 desse trabalho, os autores Turnbull e Zahirovic-Herbert (2012) analisam que mercado tradicional de imóveis que as variáveis “vaga” e “casas ocupadas”, nesse trabalho entendida como situação de ocupação do imóvel por terceiros, porém com nomenclatura diferente pela variável “situacao”, possuem correlação e afetam a variável “arremate”. Entretanto, com base na matriz de correlação exposta na Figura 25, não existem evidências estatísticas que relacionem ou correlacionem essas variáveis no mercado de leilões.

Por outro lado, os autores Bian *et al.* (2015) relacionaram a variável taxa de desconto sobre a propriedade no mercado tradicional de imóveis, aborda nesse trabalho como “abaixo”, relacionada ao percentual de desconto aplicado sobre o valor de mercado dos imóveis expostos em praças de leilões, com a variável “arremate”. Com base na matriz de correlação exposta na Figura 25, existem evidências estatísticas que relacionem ou correlacionem essas variáveis no mercado de leilões.

Kok *et al.* (2018) relacionaram a variável taxa de juros, aborda nesse trabalho como “selic”, com a variável “arremate”. Entretanto, com base na matriz de correlação exposta na Figura 25, não existem evidências estatísticas que relacionem ou correlacionem essas variáveis no mercado de leilões.

Neste exemplo, a correlação pode ser usada para explicar a relação entre os comportamentos medianos de duas variáveis, permitindo uma análise exploratória dos dados. Porém, no contexto da análise de dados, as medidas de relevância ainda são úteis na etapa de pré-processamento dos dados e, mais especificamente, para selecionar atributos interessantes a serem submetidos a algoritmos de classificação ou agrupamento. Portanto, mostra-se útil para realizar a redução da dimensionalidade (DA SILVA *et al.,* 2017).

5.2 AVALIAÇÃO DO DESEMPENHO DAS TÉCNICAS INTELIGENTES

Após a exposição dos dados, faz-se necessário avaliar as técnicas inteligentes mais promissoras, em relação aos dados de leilão, para posterior aplicação na AHI. Nesse âmbito existem métodos para realizar essa tarefa.

Um dos processo de validação cruzada bastante disseminado na literatura acadêmica, por exemplo, é o método *K-fold*, cujo objetivo principal é estimar o desempenho dos algoritmos AM em qualquer conjunto de dados, particionando a base em tamanhos menores para teste e treino. Nesse sentido, divisões diferentes de dados também levarão a resultados diferentes, o que leva a consistência do modelo preditivo (DA SILVA *et al.,* 2017).

A validação cruzada *k-fold* é uma maneira checar o desempenho de um modelo de AM. Isso envolve repetir o processo de validação cruzada várias vezes e relatar o resultado médio de todas as dobras para todas as execuções. Por outro lado, se calculado usando erro padrão, espera-se que o resultado médio seja uma estimativa mais precisa do verdadeiro desempenho médio do modelo no conjunto de dados.

Com base nos dados de entrada, três técnicas inteligentes foram avaliadas seguindo o critério de 10 dobras, conforme Tabela 1 a seguir, empregando o *software* WEKA:

Tabela 1 - Análise *K-Fold* em 10 dobras

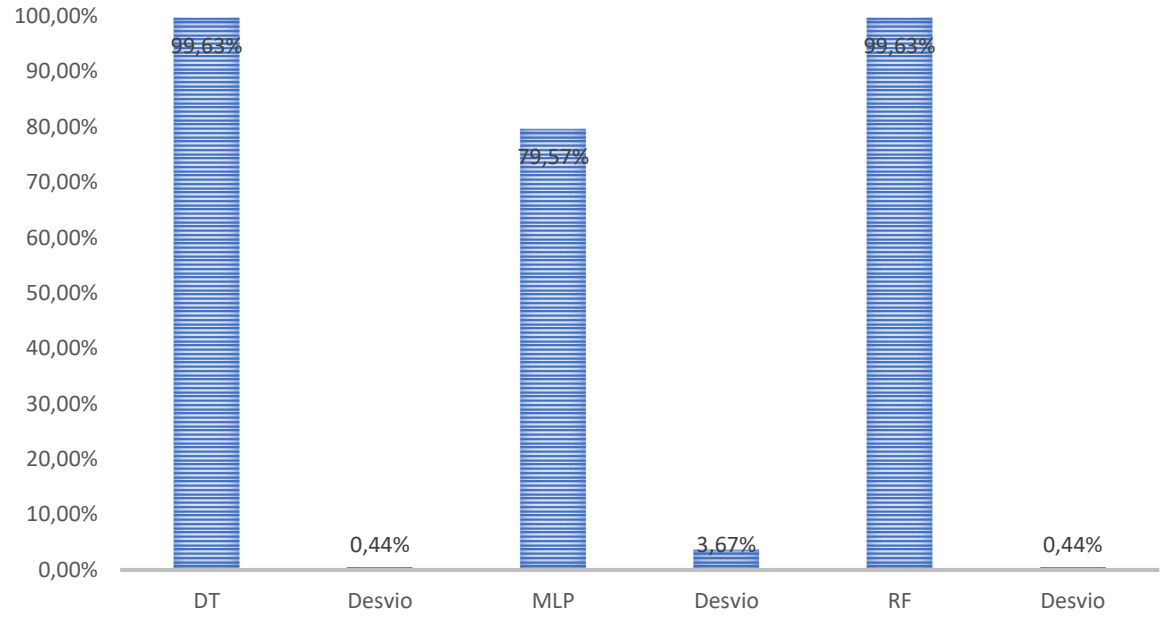
K-Fold	**DT**	**Desvio**	**MLP**	**Desvio**	**RF**	**Desvio**
1	**99,57**	**0,42**	78,77	3,87	**99,57**	**0,42**
2	**99,69**	**0,52**	78,58	3,59	**99,69**	**0,52**
3	**99,51**	**0,64**	79,07	3,03	**99,51**	**0,64**
4	**99,75**	**0,32**	78,46	4,89	**99,75**	**0,32**
5	**99,32**	**0,79**	81,67	3,08	**99,32**	**0,79**
6	**99,81**	**0,30**	79,88	3,22	**99,81**	**0,30**
7	**99,75**	**0,32**	79,20	3,11	**99,75**	**0,32**
8	**99,38**	**0,41**	81,36	5,18	**99,38**	**0,41**
9	**99,69**	**0,33**	79,44	3,19	**99,69**	**0,33**
10	**99,81**	**0,30**	79,26	3,49	**99,81**	**0,30**

Fonte: Autor.

As técnicas Árvore de decisão (*Decision Tree* - DT) e Florestas aleatórias (*Randon Forest* - RF) se desviaram pouco em relação as dobras de validação. Por outro lado, a técnica Rede Neural Artificial (*Multilayer Perceptron* - MLP) além de apresentar um desvio maior, não se ajustou tão bem aos dados imobiliários coletados.

Com base no exposto na Tabela 1 acima, o Gráfico 2 abaixo, ilustra a média dos desvios apresentados das técnicas inteligentes, empregadas nos dados de entrada.

Gráfico 2 - Análise *K-Fold* em 10 dobras



Fonte: Autor.

Por outro lado, outro processo metodológico bastante empregado para estimar o desempenho de algoritmos baseados em AM, denomina-se *Holdout*.

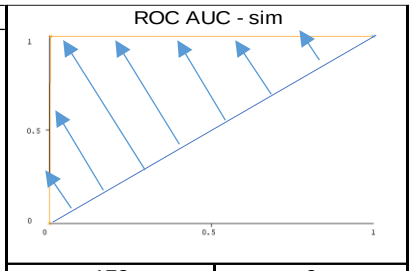
O método *Holdout* pressupõe a criação de dois subconjuntos de dados a partir do conjunto de dados disponível para uso na indução do modelo. Um dos subconjuntos será usado para treinamento por indução do modelo preditivo com 2/3 dos dados e o segundo, para teste, com 1/3 dos dados (AZEVÊDO, 2018).

Segundo Da Silva *et al.* (2017), esse tipo de distribuição empregada para validação pode trazer uma situação favorável ao modelo, porque se sabe que o conjunto de dados é uma amostra de dados possível em um contexto específico, sendo interessante para maximizar a confiabilidade e consistência da técnica inteligente empregada.

O Quadro 10 a seguir mostra resultados promissores e consistentes da técnica de DT no *software* WEKA, empregando os dados de entrada para a predição do arremate imobiliário no método *Holdout*.

Quadro 10 - Árvore de decisão

Modelo	***Decision Tree***
Número total de instâncias	540
Divisão do conjunto	***Holdout***
Treino	2/3 - 1.080 amostras
Teste	1/3 - 540 amostras
Tempo de contrução do modelo (segundos)	0,01
Tempo empregado para testar o modelo (segundos)	0,00
Classificação correta de instâncias	**99,63%** **538**
Classificação incorreta de instâncias	**0,37%** **2**
Estastística Kappa	0,9914
Erro absouto médio	0,0072
Erro médio quadrático da raiz	0,0605
Erro absoluto relativo	1,6526%
Erro quadrático relativo da raiz	13,0232%



Classe	TVP	TFP	Precisão	Recall	F-measure	MCC	ROC	PRC
sim	1,000	0,005	0,989	1,000	0,994	0,992	0,997	0,989
nao	0,995	0,000	1,000	0,995	0,997	0,992	0,997	0,988
Média ponderada	0,996	0,002	0,996	0,996	0,996	0,992	0,997	0,995

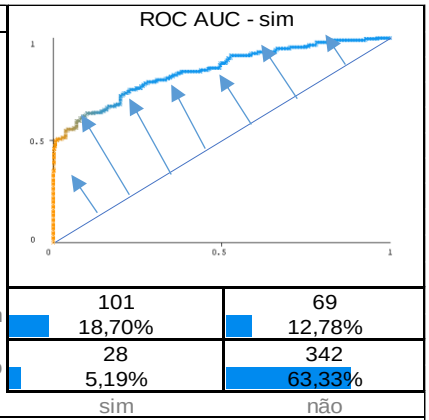
Fonte: Autor.

O modelo DT acertou 538 instâncias corretas e errou apenas 2 classificações. O modelo apresentou boa precisão de acerto para ambas as classes (Sim - arrematado e Não - arrematado), levando em consideração o valor da taxa de falsos positivos e verdadeiros positivos.

O Quadro 11 a seguir mostra os resultados não tão promissores e consistentes da técnica de MLP no *software* WEKA, empregando os dados de entrada para a predição do arremate imobiliário no método *Holdout*.

Quadro 11 - *Multilayer Perceptron*

Modelo	***Multilayer Perceptron***
Número total de instâncias	540
Divisão do conjunto	***Holdout***
Treino	2/3 - 1.080 amostras
Teste	1/3 - 540 amostras
Tempo de contrução do modelo (segundos)	2,81
Tempo empregado para testar o modelo (segundos)	0,00
Classificação correta de instâncias	**82,04%** **443**
Classificação incorreta de instâncias	**17,96%** **97**
Estastística Kappa	0,5546
Erro absouto médio	0,1931
Erro médio quadrático da raiz	0,3906
Erro absoluto relativo	44,3323%
Erro quadrático relativo da raiz	84,0686%



Classe	TVP	TFP	Precisão	Recall	F-measure	MCC	ROC	PRC
sim	0,594	0,076	0,783	0,594	0,676	0,565	0,838	0,792
nao	0,924	0,406	0,832	0,924	0,876	0,565	0,838	0,899
Média ponderada	0,820	0,302	0,817	0,820	0,813	0,565	0,838	0,866

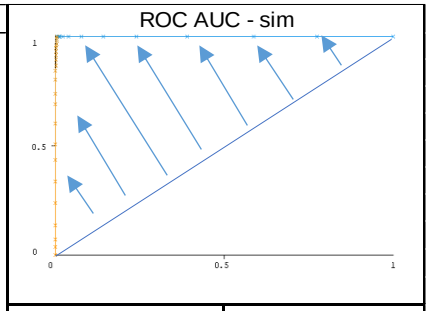
Fonte: Autor.

Apesar da performance ser inferior a técnica DT, o modelo MLP acertou 443 instâncias corretas e errou 97 classificações. O modelo apresentou precisão regular de acerto para ambas as classes (Sim - arrematado e Não - arrematado), levando em consideração o valor da taxa de falsos positivos e verdadeiros positivos.

O Quadro 12 a seguir mostra resultados promissores e consistentes da técnica de RF no *software* WEKA, empregando os dados de entrada para a predição do arremate imobiliário no método *Holdout*.

Quadro 12 - *Random Forest*

Modelo	***Random Forest***	
Número total de instâncias		540
Divisão do conjunto		***Holdout***
Treino		2/3 - 1.080 amostras
Teste		1/3 - 540 amostras
Tempo de contrução do modelo (segundos)		0,13
Tempo empregado para testar o modelo (segundos)		0,01
Classificação correta de instâncias	**99,63%**	**538**
Classificação incorreta de instâncias	**0,37%**	**2**
Estastística Kappa		0,9914
Erro absouto médio		0,0413
Erro médio quadrático da raiz		0,0753
Erro absoluto relativo		9,4750%
Erro quadrático relativo da raiz		16,2188%



	sim	não
sim	170 31,48%	0 0,00%
não	2 0,37%	368 68,15%

Classe	TVP	TFP	Precisão	Recall	F-measure	MCC	ROC	PRC
sim	1,000	0,005	0,989	1,000	0,994	0,991	0,999	0,999
nao	0,995	0,000	1,000	0,995	0,997	0,991	0,999	1,000
Média ponderada	0,996	0,002	0,996	0,996	0,996	0,991	0,999	0,999

Fonte: Autor.

O modelo RF acertou 538 instâncias corretas e errou apenas 2 classificações. O modelo apresentou boa precisão de acerto para ambas as classes (Sim - arrematado e Não - arrematado), levando em consideração o valor da taxa de falsos positivos e verdadeiros positivos.

Com objetivo de visualizar de forma concatenada os resultados das validações pelo método *K-fold*, a Tabela 2 apresenta os principais resultados das técnicas empregadas nesse trabalho acadêmico.

Tabela 2 - Sinopse das técnicas inteligentes na validação *K-Fold* em 10 dóbras

***K-Fold* | 10**	**TVP**	**TFP**	**Precisão**	**Recall**	**F-Measure**	**ROC**	**PRC**
DT	0,996	0,002	0,996	0,996	0,996	0,996	0,993
MLP	0,785	0,274	0,785	0,785	0,785	0,826	0,843
RF	0,996	0,002	0,996	0,996	0,996	0,998	0,996

Fonte: Autor.

Com objetivo de visualizar de forma concatenada os resultados das validações pelo método *Holdout*, a Tabela 3 apresenta os principais resultados das técnicas empregadas nesse trabalho acadêmico.

Tabela 3 - Consistência das técnicas inteligentes na validação *Holdout*

***Holdout* | 2/3**	**TVP**	**TFP**	**Precisão**	**Recall**	**F-Measure**	**ROC**	**PRC**
DT	0,996	0,002	0,996	0,996	0,996	0,997	0,995
MLP	0,820	0,302	0,817	0,820	0,813	0,838	0,866
RF	0,996	0,002	0,996	0,996	0,996	0,999	0,999

Fonte: Autor.

A composição da AHI será composta pela técnica inteligente RF, por conta da consistência apresentada nos experimentos realizados e pela técnica inteligente DT, para gerar as regras de decisão que melhor se ajustam aos dados coletados.

O primeiro ponto de vista é que os dados possuem entropia e as técnicas inteligentes selecionadas são baseadas em árvores de decisão. Isso significa que esses modelos se ajustam melhor a esses tipos de dados conforme visto no referencial teórico desse trabalho acadêmico. O segundo ponto é que a técnica inteligente RF evita o *overfitting* por conta de seus fundamentos e princípios de aleatoriedade. Pontos esses que melhoram a simbiose entre as técnicas escolhidas para compor a AHI.

O motivo da escolha ou da opção pela construção da AHI, se dá pelo fato da variável “lances”, ao ser classificada somente com uma técnica inteligente, pelas chances de haver lances ou não, classifica todas as instâncias como sendo não arrematadas por considerar as hipóteses da base de dados. Abaixo de 50% por cento de chances de ter um lance, a técnica inteligente zera a predição do número de lances e tão logo a técnica classifica todos os imóveis como sendo não arrematados.

No ponto de vista de regressão, para a predição numérica de lances, a técnica inteligente leva em consideração as hipóteses de treinamento e sempre acaba predizendo um valor para lances acima de zero. Em outras palavras, a técnica inteligente classificaria todas as instâncias como arrematadas. Com base nos argumentos apresentados, existe a necessidade de realizar uma combinação das técnicas inteligentes em uma AHI para a classificação da liquidez imobiliária no setor estudado.

5.3 ANÁLISE DA ARQUITETURA HÍBRIDA DO MODELO

De acordo com a pesquisa de Goldschmidt e Passos (2005), as técnicas inteligentes podem ser combinadas para produzir as chamadas arquiteturas híbridas. A grande vantagem desse sistema se deve à sinergia resultante da combinação de duas ou mais técnicas inteligentes.

Com base na composição e disposição das técnicas inteligentes propostas para essa AHI, conforme Goldschmidt e Passos (2005), classifica-se a AHI como tipo auxiliar. Em outras palavras, a técnica inteligente RF invoca a técnica DT apenas para gerar regras que separem os dados em arrematados e não arrematados com base nos lances preditos. Ao término do processo preditivo a técnica RF classifica com mais robustez a liquidez dos imóveis urbanos em leilão.

Por outro lado, algumas informações, para a predição de dados novos, não são disponibilizadas previamente a predição do arremate imobiliário em leilões como:

I. “abaixo” - Relativo ao desconto oferecido para liquidar em menos tempo o imóvel;
II. “exposicao” - Relativo ao tempo que um imóvel fica exposto em um leilão de imóveis;
III. “lances” - Relativo ao número de lances que um imóvel venha ter em um leilão de imóveis;

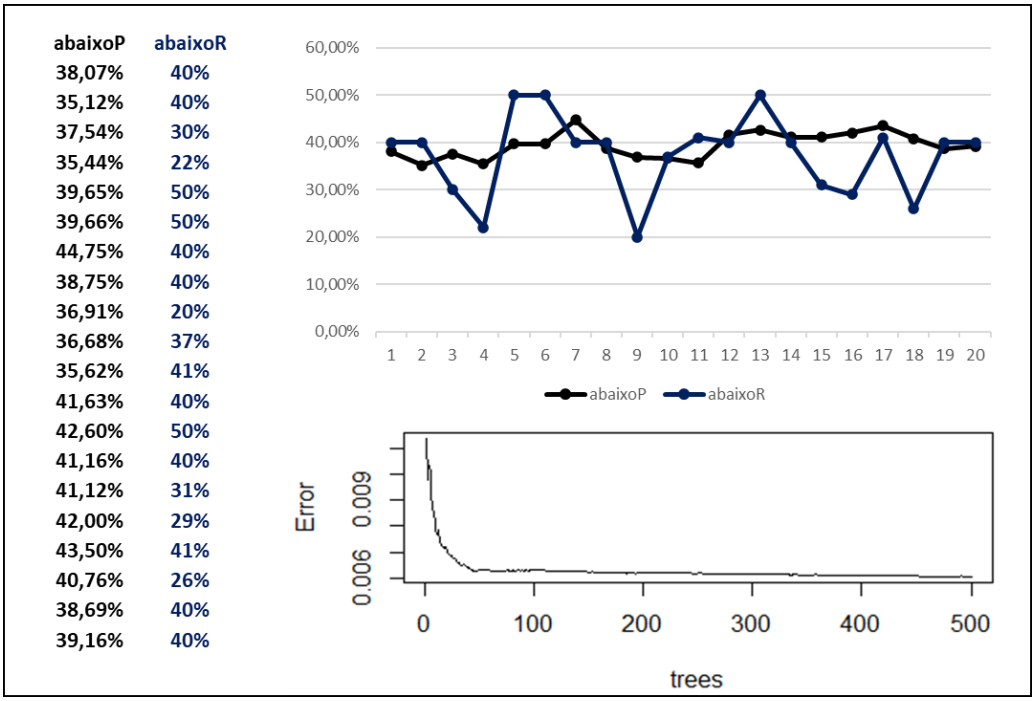
O modelo híbrido possui 18 atributos de entrada e 8 de saída, computando os 3 atributos acima descrito, que compuseram um total de 26 atributos com a classificação da liquidez a ser mapeada pela MPID posteriormente. Logo, observa-se que haverá enriquecimento da base de dados por consequência das predições realizadas pela AHI.

Com base nas informações apresentadas, essa análise foi subdividida em 6 etapas de construção da AHI para classificação de liquidez de imóveis urbanos em leilões.

5.3.1 Etapa 1 - Percentual abaixo

Por regressão linear múltipla, com 18 atributos de entrada e empregando RF, busca-se prever o percentual abaixo (abaixoP) praticado pelo mercado de leilões e comparar com o percentual abaixo real (abaixoR). Logo, a Figura 25 abaixo, avalia e compara o desconto real dado, pelo predito.

Figura 25 - Resultados da regressão para predição do percentual de desconto



abaixoP	abaixoR
38,07%	40%
35,12%	40%
37,54%	30%
35,44%	22%
39,65%	50%
39,66%	50%
44,75%	40%
38,75%	40%
36,91%	20%
36,68%	37%
35,62%	41%
41,63%	40%
42,60%	50%
41,16%	40%
41,12%	31%
42,00%	29%
43,50%	41%
40,76%	26%
38,69%	40%
39,16%	40%

Fonte: Autor.

Com o total de 500 árvores geradas no modelo de regressão da técnica RF e número de variáveis experimentadas totalizando sete, a soma residual dos quadrados, usada para ajudar o modelo a decidir se um modelo estatístico é um bom ajuste para seus dados, foi de 0.00043 com o percentual de variância explicada de 95,58%.

A variável “abaixo” possui na base coletadas valor mínimo de desconto de 10%, valor médio de 41% e máximo de 87% em alguns casos.

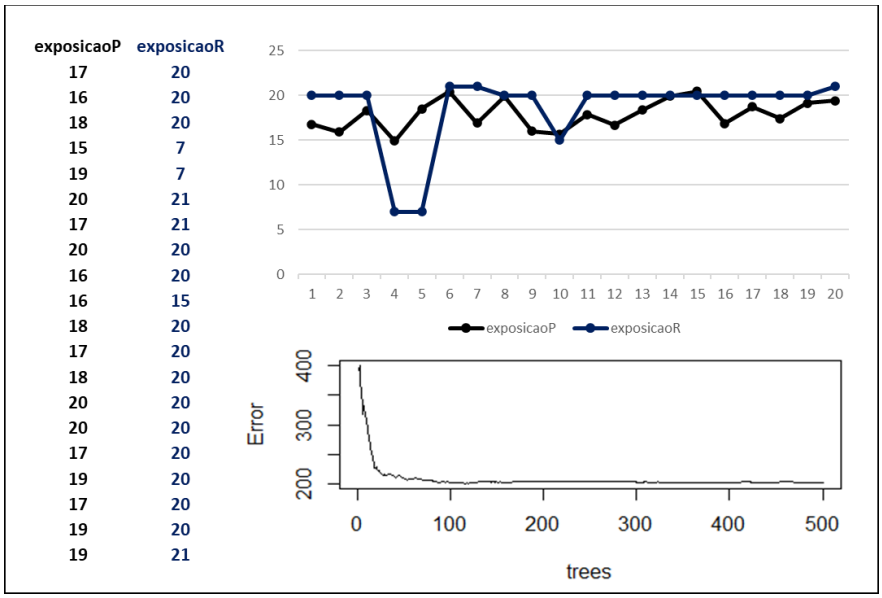
Com base na regressão realizada, e no modelo implementado, as cinco variáveis que mais ajudam o modelo a prever o percentual de desconto praticado pelo mercado de leilões são: valor de última praça, valor de primeira praça, área, valor de mercado e idade do imóvel.

5.3.2 Etapa 2 - Tempo de exposição

Por regressão linear múltipla, agora com 19 atributos de entrada e empregando RF, busca-se prever o tempo de exposição praticado pelo mercado de leilões. Essa variável foi tratada como conceito TOM, abordado na literatura e comentado nesse trabalho acadêmico, porém, denominada como a variável “exposicao” na base de dados coletada.

A Figura 26 abaixo, avalia e compara o tempo real de exposição do imóvel no mercado de leilões pelo predito.

Figura 26 - Resultados da regressão para predição do tempo de exposição



Fonte: Autor.

Com o total de 500 árvores geradas no modelo de regressão da técnica RF e número de variáveis experimentadas totalizando sete, a soma residual dos quadrados, usada para ajudar o modelo a decidir se um modelo estatístico é um bom ajuste para seus dados, foi de 225,43 com o percentual de variância explicada de 35,38%.

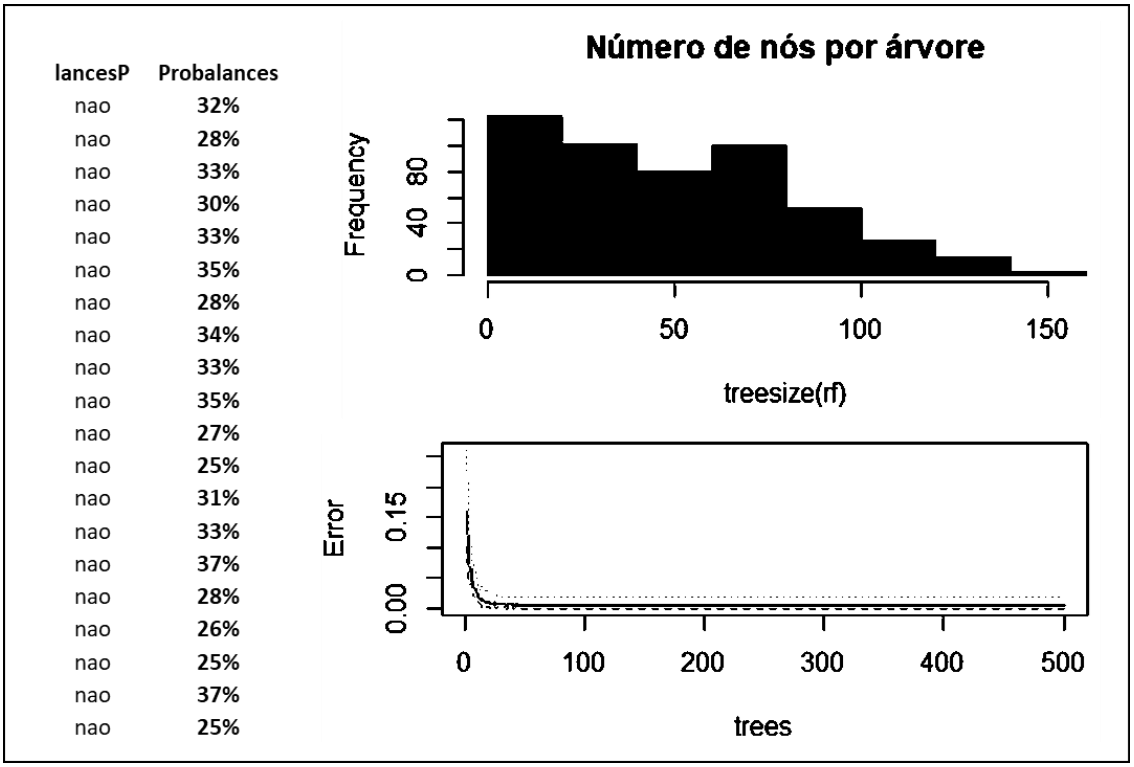
A variável “exposicao” possui na base coletada valor mínimo de desconto de zero dias, valor médio de vinte dias e máximo de cento e setenta e três dias.

Com base na regressão realizada, e no modelo implementado, as cinco variáveis que mais ajudam o modelo a prever o tempo de exposição dos imóveis no mercado de leilões são: idade, valor de primeira praça, percentual abaixo, incc e valor de mercado.

5.3.3 Etapa 3 - Possibilidade e probabilidade de lances

Por classificação, agora com 20 atributos de entrada e empregando RF, busca-se prever a possibilidade e a probabilidade de um imóvel obter lance pelo mercado de leilões. Logo, a Figura 27 abaixo, compara imóveis reais em leilões que tiveram ou não lances com as classes previstas.

Figura 27 - Resultados das classificações para lances



Fonte: Autor.

Com o total de 500 árvores geradas no modelo de classificação da técnica RF e número de variáveis experimentadas totalizando quatro, teve erro *out-of-bag* (OOB), também chamado de estimativa *out-of-bag*, cujo método serve para medir o erro de previsão de florestas aleatórias, de 0,47%.

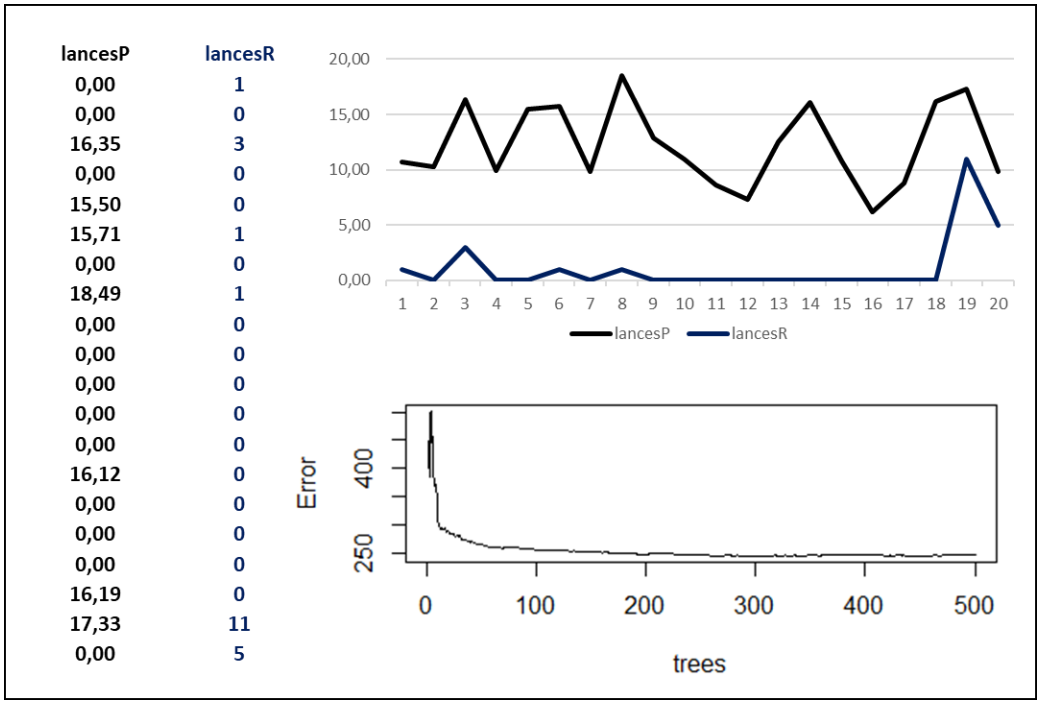
A variável “lances” possui na base coletadas valor mínimo de desconto de zero lances, valor médio de cinco lances e máximo de duzentos e quarenta lances.

Com base na classificação realizada, e no modelo implementado, as cinco variáveis que mais ajudam o modelo a prever o a probabilidade e a possibilidade de um imóvel ter lances em um leilão de imóveis são: percentual de desconto ou abaixo, valor de mercado, valor de última praça, valor de primeira praça e área.

5.3.4 Etapa 4 - Quantidade de lances

Por regressão linear múltipla, agora com 22 atributos de entrada e empregando RF, busca-se prever a quantidade de lances que um imóvel possa ter em um leilão de imóveis. A Figura 28 abaixo, avalia e compara lances reais com lances preditos.

Figura 28 - Resultados da regressão para predição de lances



Fonte: Autor.

Com o total de 500 árvores geradas no modelo de regressão da técnica RF e número de variáveis experimentadas totalizando sete, a soma residual dos quadrados, usada para ajudar o modelo a decidir se um modelo estatístico é um bom ajuste para seus dados, foi de 243,20 com o percentual de variância explicada de 23,04%.

A variável “exposicao” possui na base coletadas valor mínimo de desconto de zero lances, valor médio de cinco lances e máximo de duzentos e quarenta lances.

Com base na regressão realizada, e no modelo implementado, as cinco variáveis que mais ajudam o modelo a prever a quantidade de lances dos imóveis no mercado de leilões são: valor de mercado, percentual abaixo, valor de primeira praça, área e valor de última praça.

5.3.5 Etapa 5 - Associação de lances preditos e reais da base de dados

A composição da AHI será composta pela técnica inteligente RF, por conta da consistência apresentada nos experimentos realizados e pela técnica inteligente DT, para gerar as regras de decisão que melhor se ajustam aos dados coletados.

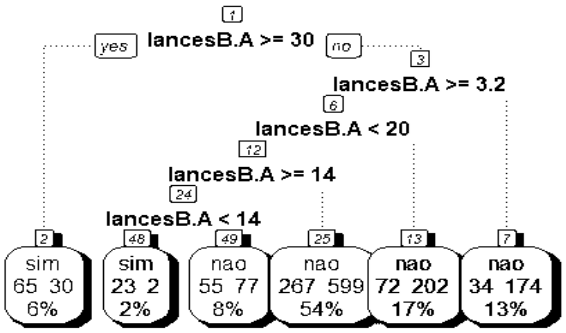
A variável “lances”, ao ser classificada somente com uma técnica inteligente, pelas chances de haver lances ou não, classifica todas as instâncias como sendo não arrematadas por considerar as hipóteses advindas da base de dados. Abaixo de 50% por cento de chances de ter um lance, a técnica inteligente zera a predição do número de lances e tão logo a técnica classifica todos os imóveis como sendo não arrematados.

Por outro lado, sob o ponto de vista de regressão, para a predição do número de lances, a técnica inteligente leva em consideração as hipóteses de treinamento e sempre acaba predizendo um valor para lances acima de zero como visto nos experimentos anteriores. Em outras palavras, a técnica inteligente classificaria todas as instâncias como arrematadas.

Com base nos argumentos apresentados, existe a necessidade de realizar uma combinação das técnicas inteligentes em uma AHI para a classificação da liquidez imobiliária no setor estudado. Para tanto, a técnica de DT comparou lances preditos, pela técnica inteligente RF, com imóveis arrematados e não arrematados em leilão, gerando regras que estabelecem as margens de lances preditos que de fato, podem ser convertidos em lances reais.

Através da biblioteca “rpart” da linguagem de programação R, a Figura 29 abaixo ilustra as regras criadas pela técnica inteligente DT, que separam os dados.

Figura 29 - Regras de lances



Fonte: Autor.

Antes da etapa 6, agora com 23 atributos de entrada e empregando RF, cria-se uma nova coluna de atributos, cujo objetivo é zerar lances previsto abaixo de 14 lances, conforme regras geradas pela técnica inteligente DT.

5.3.6 Etapa 6 - Classificação do arremate e probabilidades do evento

Por classificação, agora com 24 atributos na base de dados, busca-se classificar a liquidez desses imóveis através do arremate e da probabilidade gerada por essa classificação. A Tabela 4 abaixo, compara resultados reais de imóveis arrematados com as classificações preditas em leilões, bem como as probabilidades de cada instância analisada.

Tabela 4 - Resultados das classificações do arremate em leilões

vup (R$)	lancesP	lancesA	lancesR	arrematado P	arrematado R	arremateProba
615.050,04	10,72	0,00	1	nao	sim	13,10%
536.733,36	10,28	0,00	0	nao	nao	6,00%
3.742.412,02	16,35	16,35	3	sim	sim	76,50%
375.681,20	9,92	0,00	0	nao	nao	7,70%
128.291,18	15,50	15,50	0	sim	nao	75,10%
73.659,38	15,71	15,71	1	sim	sim	76,70%
1.823.005,47	9,83	0,00	0	nao	nao	9,80%
125.007,68	18,49	18,49	1	sim	sim	72,80%
1.988.040,50	12,86	0,00	0	nao	nao	9,50%
983.488,19	10,93	0,00	0	nao	nao	13,30%
188.140,72	8,64	0,00	0	nao	nao	5,00%
570.960,18	7,32	0,00	0	nao	nao	4,60%
521.818,59	12,56	0,00	0	nao	nao	14,70%
102.212,66	16,12	16,12	0	sim	nao	77,90%
61.799,87	10,76	0,00	0	nao	nao	9,00%
411.250,00	6,17	0,00	0	nao	nao	3,90%
496.940,47	8,82	0,00	0	nao	nao	6,50%
5.048.061,76	16,19	16,19	0	sim	nao	77,60%
146.937,86	17,33	17,33	11	sim	sim	79,00%
164.678,60	9,81	0,00	5	nao	sim	7,40%

Fonte: Autor.

Com o total de 500 árvores geradas no modelo de classificação da técnica RF e número de variáveis experimentadas totalizando quatro teve erro *out-of-bag* (OOB), também chamado de estimativa *out-of-bag*, cujo método serve para medir o erro de previsão de florestas aleatórias, de 0,37%.

A variável “arrematado” possui na base coletada valor arrematado e não arrematado.

Com base na classificação realizada, e no modelo implementado, as cinco variáveis que mais ajudam o modelo a prever o a probabilidade e a possibilidade de um imóvel ter lances em um leilão de imóveis são: lance, valor de mercado, percentual abaixo, valor de primeira praça e área.

Para mapear e sintetizar os resultados preditivos da AHI, a seguir será mostrado como a MPDI apoia as decisões, facilitando o entendimento e interpretação das predições.

5.4 APLICAÇÃO DAS CLASSIFICAÇÕES NA MPID

Abaixo o modelo MPID a ser empregado após o término do modelo final preditivo, cujos os dados serão preditos e interpretados conforme Figura 30 abaixo:

Figura 30 - MPID no apoio das previsões de arremate do modelo

	Risco de ameaças nao						Risco de oportunidades sim					
Probabilidade	40%	0	0	0	0	0	0	0	0	0	0	50%
	30%	0	0	0	0	0	0	0	0	0	0	60%
	20%	1	2	0	0	0	1	4	0	0	2	70%
	10%	2	1	2	1	1	0	0	0	0	0	80%
	5%	0	1	1	1	0	0	0	0	0	0	90%
	0%	Muito alto	Alto	Moderado	Baixo	Muito baixo	Muito baixo	Baixo	Moderado	Alto	Muito alto	100%
		Impacto financeiro										Probabilidade

Fonte: Autor.

Com a MPID pode-se extrair, em termos de análise quantitativa da carteira de imóveis dos 20 imóveis novos que:

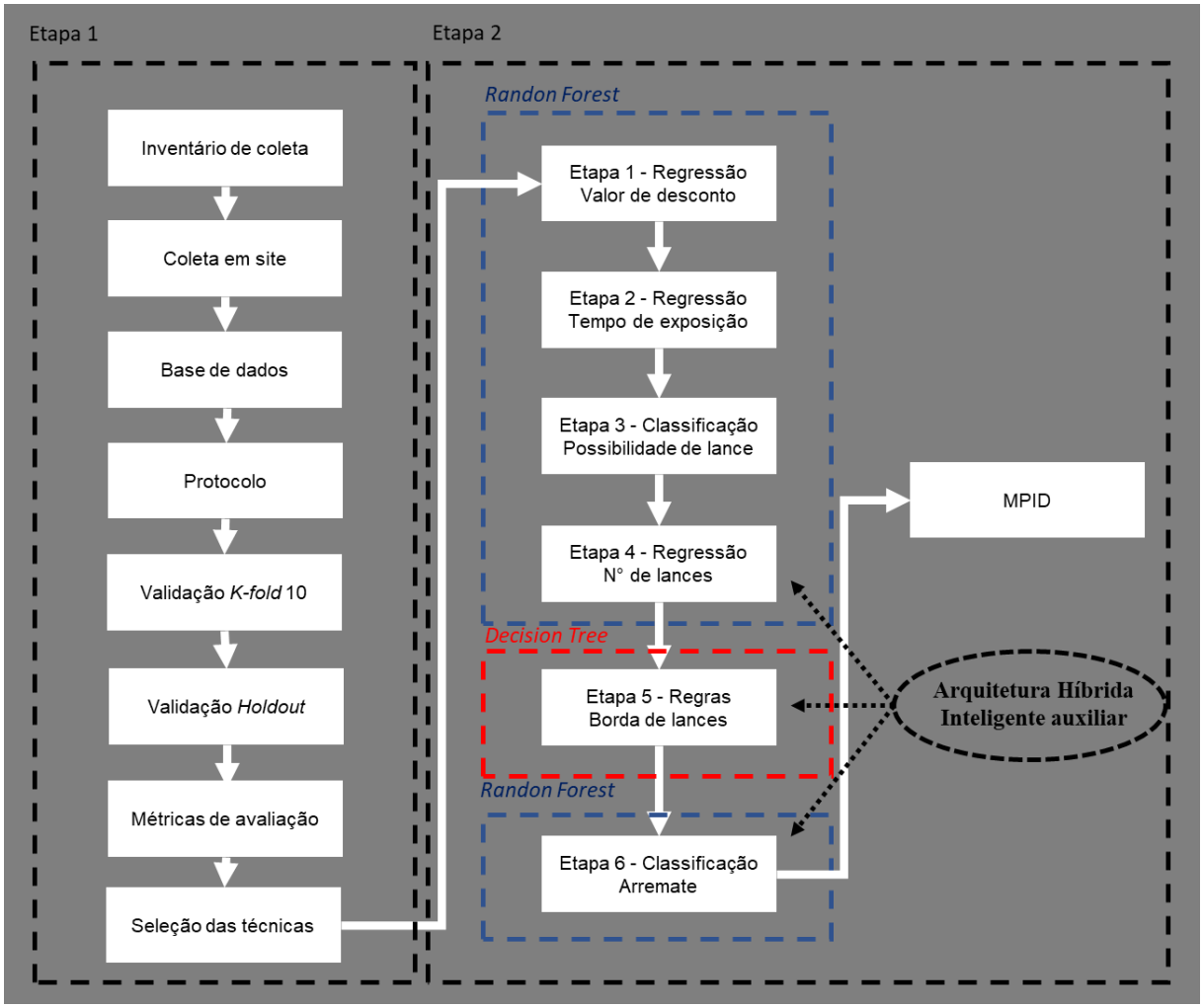
I. possui 6 ameaças, ou seja, imóveis com risco negativo de liquidez;
II. possui 7 imóveis, com pouco grau de incerteza, no entanto, com risco negativo de liquidez;
III. não possui nenhum imóvel, com boa margem de incerteza, no entanto, com risco negativo de liquidez;
IV. possui 1 imóveis, com boa margem de incerteza, no entanto, com risco positivo de liquidez;
V. possui 4 imóveis, com pouco grau de incerteza, no entanto, com risco positivo de liquidez;
VI. possui 2 oportunidades, ou seja, imóveis com risco positivo de liquidez.

Com a MPID pode-se extrair, em termos de valores monetários e percentuais que 48% do valor monetário, correspondente a quantia global de R$ 8.737.587,19, encontra-se em situação de ameaça. Ou seja, foram classificados como imóveis com probabilidades de não serem arrematados em um leilão de imóveis. Por outro lado, 52% do valor monetário correspondente a quantia global de R$ 9.366.582,54.

Ao todo foi mapeada a quantia monetária de R$ 18.104.169,73 em 20 imóveis urbanos em leilão. De acordo com as análises, conclui-se que existem mais ameaças do que oportunidades em termos de quantidade de imóveis. Logo, entende-se que esse conjunto de dados representa um risco a um leilão e a um possível ou eventual credor.

Abaixo a Figura 31 faz uma sinopse do experimento mostrando em detalhes a composição da AHI para posterior aplicação na MPID.

Figura 31 - Condesado do experimento realizado



Fonte: Autor.

5.5 VERIFICAÇÃO DAS PROPOSIÇÕES DA PESQUISA

A consolidação das proposições visa estabelecer uma análise mais aprofundada e fundamentada dos principais resultados de cada experimento, em dois momentos distintos. O primeiro considera analisar os resultados das classificações feitas pela AHI. O segundo consiste em verificar o mapeamento das classificações pela MPID no apoio das decisões a serem tomadas.

Treinando a técnica RF da AHI com o número de árvores padrão da biblioteca “randomForest” de 500 árvores, não se obtém um *score* nos 20 imóveis de teste de 75%. Portanto, quanto maior o número de árvores empregada nas etapas preliminares preditivas, melhor serão as previsões e mais robusto ficará a AHI. Foi possível obter esse *score* aplicando 900 árvores para construção do modelo.

Levando em conta o que foi observado no experimento, a nível de conhecimento gerado, associa-se que imóveis com um determinado número de lances previstos, podem de fato ter um lance real. Portanto, define-se nesse trabalho, um novo conceito denominado borda de lances, cujo objetivo é encontrar a amplitude de lances previstos que determinam se um imóvel de fato terá lances reais em um leilão. A avaliação do conhecimento extraído é útil e pode ser aplicado nos setores de leilão e bancário.

Portanto, com relação à proposição 1, que visa estabelecer a possibilidade de se classificar a liquidez imobiliária urbana, a partir de aplicação de AHI, pode-se afirmar que é válida essa proposição por conta de os resultados apresentados serem consistentes e robustos.

Ainda assim, é possível obter resultados ainda mais consistentes fazendo a técnica inteligente RF criar mais árvores. Não foi observado nenhuma diferença visível quando se aplicou 2.000 árvores. No contexto desse experimento, foi realizado com 900 árvores para que fosse possível verificar a partir de quantas árvores o modelo acerta mais.

O segundo momento consistiu em verificar o mapeamento das classificações pela MPID no apoio das decisões a serem tomadas.

A MPID além de mapear uma classificação em risco ou em ameaça, fornece a probabilidade de o evento ocorrer, sendo muito útil no âmbito estratégico da tomada de decisão. Tão logo se verifica tal aplicação, o modelo entrega outras análises como a divisão de 6 setores de risco: zona de ameaça, zona intermediária de ameaça, zona nebulosa de ameaça, zona nebulosa de oportunidade, zona intermediária de oportunidade e zona de oportunidades. Por outro lado, a MPID permite quantificar os imóveis em cada um dos quadrantes determinados, facilitando a leitura e entendimento do decisor.

Outra possível análise que pode ser feita é a distribuição percentual desses imóveis classificados. Com base na distribuição quantitativa feita pela MPID, é possível estipular o percentual com base no número de imóveis analisados.

Essa análise pode ser comparada com a distribuição percentual quantitativa e a distribuição monetária para comparação do equilíbrio desses imóveis. Seja a quantidade que for. Esse processo pode ser feito com auxílio de planilhas que automatizem essas análises com a função “procv” (procura valor). Essa análise mostra que nem sempre a distribuição quantitativa é igual à distribuição monetária, pois, os imóveis são diferentes bem como suas características e preço.

Deste modo, o emprego de Matrizes de Probabilidade e Impacto Duplas (MPID), podem apoiar o processo de tomada de decisão, com base na análise da liquidez imobiliária urbana no setor de leilões empregando AHI.

Com base nos argumentos apresentados, essa pesquisa acadêmica verificou e confirmou a possibilidade de se classificar a liquidez imobiliária urbana, a partir de aplicação de AHI e mapeando os resultados com uma MPID no apoio do processo de tomada de decisão no setor de leilões imobiliários em ameaças e oportunidades, estabelecendo resposta a questão de pesquisa: **como classificar liquidez imobiliária urbana em leilões através de AHI, apoiando as decisões com MPID?**

6 CONCLUSÕES

Um aspecto importante determinado na pesquisa foi o estabelecimento do inventário para os dados de entrada do experimento científico realizado, que possibilitaram confirmar as propostas formuladas nesse trabalho acadêmico. O objetivo de seu desenvolvimento possibilitou determinar as variáveis fundamentais de mercado com especialista, entender o processo de transação de imóveis no setor de leilões, onde são disponibilizados esses dados e quais as custas envolvidas no processo de liquidação de imóveis. Pode-se considerar conhecimento especialista embarcado no inventário.

Em virtude dos dados que foram coletados em campo foi necessário realizar enriquecimento dos dados a serem preditos. Isso ocorreu, pois, uma parcela desses dados não é disponibilizada previamente a uma praça de arremate no setor de leilões. Os dados não disponíveis são o percentual abaixo de desconto, tempo de exposição e lances. Isso implica que a AHI realizou etapas preditivas anteriores a predição e classificação final do arremate imobiliário.

Durante o desenvolvimento deste trabalho, verificou-se a importância da variável lance e do auxílio fundamental das regras das DT na atuação conjunta com o algoritmo RF na AHI, classificando de modo consistente o arremate imobiliário. A MPDI sumarizou essa classificação em ameaças e oportunidades, subsidiando assim as decisões.

Levando em conta os 18 atributos dos dados de entrada, dispostos na Tabela 2 desse trabalho acadêmico, ao final do processo de classificação, a AHI forneceu o percentual de desconto sobre o valor de mercado, o tempo de exposição que o imóvel pode ficar exposto em uma praça de leilão imobiliário, o número de lances que esse imóvel eventualmente possa ter, os limites de lances previstos que serão de fato convertidos em um lance e a probabilidade das classificações sobre o arremate de imóveis no setor de leilões.

Na etapa de condução dos experimentos foi possível notar que os algoritmos baseados em árvores de decisão, obtiveram mais êxito por conta da entropia dos dados. Deste modo, foi necessário combinar as técnicas DT e RF, tornando o modelo mais robusto em termos de acurácia e demonstrando as regras criadas de separação dos dados.

Os experimentos com MLP, apesar da popularidade dessa técnica inteligente no âmbito científico, não apresentou boas condições de acuracidade e consistência dos resultados. Em bases maiores, a MLP terá maior custo computacional, quando comparado ao tempo de construção do modelo de DT e RF, levando em conta o que foi observado nos experimentos.

Assim foi considerado que todos os experimentos que empregaram as técnicas inteligentes para minerar os dados de leilão foram bem-sucedidas em classificar a liquidez dos imóveis, ainda que a acuracidade tenha sido menor quando aplicado a MLP em relação aos algoritmos DT e RF que compõe a AHI. Nesse sentido, em termos de acuracidade, a AHI retornou com uma taxa de acerto de 75%, sendo que classificou corretamente 15 imóveis de 20.

O mercado de leilões transaciona em grande parte dos ativos movimentados nesse setor, imóveis entre as partes interessadas. Logo, um aspecto importante a ser considerado nesse trabalho é a incerteza que o comprador irá dar um lance. Essa incerteza pode ser mensurada e observada através dos dados coletados e da AHI, pois, em parte foi eliminada a falta de conhecimento sobre eventos futuros desses imóveis. Portanto, o modelo MPID proposto por HILLSON (2002), possibilitou mapear as ameaças e as oportunidades desse setor na tomada de decisões.

Nas bases de dados do mercado imobiliário urbano existe uma grande quantidade de dados, dos quais vários padrões são difíceis de descobrir pelos métodos convencionais, suportados por planilhas de cálculo. Logo, em uma perspectiva prática, esse trabalho de pesquisa fornece uma nova ferramenta de apoio as decisões para o leiloeiro ou bancário avaliar a liquidez de suas carteiras de imóveis urbanos através de um único espectro. Logo, a contribuição se dá na medida que essas carteiras são grandes para analisar, o que dificulta a decisão e facilita a adesão a essa AHI de suporte, apoiada por uma MPID.

O diferencial entre aplicar a MPID, ao invés de uma MPI consiste em discriminar o risco não só como uma ameaça, mas poder apontar as oportunidades que um determinado imóvel apresenta em termos de liquidez. Isso ocorre pois a MPID mostra as margens probabilísticas do arremate, propiciando por exemplo, o leiloeiro e o bancário, aderirem a um valor mais atraente de entrada de um imóvel em uma praça de arrematação.

O resultado da aplicação da MPID descriminou e separou os 20 imóveis em 6 zonas de risco distintas. A primeira foi a zona de ameaça com 6 imóveis enquadrados. A segunda foi a zona de ameaça intermediária de ameaça com 7 imóveis enquadrados. A terceira é a zona nebulosa de ameaças com nenhum imóveis enquadrados. A quarta é a zona nebulosa de oportunidades com 1 imóveis enquadrados. A quinta é a zona intermediária de oportunidades com 4 imóveis enquadrados. A sexta é a zona de oportunidades com 2 imóveis enquadrados.

Baseado nos enquadramentos da MPID, a AHI mapeou 20 imóveis que somam a quantia monetária de R$ 18.104.169,73. A MPID segregou imóveis sem chances de serem liquidados, cuja soma foi computada em R$ 8.737.587,19 (48% do capital em risco compondo 13 imóveis) e imóveis com chances de serem liquidados, cuja soma foi computada em R$ 9.238.291,36 (35% do capital assegurado pelas margens de arrematação compondo apenas 7 imóveis).

Sendo assim, de acordo com as análises realizadas, concluiu-se que os 20 imóveis analisados, dado o equilíbrio apresentado, possui mais ameaças do que oportunidades, apesar da distribuição monetária ser quase equivalente. Vale ressaltar que essa distribuição dos 20 dados novos, equivale a distribuição real de mercado verificada na fase de coleta de 2/3 em risco do tipo ameaça.

Em vista dos argumentos apresentados, este trabalho possibilitou atingir o seguinte objetivo geral: **avaliar técnicas inteligentes e desenvolver uma AHI para classificação de liquidez imobiliária urbana em leilões, apoiando o processo de tomada de decisão com MPID.** Considera-se, então, o objetivo geral atingido pois os experimentos computacionais classificaram a liquidez dos imóveis em leilões para a tomada de decisões com apoio da MPID, que justificam sua aplicação e possibilitam a continuação deste trabalho.

Levando em conta o que foi observado no experimento, a nível de conhecimento gerado, associa-se que imóveis com um determinado número de lances previstos, podem de fato ter um lance real. Portanto, define-se nesse trabalho, um novo conceito denominado borda de lances, cujo objetivo é encontrar a amplitude de lances previstos que determinam se um imóvel de fato terá lances reais em um leilão. A avaliação do conhecimento extraído é útil e pode ser aplicado nos setores de leilão e bancário.

Em termos de contribuições práticas e acadêmicas, a pesquisa realizada fornecerá a base de dados contendo 1.620 imóveis urbanos em leilão para que novos experimentos sejam realizados empregando técnicas diferentes das aqui vistas, o próprio modelo híbrido como uma ferramenta para usos práticos na vida real e acadêmicos de como foi avaliado e construído, sem qualquer problema ou impedimento legal com relação aos dados.

Na prática, um leilão pode evitar custas, em caso de um imóvel não ser liquidado, como despesas administrativas, despesas de avaliação, custas jurídicas, de publicação e marketing. Apensar de não termos dados desse levantamento, é inegável a contribuição desse trabalho tendo em vista que 67,78%, representando 1.098 imóveis que não foram arrematados.

Tal constatação permite ir para além de uma matriz, o que abre espaço para novas pesquisas voltadas à otimização e aplicação dos resultados usando interpolação polar da MPDI em mapas de calor com coordenadas geográficas do imóveis, o que não estava previsto durante o planejamento da pesquisa, mas é um estímulo à continuidade desta pesquisa.

Dado o exposto, a continuidade deste trabalho se dá pelo aperfeiçoamento das técnicas empregadas, pela aplicação de novas técnicas, ou pelo estudo empírico, avaliando a aplicação dos modelos desenvolvidos em casos práticos cumprindo com o objetivo de estimular novas pesquisas.

Esta pesquisa limitou-se a classificar liquidez de imóveis urbanos em leilões apoiando o processo de tomada de decisão com MPID. Através do desenho experimental empregado, novos estudos podem ser realizados com outros tipos de patrimônios como exemplo: imóveis rurais, máquinas e equipamentos.

Portanto, os resultados apresentados nesta pesquisa não têm a pretensão de esgotar o assunto, mas sim de permitir que o mesmo modelo conceitual seja operacionalizado em novas pesquisas e que possibilitem a identificação de técnicas capazes de produzir resultados mais eficientes do que os encontrados até o momento.

REFERÊNCIAS BIBLIOGRÁFICAS

ABECIP. Informativos mensais. **Abecip**, 2020. Disponível em: <https://www.abecip.org.br/imprensa/informativos-mensais>. Acesso em: 22 set. 2020.

ALBANO, M. F; NAPOLITANO, D. M. R. Técnicas inteligentes na classificação e análise de portfólios de crédito imobiliário. **RIST - Revista Ibérica de Sistemas e Tecnologias de Informação**, v.31, p. 452-464, 2020.

ALENCAR, Leonidas Pena *et al.* Avaliação de métodos de estimativa da evapotranspiração de referência para três localidades no norte de Minas Gerais. **REVISTA ENGENHARIA NA AGRICULTURA-REVENG**, v. 19, n. 5, p. 437-449, 2011.

ALMEIDA, M. Seu dinheiro. **Exame**, 2020. Disponível em: <https://exame.com/seu-dinheiro/leilao-de-imoveis-do-santander-oferece-ate-70-de-desconto/>. Acesso em: 5 jun. 2020.

ANGELONI, M. T. Technologies de transmissión de linformation et consequences sur le processus des decisions des organizations. **Grenoble, France: Mémoire de DEA. École Superieure des Affaires**, 1992.

ANGELONI, Maria Terezinha. Elementos intervenientes na tomada de decisão. **Ciência da informação**, v. 32, n. 1, p. 17-22, 2003.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2:2019 Avaliação de Bens PARTE 2: Imóveis urbanos**. RIO DE JANIERO, p. 54. 2020.

AYU, Media Anugerah *et al.* A comparison study of classifier algorithms for mobile-phone's accelerometer based activity recognition. **Procedia Engineering**, v. 41, p. 224-229, 2012.

AZEVÊDO, Luana Lúcia Alves de. Métodos estatísticos em aprendizado de máquinas para problemas de classificação. 2018.

BATISTA, Gustavo Enrique de Almeida Prado *et al.* **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese de Doutorado. Universidade de São Paulo.

BATTITI, Roberto. First-and second-order methods for learning: between steepest descent and Newton's method. **Neural computation**, v. 4, n. 2, p. 141-166, 1992.

BERRY, Michael JA; LINOFF, Gordon S. **Data mining techniques: for marketing, sales, and customer relationship management**. John Wiley & Sons, 2004.

BERSTEIN, P. L. Desafio aos Deuses: A fascinante história do risco (Trad. I. Korylowsky) Rio de Janeiro. **RJ, Brasil: Campus**, 1997.

BIGUS, Joseph P. **Data mining with neural networks: solving business problems from application development to decision support**. McGraw-Hill, Inc., 1996.

BREIMAN, Leo. Random forests. **UC Berkeley TR567**, 1999.

CAJIAS, Marcelo; FREUDENREICH, Philipp. Exploring the determinants of liquidity with big data–market heterogeneity in German markets. **Journal of Property Investment & Finance**, 2018.

CARRILLO, Paul E.; WILLIAMS, Benjamin. The repeat time-on-the-market index. **Journal of Urban Economics**, v. 112, p. 33-49, 2019.

CASTRO, Paulo André Lima de. **Uma infra-estrutura para agentes arrematantes em múltiplos leilões simultâneos**. 2003. Tese de Doutorado. Universidade de São Paulo.

CHENG, Ping; LIN, Zhenguo; LIU, Yingchun. A model of time-on-market and real estate price under sequential search with recall. **Real Estate Economics**, v. 36, n. 4, p. 813-843, 2008.

CHENG, Ping; LIN, Zhenguo; LIU, Yingchun. Home price, time-on-market, and seller heterogeneity under changing market conditions. **The Journal of Real Estate Finance and Economics**, v. 41, n. 3, p. 272-293, 2010.

CHENG, Ping; LIN, Zhenguo; LIU, Yingchun. Liquidity risk of private assets: Evidence from real estate markets. **Financial Review**, v. 48, n. 4, p. 671-696, 2013.

CHENG, Ping; LIN, Zhenguo; LIU, Yingchun. Performance of thinly traded assets: A case in real estate. **Financial Review**, v. 48, n. 3, p. 511-536, 2013.

CLÉSIO, F. Custo Computacional. **Wordpress**, 2020. Disponível em: <https://mineracaodedados.wordpress.com/tag/custo-computacional/>. Acesso em: 26 set. 2020.

COX, L. What's wrong with risk matrices?. **Risk analysis**, v. 28, n. 2, p. 497-512, 2008.

DA SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. Elsevier Brasil, 2017.

DA SILVA, Wesley Vieira *et al.* Avaliação da escolha de um fornecedor sob condição de riscos a partir do método de árvore de decisão. **REGE Revista de Gestão**, v. 15, n. 3, p. 77-94, 2008.

DA SILVA, Wilamis Kleiton Nunes; DE MEDEIROS SANTOS, Araken. Estratégias de construções de comitês de classificadores multirrótulos no aprendizado semissupervisionado multidescrição. **Revista de Informática Teórica e Aplicada**, v. 24, n. 2, p. 71-100, 2017.

DASH, Ranjita Kumari. Selection of the best classifier from different datasets using WEKA. **International Journal of Engineering Research and Technology**, v. 2, n. 3, 2013.

DEL GIUDICE, Vincenzo *et al.* Real estate investment choices and decision support systems. **Sustainability**, v. 11, n. 11, p. 3110, 2019.

FAN, Jerome; UPADHYE, Suneel; WORSTER, Andrew. Understanding receiver operating characteristic (ROC) curves. **Canadian Journal of Emergency Medicine**, v. 8, n. 1, p. 19-20, 2006.

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

FLEISS, Joseph L.; LEVIN, Bruce; PAIK, Myunghee Cho. **Statistical methods for rates and proportions**. john wiley & sons, 2013.

FRANK, Eibe; MARK, A. Hall, and Ian H. **Witten, The WEKA Workbench, Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann,** 2016.

GAMA, João. Functional trees. **Machine learning**, v. 55, n. 3, p. 219-250, 2004.

GAN, Quan. Optimal selling mechanism, auction discounts and time on market. **Real Estate Economics**, v. 41, n. 2, p. 347-383, 2013.

GARDNER, Matt W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. **Atmospheric environment**, v. 32, n. 14-15, p. 2627-2636, 1998.

GENUER, Robin; POGGI, Jean-Michel; TULEAU-MALOT, Christine. Variable selection using random forests. **Pattern recognition letters**, v. 31, n. 14, p. 2225-2236, 2010.

GIANNOTTI, Claudio; GIBILARO, Lucia; MATTAROCCI, Gianluca. Liquidity risk exposure for specialised and unspecialised real estate banks. **Journal of Property Investment & Finance**, 2011.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. Editora Atlas SA, 2008.

GISLASON, Pall Oskar; BENEDIKTSSON, Jon Atli; SVEINSSON, Johannes R. Random forests for land cover classification. **Pattern Recognition Letters**, v. 27, n. 4, p. 294-300, 2006.

GODARD, Olivier *et al.* Treatise on New Risks. **Precaution, Crisis Management and Insurance, Paris: Editions Gallimard, Folio-Actuel, Inedit**, v. 100, p. 620, 2002.

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

GRÜBLER, M. Entendendo o funcionamento de uma Rede Neural Artificial. Medium, 2018. Disponível em: <https://medium.com/brasil-ai/entendendo-o-funcionamento-de-uma-rede-neural-artificial-4463fcf44dd0>. Acesso em: 11 nov. 2020.

HALLAK, Ricardo; PEREIRA FILHO, Augusto José. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. **Revista Brasileira de Meteorologia**, v. 26, n. 4, p. 591-608, 2011.

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

HAYKIN, S. **Redes Neurais - Princípios e Práticas**. Bookman. 2a edição. Porto Alegre, 2010.

HE, Chao; WRIGHT, Randall; ZHU, Yu. Habitação e liquidez. **Review of Economic Dynamics** , v. 18, n. 3, pág. 435-455, 2015.

HE, X. *et al.* Search Benefit in Housing Markets: An Inverted U-Shaped Price and TOM Relation. **Real Estate Economics**, p. 1-36, 2017.

HEBB, Donald Olding. **The organization of behavior: a neuropsychological theory**. J. Wiley; Chapman & Hall, 1949.

Hengshu *et al.* Days on market: Measuring liquidity in real estate markets. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. 2016. p. 393-402.

HILLSON, D. Extending the risk process to manage opportunities. **International Journal of Project Management**, v.20, p. 235-240. 2002.

HILLSON, D. **Managing risk in projects**. Gower Publishing, Ltd., 2009.

JERÓNIMO, Helena Mateus. A peritagem científica perante o risco e as incertezas. **Análise Social**, n. 181, p. 1143-1165, 2006.

KERZNER, Harold. **Gerenciamento de Projetos: uma abordagem sistêmica para planejamento, programação e controle**. Editora Blucher, 2011.

KEYNES, John Maynard. **A treatise on probability**. Macmillan and Company, limited, 1921.

KLEMPERER, Paul. Auction theory: A guide to the literature. **Journal of economic surveys**, v. 13, n. 3, p. 227-286, 1999.

KOHAVI, Ron *et al.* Um estudo de validação cruzada e bootstrap para estimativa de precisão e seleção de modelo. In: **Ijcai** . 1995. p. 1137-1145.

KOK, Shiau Hui; ISMAIL, Normaz Wana; LEE, Chin. The sources of house price changes in Malaysia. **International Journal of Housing Markets and Analysis**, 2018.

KOVÁCS, Z. L.; **Redes Neurais Artificiais: Fundamentos e Aplicações**. Livraria da Física, 4ª edição, São Paulo - SP. 2006.

LAKATOS, Eva Maria; MARCONI, M. de A. Fundamentos de metodologia científica. 5. reimp. **São Paulo: Atlas**, v. 310, 2007.

LIU, Xiaohui; CHENG, Gongxian; WU, John Xingwang . Analyzing outliers cautiously. **IEEE Transactions on Knowledge and Data Engineering**, v. 14, n. 2, p. 432-437, 2002.

LOPES, Sérgio Miguel Neves. **Plataforma para gestão processual de processos de insolvência e leilões**. 2016. Tese de Doutorado.

LORENA, Ana Carolina; DE CARVALHO, André CPLF. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43-67, 2007.

MACHADO, Emerson Lopes. Redução de custo computacional em classificações baseadas em transformadas aprendidas. 2015.

MANSUR, M.; RADAM, A.; YAAKUB, Y. Struktur Pasaran, Gelagat Firma dan Produktiviti: Kajian Empirik Firma Insurans Am di Malaysia. **Jurnal Ekonomi Malaysia**, v. 51, n. 1, p. 133-144, 2017.

MARKO, R. Preço dos imóveis. **Sindusconsp**, 2020. Disponível em: <https://sindusconsp.com.br/precos-dos-imoveis-residenciais-seguem-em-alta-em-maio/>. Acesso em: 22 set. 2020.

MARKOWSKI, Adam S.; MANNAN, M. Sam. Fuzzy risk matrix. **Journal of hazardous materials**, v. 159, n. 1, p. 152-157, 2008.

MATTHEWS, Brian W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. **Biochimica et Biophysica Acta (BBA)-Protein Structure**, v. 405, n. 2, p. 442-451, 1975.

MCGREAL, S. *et al.* Measuring the influence of space and time effects on time on the market. **Urban Studies**, v. 53, n. 13, p. 2867-2884, 2016.

MITCHELL, Tom M. *et al.* **Machine learning**. 1997. Burr Ridge, IL: McGraw Hill, v. 45, n. 37, p. 870-877, 1997.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

NAPOLITANO, Domingos Márcio Rodrigues *et al.* **A tomada de decisão em projetos: um estudo exploratório sobre o processo de identificação de riscos**. 2014.

NAPOLITANO, Domingos Marcio Rodrigues; SASSI, Renato José. Modelo de sistema de inferência Fuzzy baseado em matrizes de probabilidade e impacto para classificar riscos em projetos. **Navus: Revista de Gestão e Tecnologia**, v. 8, n. 4, p. 69-89, 2018.

OLIVEIRA, Marcelo Costa; AZEVEDO-MARQUES, Paulo Mazzoncini de; CIRNE FILHO, Walfredo da Costa. Grades computacionais na recuperação de imagens médicas baseada em conteúdo. **Radiol Bras**, São Paulo , v. 40, n. 4, p. 255-261, Aug. 2007.

PMI. **Um Guia do Conhecimento no Gerenciamento de Projetos - Guia PMBoK.** Sexta Edição em Português. Newton Square, PA, USA. Project Management Institute, Inc. 2017.

POWERS, David Martin. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2011.

PRATI, R. C.; BATISTA, GEAPA; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215-222, 2008.

PRODANOV, Cleber Cristiano; DE FREITAS, Ernani Cesar. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição**. Editora Feevale, 2013.

QUINLAN, J. Ross; CAMERON-JONES, R. Mike. FOIL: A midterm report. In: **European conference on machine learning**. Springer, Berlin, Heidelberg, 1993. p. 1-20.

RAMSEY, Frank Plumpton. **Os fundamentos da matemática e outros ensaios lógicos**. K. Paul, Trench, Trubner & Company, Limited, 1931.

RICH, Elaine; KNIGHT, Kevin-Inteligência. Artificial-2 a edição. 1993.

ROSENBLATT, M. The Perceptron: A probabilistic model for information storage and organization in the Brain. **Psychological review**, v.65, n.6, p. 386-408, 1958.

SANDHOLM, Tuomas; HUAI, Qianbo. Nomad: mobile agent system for an Internet-based auction house. **IEEE internet Computing**, v. 4, n. 2, p. 80-86, 2000.

SALVETTI, N. **Proágil-29110: Processo ágil aderente à norma ISO/IEC 29110 baseado em Scrum e princípios Lean.** Programa de Pós Graduação e Gestão do Conhecimento. Universidade Nove de Julho, 2019.

SASSI, Renato José. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados: teoria dos rough sets e redes neurais artificiais mapas auto-organizáveis**. 2006. Tese de Doutorado. Universidade de São Paulo.

SINGH, Archana; SHARMA, Apoorva; DUBEY, Gaurav. Big data analytics predicting real estate prices. **International Journal of System Assurance Engineering and Management**, p. 1-12, 2020.

SOUZA, FJ de. **Modelos Neuro-Fuzzy Hierárquicos**. 1999. Tese de Doutorado. Tese de Doutoramento, Departamento de Engenharia Elétrica da PUC, Rio de Janeiro.

TAJANI, Francesco; MORANO, Pierluigi; NTALIANIS, Klimis. Automated valuation models for real estate portfolios. **Journal of Property Investment & Finance**, 2018.

TANG, Fangqin; REN, Aizhu. Agent-based evacuation model incorporating fire scene and building geometry. **Tsinghua Science and Technology**, v. 13, n. 5, p. 708-714, 2008.

TURNBULL, Geoffrey K.; ZAHIROVIC-HERBERT, Velma. The transitory and legacy effects of the rental externality on house price and liquidity. **The Journal of Real Estate Finance and Economics**, v. 44, n. 3, p. 275-297, 2012.

TURNBULL, Geoffrey K.; ZAHIROVIC-HERBERT, Velma. Why do vacant houses sell for less: holding costs, bargaining power or stigma?. Real Estate Economics, v. 39, n. 1, p. 19-43, 2011.

VAN DER AALST, Wil MP *et al.* Process mining: a two-step approach to balance between underfitting and overfitting. **Software & Systems Modeling**, v. 9, n. 1, p. 87, 2010.

WIDEMAN, R. Max. **Project and program risk management: a guide to managing project risks and opportunities**. 1992. Tese de Doutorado. Univerza v Mariboru, Ekonomsko-poslovna fakulteta.

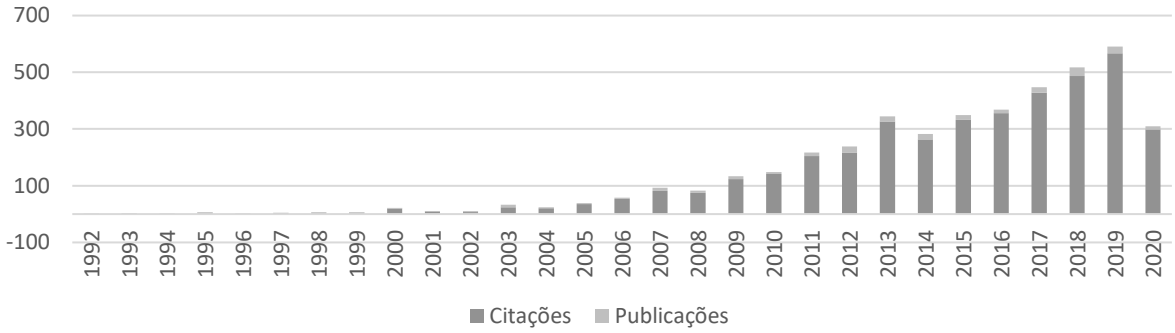
APÊNDICE A - BIBLIOMETRIA (LIQUIDAÇÃO DE IMÓVEIS)

Em princípio, a temática imobiliária com foco na liquidez constitui uma linha de pesquisa que vem evoluindo ao longo do tempo. Portanto, o objetivo desse anexo é realizar uma análise da produção científica por meio de técnicas e ferramentas bibliométricas como *VosViwer* e *Bibliometrix,* respondendo a 3 leis bibliométricas essenciais em apoio a essa dissertação (Lei de Bradford, Lotka e Zipf).

Em uma perspectiva das métricas acadêmicas o tema mostra-se em constante evolução, como mostram os resultados de uma pesquisa no site Web of Science (WoS) realizada dia 30 de junho de 2020.

Como resultado desta revisão, foram encontradas 263 publicações das quais são feitas 4.092 citações, sendo 3.675 sem autocitações cuja evolução no tempo é apresentada no Gráfico 3 empregando as palavras-chave “Real estate” e “Liquidity”.

Gráfico 3 - Evolução das publicações sobre liquidez imobiliária na base WoS



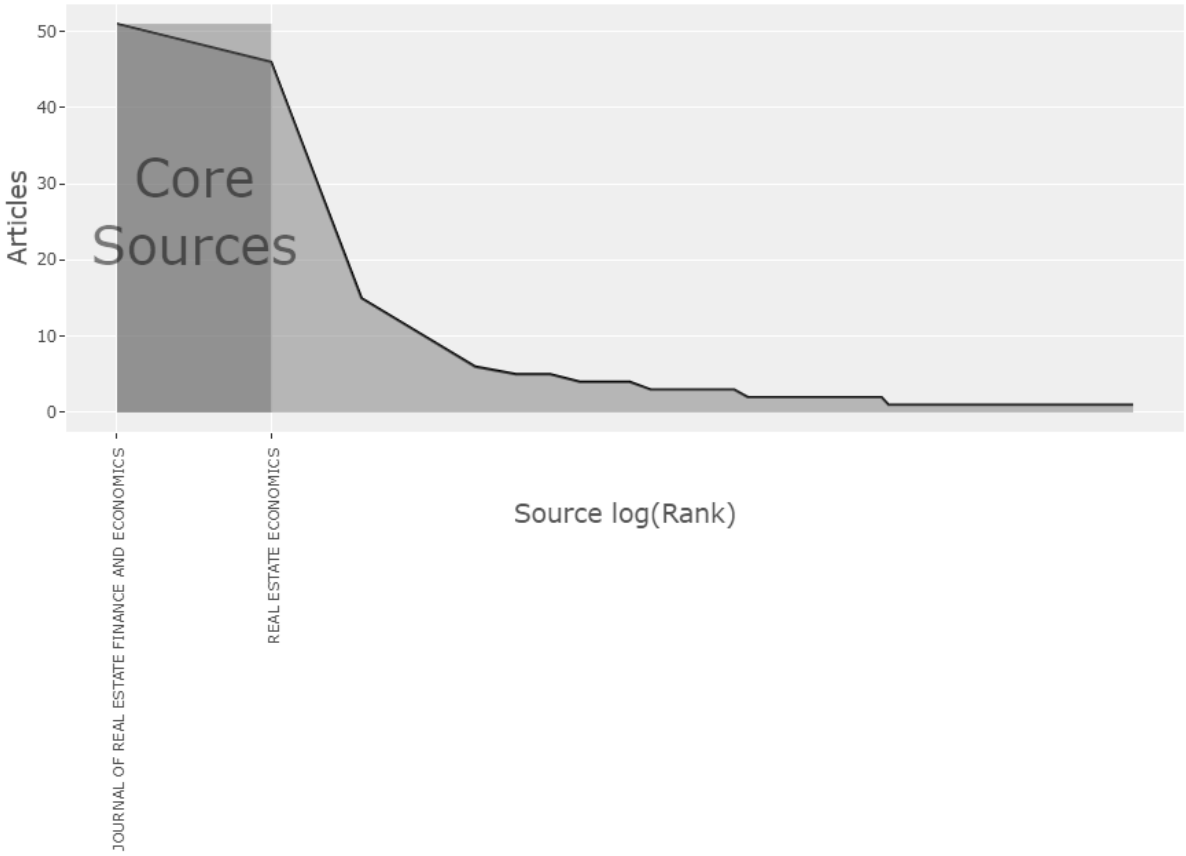
Fonte: Autor.

Um evento de destaque nesta área, aparentemente foi a publicação do artigo de CHENG *et al.* (2008), que descrevem em seu modelo uma fórmula para descobrir a relação teórica entre preços de imóveis e seu tempo exposto no mercado (*Time On Market* - TOM), utilizando métricas convencionais para retorno imobiliário de risco, descrito posteriormente por Cajias e Freudenreich (2018) e por Carrillo e Williams (2019), demonstrando que esse tema pode ser tratado em uma abordagem científica no sentido de aprimorar os resultados das decisões realizadas por meio deste instrumento. Mesmo conceito abordado por Hengshu *et al.* (2016), no entanto, denominado Days on Market (DOM), o qual se refere ao mesmo indicador, porém, com uma abordagem com foco na avaliação da popularidade de um imóvel.

Uma das principais leis da bibliometria, a lei de Bradford, foi aplicada para esta análise. A lei de Bradford é um padrão descrito pela primeira vez por Samuel C. Bradford em 1934 que estima os retornos exponencialmente decrescentes da busca por referências em revistas científicas verificadas na pesquisa de (SEMBAY *et al.*, 2019).

Isso significa que em uma determinada área de pesquisa, um pequeno número de publicações que são responsáveis por uma parte considerável do total de publicações como mostra a Figura 32 abaixo.

Figura 32 - Lei de Bradford



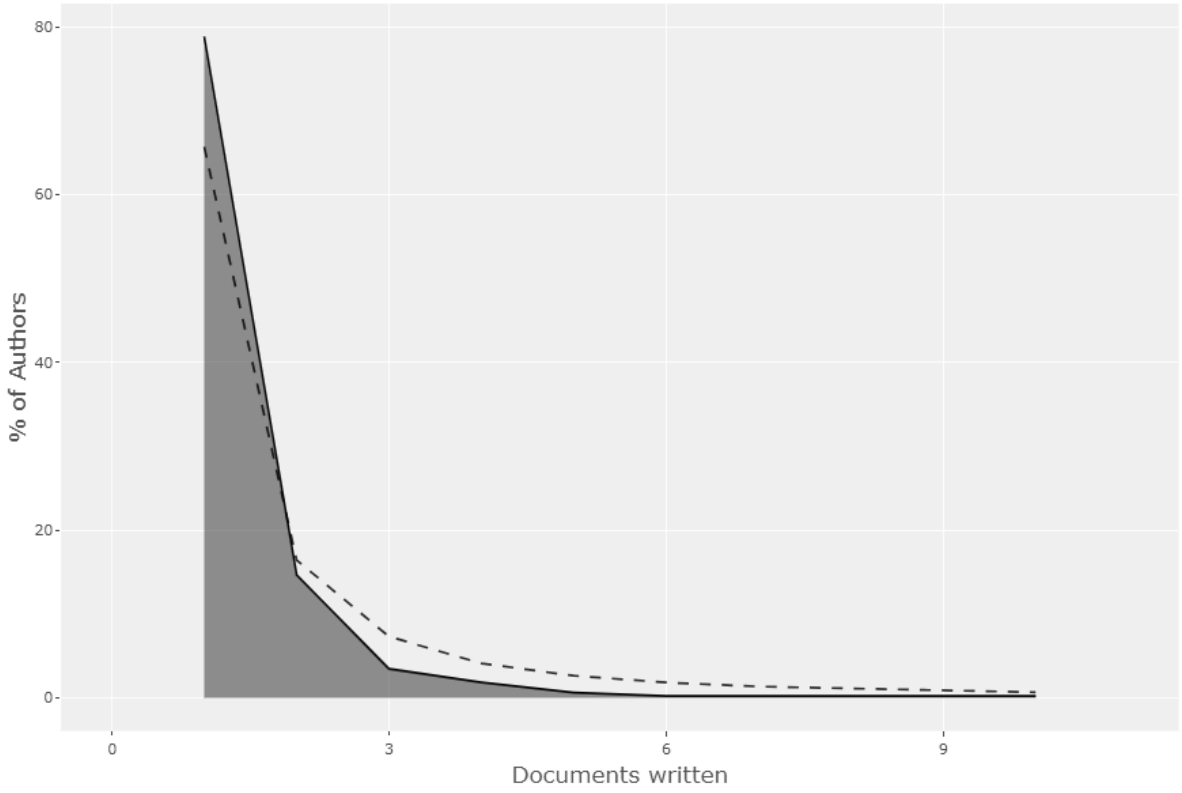
Fonte: Autor.

A lei de Bradford aplicada nesse levantamento apontou os periódicos que mais publicam na área de pesquisa em que essa dissertação discorre. Em outras palavras, os periódicos ou fontes principais, concentram os *ranks* 1 e 2 com um total de frequência de publicação de 97, enquanto 18 outros periódicos possuem 79 publicações de frequência.

Outra lei fundamental da bibliometria é a lei de Lotka. Alfred J. Lotka (1926) estudou os padrões de produtividade autor. Lotka observou que, em uma determinada área da ciência, há uma série de autores que publicam apenas um estudo, enquanto um pequeno grupo de autores prolíficos contribuem com um grande número de publicações.

Este padrão de produtividade autor não parece depender da ciência em que é aplicada a lei de Lotka. A única condição que tem de ser assumido é o período de tempo, como pode-se observar na Figura 33 abaixo.

Figura 33 - Lei de Lotka - Frequência de distribuição de produção científica



Fonte: Autor.

Por fim, a terceira lei da bibliometria conhecida na literatura, é a lei de Zipf, que é usada para previsão de a frequência de palavras em texto.

A Lei de Zipf, formulada na década de 1940 por George Kingsley Zipf, linguista da Universidade de Harvard, rege a dimensão, importância ou frequência dos elementos de uma lista ordenada.

Com base no levantamento realizado na WoS, a Tabela 5 abaixo, mostra em uma lista ordenada, as palavras-chave mais empregadas na literatura revisada para esta dissertação.

Tabela 5 - Lei de Zipf - Frequência de palavras na WoS

Palavras-chave	Ocorrências
liquidity	89
real-estate	44
performance	42
market	37
model	36
risk	34
time	27
returns	26
impact	22
prices	22

Fonte: Autor.

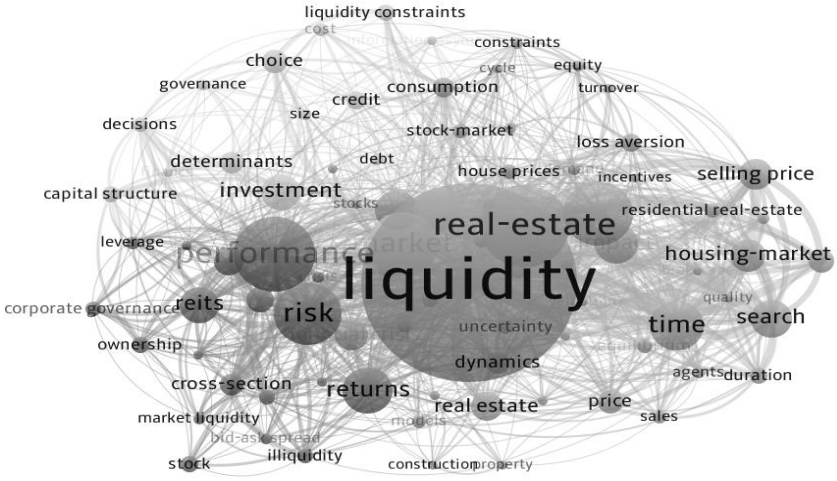
Com base no levantamento realizado na WoS, a Figura 34 abaixo, mostra em uma nuvem de palavras empregadas nos artigos presentes na literatura revisada.

Figura 34 - Nuvem de palavras

Fonte: Autor.

Os relacionamentos entre as palavras-chave levantados na WoS são exibidos na forma de uma rede na Figura 35, abaixo.

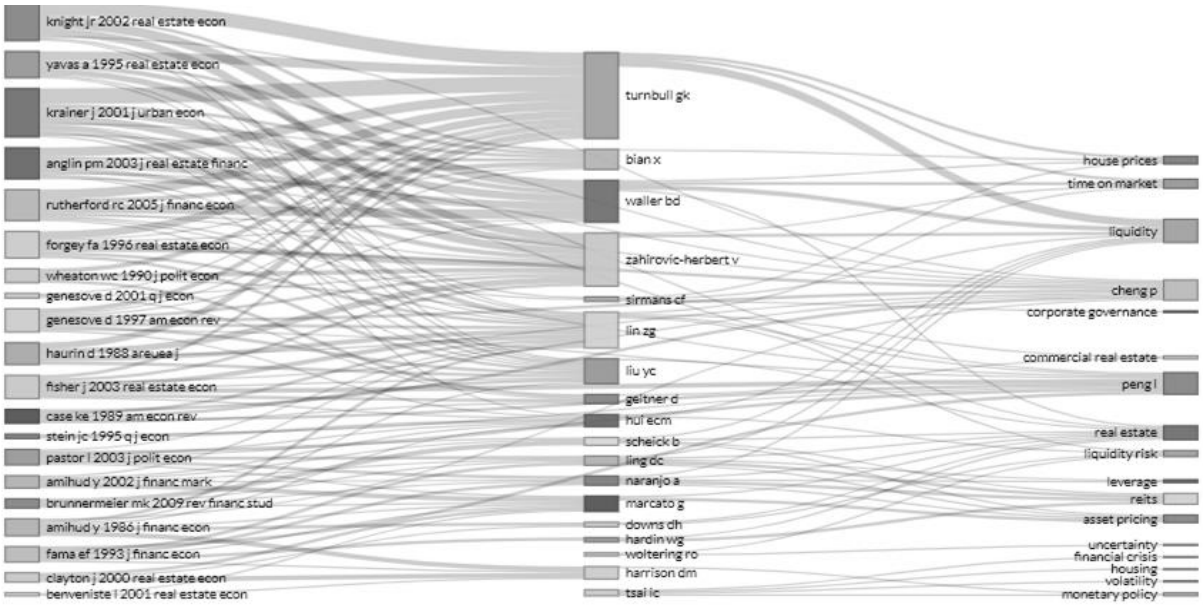
Figura 35 - Frequência de rede de palavras-chave no *software* VosViwer



Fonte: Autor.

Por fim, face aos resultados apresentados, a revisão da literatura realizada revelou que inicialmente as pesquisas eram mais genéricas e que ao longo do tempo novos autores foram aparecendo. Portanto, para entender a evolução destes temas em uma dinâmica temporal, apresenta-se a Figura 36 a evolução temática da dos artigos retornados da pesquisa no WoS analisada.

Figura 36 - Evolução temática dos artigos pesquisados

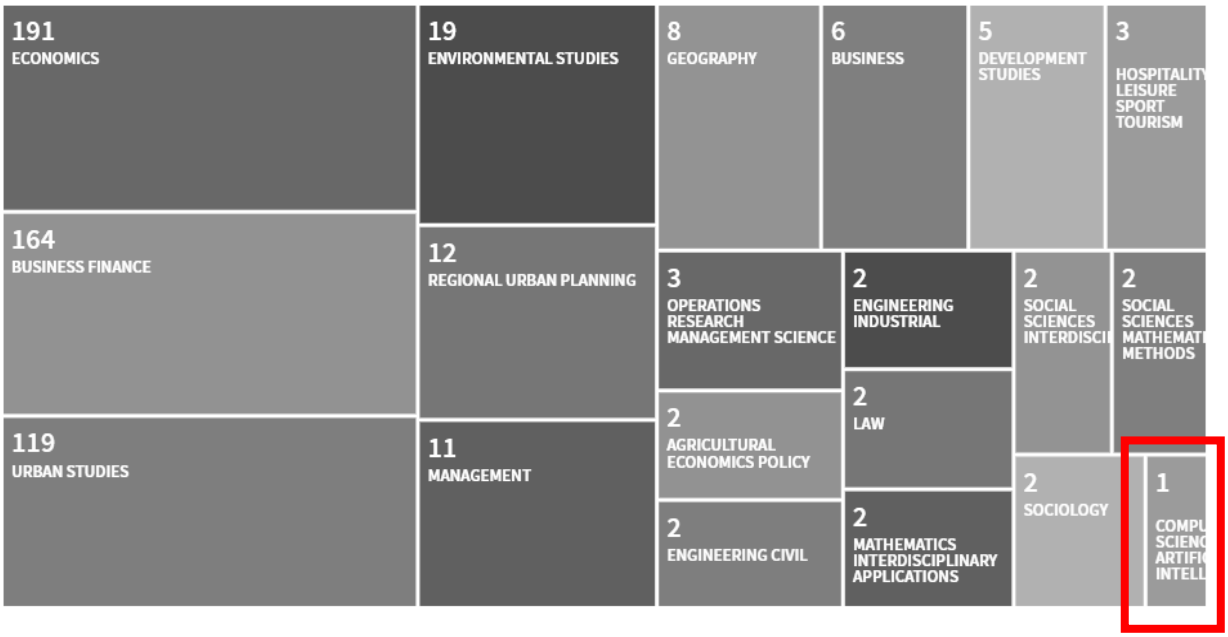


Fonte: Autor.

Dessas análises, pode-se concluir que o tema vem evoluindo ao longo do tempo, configurando um tema atual com 263 publicações das quais são feitas 4.092 citações, sendo 3.675 sem autocitações, empregando as palavras-chave “Real estate” e “Liquidity”.

Por fim, consta base também foram avaliadas as áreas do conhecimento nas quais estas pesquisas são conduzidas, o que mostrou que o tema possui grande impacto sendo muito empregados na Economia, finanças de negócios, Gestão, negócios e pouco empregado ainda na computação. A distribuição resultante é apresentada na Figura 37:

Figura 37 - Áreas de conhecimento



Fonte: Elaborado pelo autor com base em dados levantados no WoS.

APÊNDICE B - CÓDIGOS EMPREGADOS NO EXPERIMENTO

Arquitetura híbrida para classificação de liquidez imobiliária urbana em Leilões por meio de técnicas inteligentes

Objetivo
Qualificação de mestrado e futura defesa

Finalidade
Acadêmica e prática

Escopo
Modelo híbrido preditivo

Linguem de programação empregada para a construção do modelo
Linguagem R – Versão 4.0

Aplicação das técnicas inteligente
Regressão e Classificação

Técnica inteligente selecionada
Randon Forest

Autor
Marcelo Ferreira Albano
marceloalbano@uni9.edu.br

Lattes
http://lattes.cnpq.br/3392758888478198

Github
https://github.com/mfalbano3/UNINOVE

Processo de discretização dos dados

#Legendas
>Códigos

Classe: (1 Residencial - 2 Comercial - 3 Industrial)
Município - UF: Conforme IBGE
Tipo: (1 Apartamento - 2 casa - 3 Galpão - 4 Gleba - 5 Loja - 6 Lote - 7 Sala comercial - 8 Terreno)
Situação: (1 Ocupado - 2 Livre)
Demais: (1 Sim - 2 Não)

##Observações gerais

Após as bibliotecas empregadas, será mostrado o procedimento de construção do modelo em 5 etapas, pois na predição de dados novos, algumas informações encontram-se indisponíveis como o % abaixo do mercado, o tempo de exposição, a possibilidade de lances, número de lances, o arremate e a probabilidade.

Bibliotecas empregadas

```
>library(FactoMineR)
>library(factoextra)
>library(cluster)
>library(csv)
>library(dplyr)
>library Factoshiny)
>library(xlsx)
>library(arules)
>library(caTools)
>library(rpart)
>library(rpart.plot)
>library(e1071)
>library(ROCR)
>library(pROC)
>library(PRROC)
>library(caret)
>library(corrplot)
>library(Hmisc)
>library(PerformanceAnalytics)
>library(randomForest)
```

5 - Etapas do processo na Arquitetura híbrida

##Etapa 1

Por **regressão** linear múltipla, tendo 18 atributos de entrada, necessita-se encontrar:
Qual o percentual abaixo do mercado os imóveis são arrematados com base em seus atributos?

#Ler Base de dados - 1.600 instâncias com 22 variáveis

```
>Base <- read.xlsx("C:/Users/usuário/pasta/nome do arq.xlsx",
                   sheetIndex = 1,
                   as.data.frame =T,
                   header =T)
```

#Ler dados novos - 20 instâncias - 18 atributos variáveis

```
>df18 <- read.xlsx ("C:/Users/usuário/pasta/nome do arq.xlsx",
                   sheetIndex = 1,
                   as.data.frame =T,
                   header =T)
```

##Etapa 1 continua...

```
##Etapa 1

#Dividindo a Base em treino e teste para a
#função alvo "abaixo"

>divisao = sample.split(Base$abaixo, SplitRatio = 0.666666667)

>base_treinamento = subset(Base,divisao==TRUE)
>base_teste = subset(Base,divisao==FALSE)

>base_treinamento$abaixo <- as.numeric(base_treinamento
$abaixo)

>classificador = randomForest(x = base_treinamento[-1], y =
base_treinamento$abaixo)

>print(classificador)

#Out-of-bag
#Método de medir o erro de previsão de florestas aleatórias,
#árvores de decisão aprimoradas e outros modelos de
#aprendizado de máquina que utilizam o bootstrap agregado a
#amostras de dados de subamostras usadas para treinamento.

>set.seed(221)
>rf <- randomForest(abaixo~.,base_treinamento)
>print(rf)
>plot(rf)

#Módulo de tunnig do RF

>tuneRF(base_treinamento[,-19],base_treinamento[,19],
       stepFActor = 0.5, plot = TRUE,
       ntreeTry = 900,trace = TRUE,
       improve = 0.05)
>print(rf)

#Número de nós

>hist(treesize(rf),main = "Número de nós por árvore",
     col = "purple")

#Importância das variáveis

varImpPlot(rf)

varImpPlot(rf, sort = T, n.var = 10, main = "Top")
importance(rf)
```

##Etapa 1 continua...

```
##Etapa 1

#Features - Seleção dos 18 atributos a serem considerados

>features <- c("classe", "municipio", "uf", "tipo","situacao",
              "area", "vpp", "vup", "vm",
              "selic", "igpm", "ipca", "incc", "pib", "idade",
              "condominio", "esquina", "vagas")

#Aplicação do modelo preditivo

>rf.model <- randomForest(x = base_treinamento[,features],
                          y = base_treinamento$abaixo,
                          importance = TRUE,
                          ntree = 900,
                          trace = TRUE,
                          improve = 0.05)

#Predição múltipla instância

>predictions <- data.frame(df18)
>predictions$abaixo <- predict(rf.model, df18)

                                                  ##Encerramento da Etapa 1.
```

```
##Etapa 2)

Por regressão linear múltipla, agora com 19 atributos de
entrada, necessita-se encontrar:
Qual o DOM ou tempo de exposição desses imóveis no mercado de
leilões?

>df19 <- predictions

#Dividindo a Base em treino e teste para a função
#alvo "exposicao"

>divisao = sample.split(Base$exposicao, SplitRatio = 0.666667)

>base_treinamento = subset(Base,divisao==TRUE)
>base_teste = subset(Base,divisao==FALSE)

>classificador = randomForest(x = base_treinamento[-1],
      y = base_treinamento$exposicao)

>print(classificador)

                                                  ##Etapa 2 continua...
```

```
##Etapa 2)

#Out-of-bag
#Método de medir o erro de previsão de florestas aleatórias,
#árvores de decisão aprimoradas e outros modelos de
#aprendizado de máquina que utilizam o bootstrap agregado a
#amostras de dados de subamostras usadas para treinamento.

>set.seed(222)
>rf <- randomForest(exposicao~.,base_treinamento)
>print(rf)
>plot(rf)

#Módulo de tunnig do RF

>tuneRF(base_treinamento[,-20],base_treinamento[,20],
       stepFActor = 0.5, plot = TRUE,
       ntreeTry = 900,trace = TRUE,
       improve = 0.05)
>print(rf)

#Número de nós

>hist(treesize(rf),main = "Número de nós por árvore",
     col = "purple")

#Importância das variáveis

varImpPlot(rf)

varImpPlot(rf, sort = T, n.var = 10, main = "Top")
importance(rf)

#Features - Seleção dos 19 atributos a serem considerados

>features <- c("classe", "municipio", "uf", "tipo","situacao",
              "area", "vpp", "vup", "vm", "abaixo",
              "selic", "igpm", "ipca", "incc", "pib", "idade",
              "condominio", "esquina", "vagas")

#Aplicação do modelo preditivo

>rf.model <- randomForest(x = base_treinamento[, features],
                          y = base_treinamento$exposicao,
                          importance = TRUE,
                          ntree = 900,
                          trace = TRUE,
                          improve = 0.05)
```

##Etapa 2 continua...

```
##Etapa 2)

#Predição múltipla instância

>predictions <- data.frame(df19)
>predictions$exposicao <- predict(rf.model, df19)

                                                ##Encerramento da Etapa 2.
```

```
##Etapa 3)

Por classificação, tendo 20 atributos de entrada, necessita-se
encontrar:
Qual a probabilidade desses imóveis terem ou não lances no
mercado de leilões?

>df20 <- predictions

#Aqui temos que fazer o modelo discretizar da coluna lances em
#uma nova base, que chamaremos de Base2

>Base2 <- Base

>Base2$lances[Base2$lances>0] <- "sim"
>Base2$lances[Base2$lances==0] <- "nao"

>str(Base)
>str(Base2)

>Base2$arrematado[Base2$arrematado=="sim"] <- "1"
>Base2$arrematado[Base2$arrematado=="nao"] <- "2"

>str(Base)
>str(Base2)

#Dividindo a Base em treino e teste para função alvo "lances"

>divisao = sample.split(Base2$lances, SplitRatio = 0.66666667)

>base_treinamento = subset(Base2,divisao==TRUE)
>base_teste = subset(Base2,divisao==FALSE)

>base_treinamento$lances <- as.factor(base_treinamento$lances)
>classificador = randomForest(x = base_treinamento[-1],
      y = base_treinamento$lances)



                                                ##Etapa 3 continua...
```

```
##Etapa 3)

>print(classificador)

#Out-of-bag
#Método de medir o erro de previsão de florestas aleatórias,
#árvores de decisão aprimoradas e outros modelos de
#aprendizado de máquina que utilizam o bootstrap agregado a
#amostras de dados de subamostras usadas para treinamento.

>set.seed(223)
>rf <- randomForest(lances~.,base_treinamento)
>print(rf)
>plot(rf)

#Módulo de tunnig do RF

>tuneRF(base_treinamento[,-21],base_treinamento[,21],
       stepFActor = 0.5,
       plot = TRUE,
       ntreeTry = 900,
       trace = TRUE,
       improve = 0.05)
>print(rf)

#Número de nós

>hist(treesize(rf),main = "Número de nós por árvore",
     col = "purple")

#Importância das variáveis

>varImpPlot(rf)

>varImpPlot(rf, sort = T, n.var = 10, main = "Top")
>importance(rf)

#Features - Seleção dos 19 atributos a serem considerados

>features <- c("classe","municipio", "uf", "tipo", "situacao",
            "area", "vpp", "vup", "vm", "exposicao", "abaixo",
              "selic", "igpm", "ipca", "incc", "pib", "idade",
              "condominio", "esquina", "vagas")

#Aplicação do modelo preditivo

>rf.model <- randomForest(x = base_treinamento[, features],
      y = base_treinamento$lances, importance = TRUE,
      ntree = 900)
```

##Etapa 3 continua...

```
##Etapa 3)

#Predição múltipla instância

>redictions <- data.frame(df20)
>predictions$lancesS <- predict(rf.model, df20)
>predictions$Probalances <- predict(rf.model, df20,
      type = "prob")[,2]

                                        ##Encerramento da Etapa 3.
```

```
##Etapa 4)

Por regressão linear múltipla, agora com 22 atributos de
entrada, necessita-se encontrar:
Quantos lances esses imóveis podem ter no mercado de leilões?

> df22 <- predictions

#Dividindo a Base em treino e teste para a função
#alvo "exposicao"

>divisao = sample.split(Base$lances, SplitRatio = 0.666666667)

>base_treinamento = subset(Base,divisao==TRUE)
>base_teste = subset(Base,divisao==FALSE)

>base_treinamento$lances <- as.numeric(base_treinamento
      $lances)

>classificador = randomForest(x = base_treinamento[-1],
      y = base_treinamento$lances)

print(classificador)

#Out-of-bag
#Método de medir o erro de previsão de florestas aleatórias,
#árvores de decisão aprimoradas e outros modelos de
#aprendizado de máquina que utilizam o bootstrap agregado a
#amostras de dados de subamostras usadas para treinamento.

>set.seed(224)

                                        ##Etapa 4 continua...
```

```
##Etapa 4)

>rf <- randomForest(lances~.,base_treinamento)
>print(rf)

>plot(rf)

#Módulo de tunnig do RF

>tuneRF(base_treinamento[,-21],base_treinamento[,21],
       stepFActor = 0.5,
       plot = TRUE,
       ntreeTry = 900,
       trace = TRUE,
       improve = 0.05)
>print(rf)

#Número de nós

>hist(treesize(rf),main = "Número de nós por árvore",
     col = "purple")

#Importância das variáveis

varImpPlot(rf)

varImpPlot(rf, sort = T, n.var = 10, main = "Top")
importance(rf)

#Features - Seleção dos 22 atributos a serem considerados

>names(df22)
>features <- c("classe", "municipio", "uf", "tipo","situacao",
              "area", "vpp","vup", "vm", "exposicao","abaixo",
              "selic", "igpm", "ipca", "incc", "pib", "idade",
              "condominio", "esquina", "vagas")

#Aplicação do modelo preditivo

>rf.model <- randomForest(x = base_treinamento[, features],
                          y = base_treinamento$lances,
                          importance = TRUE,
                          ntree = 900)

#Predição múltipla instância

>predictions <- data.frame(df22)
>predictions$lances <- predict(rf.model, df22)

                                              ##Encerramento da Etapa 4.
```

```
##Etapa 5)

Por classificação, tendo 23 atributos de entrada, necessita-se
encontrar:
Qual a probabilidade desses imóveis serem ou não arrematados
No mercado de leilões?

>df23 <- predictions

#Dividindo a Base em treino e teste para a função
#alvo "exposicao"

>Base$arrematado <- as.factor(Base$arrematado)

>divisao = sample.split(Base$arrematado, SplitRatio = 0.66667)

>base_treinamento = subset(Base,divisao==TRUE)
>base_teste = subset(Base,divisao==FALSE)

>classificador = randomForest(x = base_treinamento[-1],
      y = base_treinamento$arrematado)

>print(classificador)

#Out-of-bag
#Método de medir o erro de previsão de florestas aleatórias,
#árvores de decisão aprimoradas e outros modelos de
#aprendizado de máquina que utilizam o bootstrap agregado a
#amostras de dados de subamostras usadas para treinamento.

>set.seed(225)
>rf <- randomForest(arrematado~.,base_treinamento)
>print(rf)

>plot(rf)

#Módulo de tunnig do RF

>tuneRF(base_treinamento[,-22],base_treinamento[,22],
       stepFActor = 0.5,
       plot = TRUE,
       ntreeTry = 900,
       trace = TRUE,
       improve = 0.05)
>print(rf)
```

##Etapa 5 continua...

```
##Etapa 5)

#Número de nós

>hist(treesize(rf),main = "Número de nós por árvore",
     col = "purple")

#Importância das variáveis

varImpPlot(rf)

varImpPlot(rf, sort = T, n.var = 10, main = "Top")
importance(rf)

#Features - Seleção dos 21 atributos a serem considerados

>names(df23)

>features <- c("classe","municipio", "uf", "tipo", "situacao",
               "area", "vpp","vup", "vm", "lances","exposicao",
               "selic", "igpm", "ipca", "incc", "pib", "idade",
               "condominio", "esquina", "vagas", "abaixo")

#Aplicação do modelo preditivo

>rf.model <- randomForest(x = base_treinamento[, features],
                          y = base_treinamento$arrematado,
                          importance = TRUE,
                          ntree = 1000)

#Predição múltipla instância

>predictions <- data.frame(df23)
>predictions$arrematado <- predict(rf.model, df23)
>predictions$arremateProba <- predict(rf.model,
      df23, type = "prob")[,2]

#Salvar em Excel.

>write.xlsx(predictions, " C:/Users/usuário/pasta/
      nome do arq.xlsx")


Por classificação, tendo agora 25 atributos de saída.
```

```
#Gerar árvore para as regras de lance

#Criando um classificador

dt <- predictions

>dt$Base.arrematado <- factor(dt$Base.arrematado,
levels = c('sim','nao'))


>fit <- rpart(formula = dt$Base.arrematado ~ lancesB.A,
             method = 'class',
             data = dt,
             parms = list(split = "gini"),
             cp = 0.000000002,
             control = rpart.control(
               minsplit = 1,
               minbucket = 1,
               maxdepth = 50))
>barplot(fit$variable.importance)
>fit$cptable
>printcp(fit)


>rpart.plot(fit,
           type = 0,
           extra = 101,
           box.palette = "GnBu",
           branch.lty=3,
           shadow.col = "blue",
           nn=T,
           cex = 1)
>print(fit)


                        ##Encerramento da Etapa 5 e do processo preditivo.
```