

THAYS CARMONHA DAMIN

**APLICAÇÕES DE *MACHINE LEARNING* PARA VALORAÇÃO DE
EMPREENDIMENTOS RESIDENCIAIS DE BASE IMOBILIÁRIA NA CIDADE DE
SÃO PAULO**

São Paulo

2024

THAYS CARMONHA DAMIN

**APLICAÇÕES DE *MACHINE LEARNING* PARA VALORAÇÃO DE
EMPREENDIMENTOS RESIDENCIAIS DE BASE IMOBILIÁRIA NA CIDADE DE
SÃO PAULO**

Monografia apresentada à Escola
Politécnica da Universidade de São Paulo, para
obtenção do título de Especialista em *Real Estate*
– Economia Setorial e Mercados MBA-USP

Orientador:

Prof. Dr. Abla Akkari Osso

São Paulo

2024

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo-na-publicação

Damin, Thays

APLICAÇÕES DE MACHINE LEARNING PARA VALORAÇÃO DE EMPREENDIMENTOS RESIDENCIAIS DE BASE IMOBILIÁRIA NA CIDADE DE SÃO PAULO / T. Damin -- São Paulo, 2024.

118 p.

Monografia (MBA em Real Estate – Economia Setorial e Mercados) Escola Politécnica da Universidade de São Paulo. Poli-Integra.

1.Machine Learning 2.Inteligência artificial 3.Valoração 4.EBI 5.Residencial I.Universidade de São Paulo. Escola Politécnica. Poli-Integra II.t.

DEDICATÓRIA

“Não deseje que seja mais fácil, deseje que você seja melhor”

John Earl Shoaff

Dedico este trabalho à minha amada família - Raquel, Alexandre, Arthur e Cindy. O amor e apoio inabaláveis que vocês me proporcionaram foram o combustível para minha persistência. Ao meu namorado, Jacques Massie, agradeço por estar sempre presente, mesmo quando fisicamente distante. E a Prof^a Dr. Abla Akkari, cuja orientação foi decisiva para a conclusão desta importante etapa acadêmica que enche meu coração de orgulho.

AGRADECIMENTOS

Agradeço imensamente à Prof^ª Dr. Abla Akkari por sua orientação vital e feedbacks construtivos, tornando este trabalho relevante e envolvente. Também sou grata à Prof^ª Dr. Paola Torneri e ao Prof^º Dr. Fabiano Correa por suas valiosas contribuições e experiência, sugerindo melhorias cruciais que enriqueceram este material.

Agradeço aos professores, colegas e amigos que facilitaram meu acesso aos profissionais incríveis entrevistados: Prof^º Dr. Claudio Tavares, Prof^ª Dr. Eliane Monetti, Prof^ª Dr. Paola Torneri, Prof^ª Dr. Carolina Gregório, Antônio Neto, Marcio Augusto Carvalho, Leonardo Morgatto, André Lucarelli e Bruno Ainsa; e também aqueles que me auxiliaram com sua valiosa opinião profissional sobre o tema abordado neste trabalho: Ricardo Vallilo, Thiago Robles e Paulo Gallo.

A todos os entrevistados, que por razões de confidencialidade não posso nomear, expresso minha profunda gratidão por disponibilizarem seu tempo para esta pesquisa. Sua participação foi fundamental para iluminar o panorama do mercado de EBIs residenciais em São Paulo frente às inovações tecnológicas como o ML.

Por fim, agradeço a todos que compuseram minha rede de apoio emocional, compreensivos durante minha ausência para me dedicar a este projeto e ao curso em geral. À Cindy, minha amada cachorrinha que partiu durante este trabalho, mas esteve ao meu lado durante grande parte dele, minha eterna gratidão.

RESUMO

O mercado imobiliário, marcado por decisões complexas com diversas variáveis, está se beneficiando do crescente acesso a dados e do progresso das tecnologias de *Machine Learning* (ML). Este estudo tem como foco a intersecção de dois domínios: [1] ML e [2] Valoração de imóveis residenciais para locação (EBIs), com o objetivo de [A] avaliar os benefícios do ML na melhoria dos processos de valoração desses ativos, minimizando vieses e subjetividade na aproximação do *Market Tradable Value*, e [B] examinar o grau de aderência deste segmento à tecnologia ML, especificamente na cidade de São Paulo. Como metodologia, além da revisão bibliográfica sobre as definições e estudos de caso de ML para o mercado de *Real Estate*, foi adotada a Survey, com a obtenção de dados através de entrevistas de 8 empresas selecionadas no mercado de EBIs residenciais em São Paulo, divididas em 3 categorias: C (comercialização de EBIs), LS (desenvolvimento e administração de EBIs de longa estadia) e BD (fornecimento de inteligência baseada em dados para EBIs). Os resultados indicaram que: [A] a ML melhora a precisão da avaliação imobiliária e permite identificar relações complexas que não são possíveis por modelos tradicionais; [B] 62% das empresas da amostragem entrevistada já utilizam ML e 75% veem potencial para o uso de tecnologias de IA e ML em seu segmento, sendo que, desse universo, 100% já utiliza dados para processos da empresa em geral e já estão familiarizados com o conceito de ML e, adicionalmente; [C] a barreira documental de acesso a dados de transação no mercado brasileiro é um dos principais desafios identificados pelos entrevistados para que o uso das tecnologias de ML possam ser mais difundidas no mercado imobiliário de EBIs residenciais.

Palavras chaves: *Machine Learning* (ML), *Real Estate*, Empreendimento de Base Imobiliária (EBI), Valoração, Imóveis residenciais, São Paulo

ABSTRACT

The real estate market, marked by complex decisions with several variables, is benefiting from increasing access to data and the progress of Machine Learning (ML) technologies. This study focuses on the intersection of two domains: [1] ML and [2] Valuation of residential rental properties (EBIs), with the objective of [A] evaluating the benefits of ML in improving the valuation processes of these assets, minimizing biases and subjectivity in approaching their Market Tradable Value, and [B] examining the degree of engagement of this segment to ML technology, specifically in the city of São Paulo. As a methodological approach, in addition to the literature review on the definitions and case studies of ML for the Real Estate market, the Survey was adopted, obtaining data through interviews with 8 of the 10 most relevant companies in the residential EBIs market in São Paulo, divided into 3 categories: C (sale of residential rental assets), LS (development and management of long-stay residential properties) and BD (providing data-based intelligence for the sector). The results shows that [A] ML improves the accuracy of real estate valuation and allows for the identification of complex relationships that are not possible with traditional models, [B] 62% of the companies in the surveyed sample already use ML and 75% see potential for the use of AI and ML technologies in their segment, with 100% of these already using data for general company processes and being familiar with the concept of ML, and, additionally, [C] the documentary barrier to accessing transaction data in the Brazilian market is one of the main challenges identified by respondents for the wider use of ML technologies in the residential EBI real estate market.

Keywords: Machine Learning (ML), Real Estate, Residential Rental Properties (EBI), Valuation, São Paulo

LISTA DE FIGURAS

Figura 1 Categorias de Big Data que podem ser adotados – Dados Estruturados e Dados Não-estruturados, Internos e Externos. Fonte: (NAIR & NARAYANAN, 2012).....	22
Figura 2 Ciclos característicos e transações financeiras. Fonte: retirado de Florencio de Alencar (2016).....	26
Figura 3 Modelo Básico do Enfoque de CBR – Retirado de WANGENHEIM, 2003...	32
Figura 4 Ciclo de Funcionamento de sistema RBC – Retirado de WANGENHEIM, 2003	32
Figura 5 Regressão Linear - Parâmetros de transparência e ilustração - imagem retirada de ARRIETA	41
Figura 6 - Árvore de Decisão - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP	42
Figura 7 - KNN - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP	43
Figura 8 - RBL - Parâmetros de transparência e ilustração - adaptado de ARRIETA e ROSS	44
Figura 9 - GMA - Parâmetros de transparência e ilustração – adaptado de ARRIETA	46
Figura 10 - Modelo Bayesiano - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP	47
Figura 11 Ilustração sobre resultado da redução do espaço amostras - Extraído de RUSSEL (2020), p.749.....	48
Figura 12 Modelo Tree ensembles, random forest and multiple classifier systems - Parâmetros de transparência - adaptado de ARRIETA	48
Figura 13 Randon Forest - Ilustração e Aplicação - Produção original	49
Figura 14 Bossting - - Ilustração e Aplicação – Adaptado de LUCIC	50
Figura 15 Stacking - Ilustração e Aplicação - ilustração explicativa adaptada de ARRIETA, CONWAY e ROSS	51
Figura 16 - Modelo SVM - Parâmetros de transparência, ilustração explicativa e exemplo de aplicação no mercado de valorização, em <i>Real Estate</i> - retirado de ARRIETA, BISHOP e CONWAY	52
Figura 17 Modelo Neural Networks - Parâmetros de transparência, ilustração explicativa adaptada de CONWAY	54

Figura 18 Ilustração da Estrutura de "Feedforward" e "Backpropagation" - Diagramas extraídos de CIABURRO, 2021	58
Figura 19 Quadro resumo de agrupamento das empresas selecionadas para entrevista. Fonte: Elaboração própria.	72
Figura 20 Pergunta 2 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.....	83
Figura 21 Pergunta 3 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.....	85
Figura 22 Pergunta 4 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.....	86
Figura 23 Pergunta 5 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.....	87
Figura 24 Pergunta 6 e Pergunta 3, respectivamente – indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.	91
Figura 25 Pergunta 7 e Pergunta 5, respectivamente – indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.	94
Figura 26 Pergunta 8 e Pergunta 2, respectivamente - indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.	98
Figura 27 Classificação quanto a qualidade das respostas às perguntas 6, 7 e 8. Fonte: Elaboração própria.....	101
Figura 28 Indicação de quantidade algoritmos utilizados nos estudos de caso citados no capítulo 3 e citados pelas empresas entrevistadas no questionário aplicado, respectivamente. Fonte: Elaboração própria.	105
Figura 29 Comparação entre respostas adquiridas nas entrevistas de Conway (2018) e a autora. Fonte: Elaboração Própria.	107

LISTA DE TABELAS

Tabela 1 Objetivos da XAI – Baseado em ARRIETA et al., 2019	37
Tabela 2 Conceitos para transparência – Baseado em ARRIETA et al., 2019.....	38
Tabela 3 Classificação de modelos de ML a partir do seu nível de explicabilidade – Baseado em ARRIETA et al., 2019.....	39
Tabela 4 Resumo dos estudos de caso.....	61
Tabela 5 Lista de Questões Aplicadas no Questionário (Seção 1: Foco em Aplicação na Empresa): Versão Original, Traduzida e Adaptada - com comentários sobre os ajustes realizados.....	69
Tabela 6 Lista de Questões Aplicadas no Questionário (Seção 1: Foco na Opinião Geral de Mercado): Versão Original, Traduzida e Adaptada - com comentários sobre os ajustes realizados.....	70

LISTA DE ABREVIATURAS E SIGLAS

Geral

AIA	<i>Artificial Intelligence Act</i>
AVM	<i>Automated Valuation Models</i>
FTA	<i>Financial Transparency Act</i>
IA	Inteligência Artificial
MAE	<i>Mean Absolute Error</i>
MTV	<i>Market Tradable Value</i>
ML	<i>Machine Learning</i>
RBC	Raciocínio Baseado em Casos
RGPD	Regulamento Geral sobre a Proteção de Dados
RODi	Resultado Operacional Disponível
XAI	<i>Explainable Artificial Intelligence</i>

Algoritmos e Funções

BM	<i>Bayesian Models</i>
CN	<i>Convolutional Networks</i>
DT	<i>Decision Trees</i>
GBM	<i>Gradient Boosting Machine</i>
GMA	<i>General Additive Models</i>
KNN	<i>K-Nearest Neighbours</i>
LR	<i>Linear/ Logistic Regression</i>
MCS	<i>Multiple Classifier Systems</i>
MLN	<i>Multi-Layer Network</i>
NN	<i>Neural Networks</i>
RBL	<i>Rule Based Learners</i>
ReLU	<i>Rectified Linear Unit</i>
RF	<i>Random Forest</i>
RN	<i>Recurrent Networks</i>
RNN	<i>Recurrent Neural Networks</i>
SVM	<i>Support Vector Machines</i>
TE	<i>Tree Ensemble</i>

SUMÁRIO

1	INTRODUÇÃO.....	14
1.1	JUSTIFICATIVA	14
1.2	OBJETIVOS	16
1.2.1	Objetivos Primários	16
1.2.2	Objetivos secundários.....	17
1.3	MÉTODOS DE PESQUISA.....	17
1.4	ESTRUTURA DO TRABALHO	18
2	REVISÃO BIBLIOGRÁFICA.....	20
2.1	CONTEXTO DA APLICAÇÃO DE <i>MACHINE LEARNING</i> NO MERCADO IMOBILIÁRIO.....	20
2.2	DEFINIÇÕES E CONCEITOS	25
2.2.1	EMPREENDIMENTOS DE BASE IMOBILIÁRIA (EBIs)	25
2.2.2	METODOLOGIAS DE VALORAÇÃO PARA EMPREENDIMENTOS DE BASE IMOBILIÁRIA (EBIs).....	27
2.2.3	INTELIGENCIA ARTIFICIAL (IA).....	31
2.2.4	<i>MACHINE LEARNING</i> (ML).....	35
3	APLICAÇÕES DE MODELOS DE <i>MACHINE LEARNING</i> PARA ARBITRAGEM DE VALOR EM EMPREENDIMENTOS RESIDENCIAIS	60
4	METODOLOGIA DE PESQUISA	67
4.1	METODOLOGIA.....	67
4.2	FORMULAÇÃO DO QUESTIONÁRIO.....	67
4.3	SELEÇÃO DAS EMPRESAS.....	71
4.4	APLICAÇÃO DO QUESTIONÁRIOS.....	77
4.5	APRESENTAÇÃO DOS DADOS	79
5	ANÁLISE DOS DADOS	99
6	CONCLUSÕES E RECOMENDAÇÕES PARA FUTUROS TRABALHOS.	108
6.1	RECOMENDAÇÕES PARA FUTUROS TRABALHOS.....	110

REFERÊNCIAS	112
APÊNDICE 1	116

1 INTRODUÇÃO

No capítulo inicial deste trabalho, será apresentada a justificativa para a pesquisa, destacando a importância que tecnologias como *Machine Learning* (ML) podem desempenhar na arbitragem de valor de ativos do mercado imobiliário residencial de base imobiliária. Além disso, serão apresentados os objetivos do estudo, a metodologia a ser seguida e uma visão geral da estrutura dos capítulos subsequentes. Isso permitirá o entendimento, de forma sucinta, sobre o que esperar ao longo deste trabalho.

1.1 JUSTIFICATIVA

No Empreendimento de Base Imobiliária (EBI), o principal objetivo do empreendedor é obter renda por meio da exploração do espaço físico construído (LIMA JR., MONETTI, & DE ALENCAR, 2011). Nesse sentido, seja nos mercados de investimentos primários ou nos de portfólios, a decisão de se investir é, em grande parte, apoiada quase exclusivamente na relação entre aluguel praticado e preço da transação (LIMA JR., MONETTI, & DE ALENCAR, 2011), que são diretamente associados ao poder de troca do bem e, portanto, no valor que este bem será transacionado.

Levando em conta que o EBI possui um ciclo de vida longo, existe uma série de fatores, que vão influenciar o seu valor no momento da transação, desde características físicas, de localização e qualidade do imóvel, até conceitos mais abstratos, como oferta e demanda, taxas de juros e economia em geral (LIMA JR., MONETTI, & DE ALENCAR, 2011). Nesse sentido, essa diversidade de agentes que moldam a decisão do investidor, que é o tomador do risco, pode dificultar o prognóstico do *Market Tradable Value* (MTV) para arbitragem do preço do ativo e do fluxo de recebimento de resultados provenientes da operação do empreendimento (RODi - Resultado Operacional Disponível), que é a renda proveniente do ativo para o investidor (LIMA JR., MONETTI, & DE ALENCAR, 2011).

Conforme os estudos de caso que serão apresentados no Capítulo 3, existem evidências de que a aplicação de modelos de ML apresenta menos erros de estimação quando comparadas outros modelos de valoração imobiliária que serão mais detalhadamente

discutidos no Capítulo 2. Isso ocorre em grande parte devido a capacidade de capturar relacionamentos não lineares de atributos que possuem influência sobre o valor percebido do imóvel. Além disso, os modelos de ML podem considerar não apenas os atributos estruturados como também as informações não estruturadas, como descrições de propriedades e imagens. Isso viabiliza uma valoração, muitas vezes, mais precisa e, no Capítulo 3, serão ilustrados alguns exemplos.

A Inteligência Artificial (IA), termo criado pelo professor John McCarthy, em 1955, foi definida por ele como "a ciência e engenharia de fazer máquinas inteligentes" (MANNING, 2020), já a aprendizagem de máquina (ML) é a parte da IA que estuda como os agentes computacionais podem melhorar suas recomendações com base em experiências ou dados. Para isso, o ML se baseia em ciência da computação, estatística, psicologia, neurociência, economia e teoria de controle (MANNING, 2020).

As primeiras aplicações de IA e ML em *Real Estate* datam dos anos 1960. No entanto, foi apenas nos últimos anos que seu uso tem evoluído e que seus benefícios tem se tornado cada vez mais acessíveis para o mercado imobiliário. À medida que a tecnologia da IA e do ML continua a se desenvolver, é provável que seu uso no mercado imobiliário aumente e, conseqüentemente, o uso dessas tecnologias pode representar uma grande vantagem competitiva (CONWAY, 2018).

Segundo (BAUM, 2017), o setor imobiliário é relutante em adotar mudanças rapidamente e, principalmente devido as transações acontecerem em sua maioria no mercado privado, existe um nível de protecionismo que acaba dificultando a adoção de mudanças tecnológicas. Complementarmente, em entrevista realizada com professor David Geltner (CONWAY, 2018), professor PhD no departamento de Real Estate da universidade do MIT, ele comenta que parte da hesitação das empresas do setor imobiliário ocorre também devido a atritos culturais e o desejo de não compartilhar dados.

É possível notar o avanço na adoção dessas tecnologias, conforme apontado no Capítulo 3. Entretanto, há muitos obstáculos a serem superados na aplicação da IA e do ML no setor imobiliário, como o de convencer os profissionais a adotarem as novas tecnologias ou a necessidade de ter acesso a dados imobiliários de transações realizadas, investimento de recursos para desenvolvimento do mercado de trabalho, tempo para testar e estruturar os dados a serem utilizados, entre outros.

Este trabalho busca enriquecer o repertório disponível com relação a conexão entre o mercado de *Real Estate*, mais especificamente dentro do nicho de EBIs residenciais, e as potenciais colaborações que o ML pode representar às estratégias empresariais e ao planejamento estratégico dentro do setor. A partir disso, permitindo que profissionais do mercado imobiliário no Brasil possam avaliar a viabilidade dessas soluções tecnológicas, aproveitando o potencial que o valor de seus próprios dados pode agregar aos seus resultados.

Portanto, a autor pretende definir os conceitos que sustentam o entendimento sobre o que é o ML para o público do setor imobiliário. Serão examinadas as maneiras pelas quais essas novas tecnologias vêm sendo utilizadas no setor imobiliário, a partir de revisão bibliográfica e entrevistas com profissionais de mercado, e apresentadas as potenciais aplicações futuras e os desafios para o crescimento da sua curva de adoção dentro do setor residencial de EBIs no mercado de São Paulo.

1.2 OBJETIVOS

Esse trabalho possui os seguintes objetivos:

1.2.1 Objetivos Primários

O propósito fundamental deste trabalho consiste em abordar duas questões centrais que permitam o entendimento sobre o panorama de aderência do uso de tecnologias de ML no setor imobiliário, mais especificamente, em ativos residenciais de base imobiliária voltados a valoração.

- A. Como as tecnologias de ML estão sendo aplicadas no setor residencial de base imobiliária no segmento de valoração?
- B. Quais oportunidades concretas se apresentam para a adoção de ML no setor residencial de base imobiliária?

Utilizando uma abordagem que combina (1) revisão de bibliografia especializada, tanto para leitura do contexto e conceituação de aplicação dessas tecnologias quanto para avaliar estudos de caso que apresentam resultados concretos de melhoria de precisão ao aplicar tais tecnologias, e (2) com a aplicação de entrevistas com profissionais do mercado

que buscam obter informações sobre o estágio de adoção das tecnologias de ML no mercado imobiliário de ativos residenciais de base imobiliária em São Paulo.

1.2.2 Objetivos secundários

Além do objetivo central, este trabalho visa:

- I. Comparar os resultados obtidos em entrevistas com profissionais de mercado do nicho de Valoração, com atuação em São Paulo, com aqueles encontrados na pesquisa de mestrado de Conway (2018), realizada para o Instituto de Tecnologia de Massachusetts (MIT): “*Artificial Intelligence and Machine Learning: Current Applications in Real Estate*”;
- II. Conduzir um levantamento abrangente de algoritmos de ML aplicados à arbitragem de valor para transações imobiliárias de ativos residenciais;
- III. Compilar informações recentes sobre a implementação de algoritmos cujo propósito serve ao aprimoramento da eficiência e precisão de arbitragem de valor para transações imobiliárias de EBI residenciais.

1.3 MÉTODOS DE PESQUISA

A metodologia deste estudo é SURVEY, que é baseada num instrumento de pesquisa, normalmente um questionário, a fim de obter dados ou informações sobre características, ações ou opiniões de determinado grupo de pessoas, que representam uma população-alvo (FONSECA, 2002).

Esta abordagem consiste em (1) uma revisão bibliográfica, complementada pelo levantamento de aplicação das tecnologias de IA e ML em estudos de caso, (2) realização de entrevistas e (3) construção de uma matriz de correlações entre as áreas do setor de *Real Estate* e as técnicas de ML aplicadas.

1. Inicialmente, a revisão bibliográfica envolveu uma pesquisa ampla de literatura acadêmica, que incluiu fontes como livros consolidados e artigos recentes. Cada fonte foi abordada com foco nos seguintes aspectos: contexto, conceito e exemplos de estudos de caso. Essa revisão serviu como base tanto para o capítulo

2, cujo objetivo é definir conceitos como IA e ML para o público de investidores e profissionais do mercado imobiliário, quanto para o Capítulo 3, que explora como essas novas tecnologias estão sendo utilizadas na indústria imobiliária, além de considerar possíveis aplicações futuras e os desafios associados.

2. A segunda etapa, seguindo a estrutura do questionário de Conway (2018), envolve a aplicação de entrevistas. Essas entrevistas têm como foco responder a duas questões centrais:

A. Como as tecnologias de IA e ML estão sendo aplicadas no setor residencial de base imobiliária?

B. Quais oportunidades concretas se apresentam para a adoção de IA e ML no setor residencial de base imobiliária?

Para abordar essas questões, foi desenvolvido um questionário com 8 perguntas, mais detalhadas no Capítulo 4, com o propósito de avaliar a aderência das empresas selecionadas quanto ao conhecimento e a aplicação das tecnologias em questão. Além disso, as entrevistas buscam identificar como essas tecnologias já estão sendo implementadas nas empresas dos entrevistados e explorar as possibilidades futuras que elas podem oferecer.

3. Após a realização da pesquisa, serão elaboradas matrizes, que ilustrarão os resultados e indicarão o nível de adoção de cada empresa diante das tecnologias de ML, acompanhadas por uma análise que incluirá informações mencionadas durante a entrevista.

1.4 ESTRUTURA DO TRABALHO

Capítulo 1: Introdução

Este capítulo inicial apresenta a justificativa da pesquisa, ressaltando a relevância da utilização de tecnologias como o ML para aprimorar os processos de arbitragem de preço de transação para venda e para locação de ativos residenciais EBI. Além disso, delinea os objetivos do estudo, descreve a metodologia adotada e fornece uma visão geral da

estrutura dos capítulos subsequentes, oferecendo uma síntese do que os leitores podem esperar ao longo deste trabalho.

Capítulo 2: Revisão Bibliográfica

Este capítulo envolveu uma pesquisa ampla sobre o papel do ML no setor imobiliário, com ênfase na arbitragem de preço de transação para venda e para locação de ativos residenciais EBI. Começando com uma apresentação do contexto, serão analisados os principais benefícios e desafios associados a essas tecnologias no mercado imobiliário e examinadas algumas perspectivas futuras. Seguido para a exploração dos conceitos relacionados aos Empreendimentos de Base Imobiliária e as principais metodologias de valoração focadas nesse tipo de ativo. Posteriormente, serão apresentados os conceitos essenciais e principais algoritmos que sustentam o ML.

Capítulo 3: Aplicações de Modelos de ML para Arbitragem de Valor em EBIs

Neste capítulo serão pontuados exemplos práticos das aplicações das tecnologias de ML e exploradas quais as melhorias trazidas aos processos de valoração de bens residenciais de base imobiliária.

Capítulo 4: Survey

Neste capítulo será apresentada a formulação do questionário utilizado e apresentados os dados obtidos a partir das respostas coletadas.

Capítulo 5: Análise dos Dados

Neste capítulo será feita uma análise aprofundada dos dados coletados por meio das respostas ao questionário.

Capítulo 6: Conclusão e Recomendação para Futuros Trabalhos

Neste capítulo serão apresentadas tanto as principais conclusões derivadas das análises realizadas ao longo da pesquisa, quanto as recomendações para possíveis investigações futuras, visando contribuir com a ampliação e aprofundamento do conhecimento no campo de estudo.

2 REVISÃO BIBLIOGRÁFICA

Esta seção contém uma base teórica que é fundamental para uma compreensão aprofundada dos estudos de caso apresentados no Capítulo 3. Inicialmente, serão conceituados os Empreendimentos de Base Imobiliária (EBIs) e as metodologias de valoração para arbitragem de valor desses ativos. Após, será feita uma breve contextualização da conexão entre ML e as oportunidades que essa tecnologia oferece para a melhoria desses processos. Posteriormente, serão abordados os conceitos e os algoritmos que constituem a aplicação do ML.

2.1 CONTEXTO DA APLICAÇÃO DE *MACHINE LEARNING* NO MERCADO IMOBILIÁRIO

Nesta seção, será apresentado um breve contexto que engloba uma visão geral da relação entre o ML com o mercado imobiliário em geral, abordando oportunidades e obstáculos.

Segundo o levantamento sobre Tendências Tecnológicas, realizado pela consultoria McKinsey, foram identificadas 15 categorias principais de tecnologias que se destacaram em 2022. Dentro das categorias listadas, a única em que o setor imobiliário, enquanto um investimento com aspectos econômicos e financeiros, foi considerado da lista de indústrias mais afetadas, foi o grupo “Inteligência Artificial Aplicada”, ou “*Applied AI*” – em que o *Machine Learning* está inserido.

Para uma análise comparativa da situação de cada grupo categorizado, o relatório criou uma classificação com base em quatro parâmetros:

- Grau de inovação (“*Innovation Score*”), medido pelo número de patentes e pesquisas relacionadas a essas tecnologias;
- Nível de interesse (“*Interest Score*”), avaliado por meio da quantidade de publicação de notícias e volume de pesquisas *online*;
- Nível de engajamento (“*Adoption Rate*”), avaliado pela quantidade de empresas que aplicam cada tecnologia, e

- Quantidade de investimento (“*Equity Investments*”), considerando o capital levantado tanto no mercado privado quanto no mercado público.

De acordo com o **Gráfico 1**, a Inteligência Artificial aplicada se destaca entre os três maiores com “equity investment”, ou seja, com o maior nível de investimento. Dentre eles, a IA aplicada, é o único que possui um alto nível de adoção, estando na escala 5 do quesito “adoption rate”. Isso indica que o investimento nessa tecnologia foi distribuído entre um número maior de investidores, demonstrando um maior alcance e menor concentração do capital que está sendo investido. Além disso, conforme os eixos do gráfico, a IA aplicada é a tendência com maior nível de inovação (eixo y) e ocupa a terceira posição em termos de interesse (eixo x), com uma grande quantidade de artigos e patentes, além de uma alta demanda *online* sobre o tema.

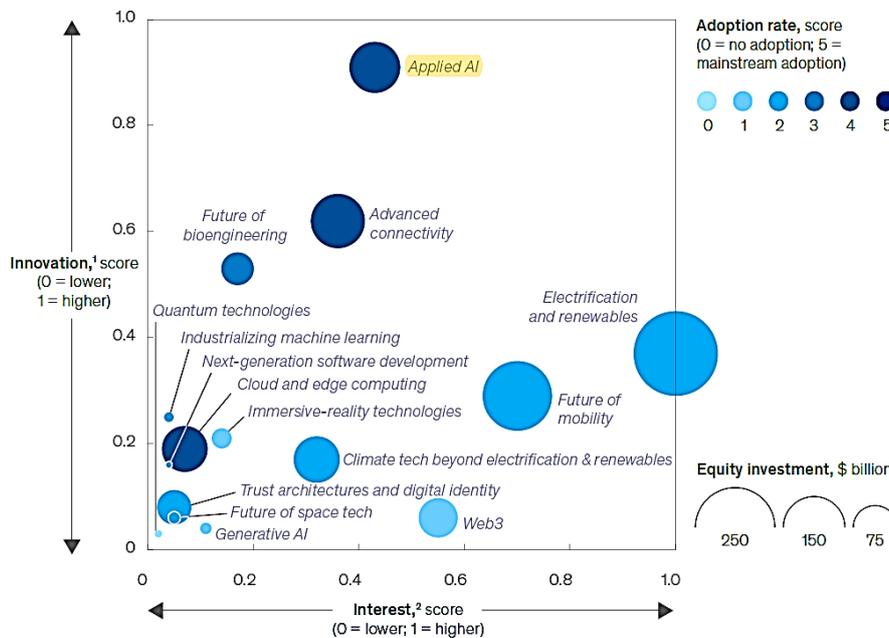


Gráfico 1 Classificação de principais tecnologias inovadoras em 2022 conforme: grau de inovação x nível de interesse x nível de engajamento x quantidade de investimentos. Fonte: (MCKINSEY & COMPANY, 2023)

As tecnologias de IA, como o Aprendizado de Máquina, estão modificando o cenário competitivo de muitas indústrias e pesquisas recentes destacam que incorporar inteligência artificial em processos existentes pode produzir resultados melhores e ser mais eficiente no setor imobiliário, com destaque para métodos de prognóstico de valor de ativos (CHAN, HOGABOAM & CAO, 2022). Esses exemplos serão detalhados no Capítulo 3, em que serão abordados os exemplos de aplicação de tecnologias de ML no mercado imobiliário de EBIs.

Como parte fundamental do progresso dessas tecnologias, o crescimento dos dados revela que o Big Data está expandindo em uma taxa sem precedentes¹. O crescimento exponencial na quantidade de dados originou tanto dados estruturados (i.e. informações organizadas em planilhas) quanto não-estruturados (i.e. fotos, vídeos, textos), conforme apontado na **Figura 1**, abaixo.

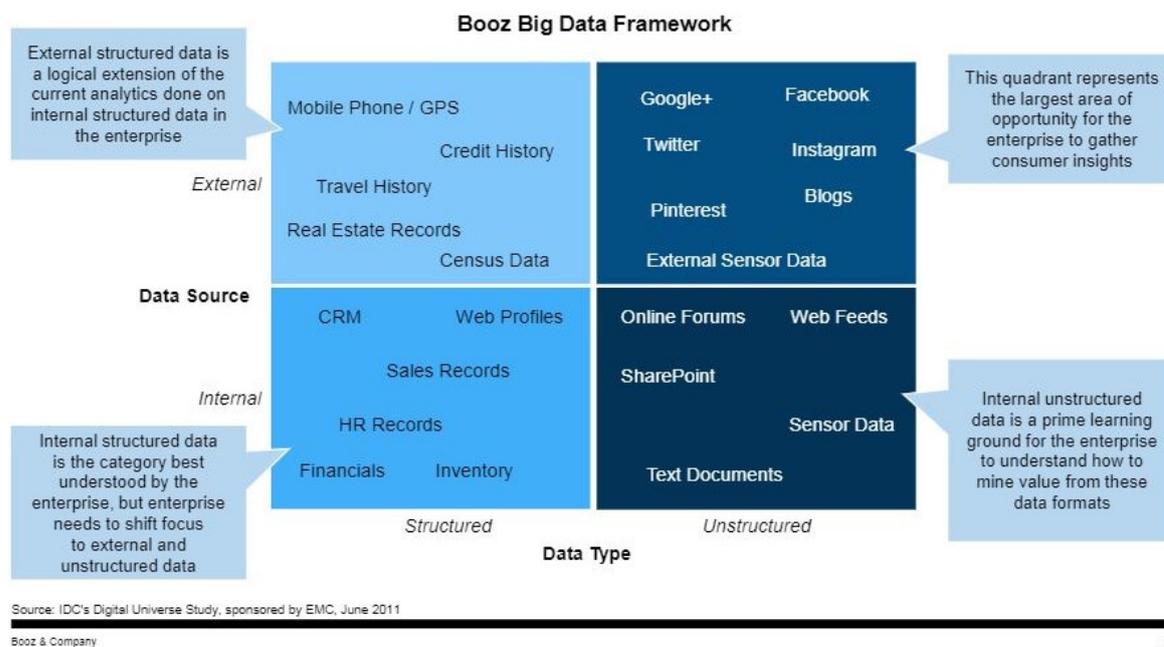


Figura 1 Categorias de Big Data que podem ser adotados – Dados Estruturados e Dados Não-estruturados, Internos e Externos. Fonte: (NAIR & NARAYANAN, 2012)

Embora os dados brutos possam não ter valor por si só, as informações que podem ser extraídas deles através de métodos avançados, utilizando análises preditivas, inteligência artificial, mineração de dados, ciência de dados e etc, são uma fonte de grande valor para corretores e desenvolvedores imobiliários, ainda mais quando existe a contextualização de tais dados, tornando-os georreferenciados, ou seja, associados a uma referência de localização (CHAN, HOGABOAM & CAO, 2022).

A fim de identificar os tipos de dados dentro do setor imobiliário que podem ser utilizados para definir o preço de um ativo imobiliário, Winson-Geidman e Krause (2016) explicam que existem três tipos de dados imobiliários: dados financeiros, dados transacionais e dados físicos. Os dados financeiros englobam ações ou fundos relacionadas ao mercado imobiliário, influenciados por taxas de juros, as políticas governamentais relacionadas ao

¹ Fonte da informação: <https://explodingtopics.com/blog/big-data-trends> (extraída em Janeiro/2024)

setor imobiliário e a economia em geral, os dados transacionais referem-se a informações como hipotecas e aluguéis, e os dados físicos incluem informações de localização e características estruturais do objeto em questão. Esses três tipos de dados imobiliários constituem o conjunto de big data disponível para as empresas utilizarem técnicas de ML na arbitragem do valor de uma casa, por exemplo.

Estima-se que aproximadamente 80% dos dados mantidos pelas empresas são não-estruturados (Dayley & Logan (2015) apud POTRAWA & TETEREVA (2022)), esse percentual é representado principalmente pelo uso difundido de whatsapp. Em paralelo, avalia-se que o volume de dados não-estruturados está crescendo 15 vezes mais rapidamente do que os estruturados (Nair & Narayanan (2012) apud POTRAWA & TETEREVA (2022)). Conforme apontado por Potrawa (2022), isso pode ser importante para descobrir novos fatores e dependências que não são capturados pelo uso de dados estruturados convencionais e, igualmente, através de modelos estatísticos tradicionais, que serão explicados mais detalhadamente no Capítulo 2.

Vale observar que uma das maiores dificuldades para análise de dados no mercado brasileiro, focando mais precisamente em São Paulo, se encontra no acesso aos dados de transações imobiliárias e na divulgação pública das transações e sua transparência com relação à divulgação dos valores negociados (TAKAOKA, 2009). Conforme citado em “Valuation of income Real Estate Properties do loan guarantee...”, por FLORENCIO e ALENCAR em 2016:

“Ademais, no Brasil, além da falta de transações abrangendo tais empreendimentos [EBI], quando estas ocorrem, o acesso à informação nem sempre é fácil e o conjunto de fatores que levaram a negociação a um certo patamar de valor nem sempre está claro, o que pode conduzir a subjetividade do processo de avaliação [...]”

(FLORENCIO & ALENCAR, 2016)

Ainda assim, mesmo assumindo o acesso a tais dados, existem desafios a serem superados para a adoção em larga escala do ML no mercado imobiliário. De acordo com Conway (2018), baseado na análise de 80 empresas do setor imobiliário nos Estados Unidos, sendo 27 delas entrevistadas pela pesquisadora para obter mais detalhes sobre o uso de tecnologias de IA e ML, um dos pontos destacados foi sobre a necessidade de demonstrar

resultados concretos e evidências de retorno sobre o investimento para conquistar a confiança dos profissionais do ramo imobiliário.

Todavia, somente crescem estudos acadêmicos e exemplos de empresas maduras ou em ascensão no mercado imobiliário que buscam validar o potencial que a aplicação de tecnologias como o ML podem oferecer como diferencial competitivo, conforme será apresentado no Capítulo 3, em que o grau de precisão para arbitragem de valor de ativos imobiliários por algoritmos de ML se mostra superior aos métodos tradicionais de valoração.

Contudo, um fator importante a ser considerado é que não é o modelo que dita a arbitragem de valor do ativo, mas sim o profissional que deve ter esse discernimento, utilizando a tecnologia como ferramental para melhorar e justificar sua tomada de decisão (DAMORAN, 2012). Sendo assim, levando em conta a importância da responsabilização e regulamentação do mercado de *Real Estate* para processo de valoração, a transparência dos processos é um fator importante para que o setor imobiliário absorva tais tecnologias.

Ainda que não exista uma regulamentação específica no Brasil para controlar os limites no uso do ML no mercado imobiliário, em 2020, houve o primeiro registro de uma regulação em IA para a jurisdição feita por uma Resolução do Conselho Nacional de Justiça, a Resolução 332. Nessa Resolução, são apontados aspectos éticos de transparência e governança que buscam reconhecer potenciais usos para essa tecnologia e destacar alguns cuidados que devem ser levados em consideração para a sua aplicação respeitando os direitos fundamentais (PEIXOTO, 2020).

Corroborando com a preocupação em garantir a transparência e a ética no uso dessas tecnologias, algumas iniciativas legislativas recentes podem ser destacadas, como a Lei de Transparência Financeira dos EUA de 2021 (FTA - "Financial Transparency Act") ou a Lei de Inteligência Artificial da União Europeia (AIA - "Artificial Intelligence Act"), estabelecem níveis de transparência que devem ser divulgados para aplicação na prática da tomada de decisões apoiada pela IA (WEBER, CARL, & HINZ, 2023, p. 3).

Por isso, campos de estudo como o XAI (“Explainable Artificial Intelligence”) estão emergindo com o objetivo de mitigar problemas chamados “black-box”², em que a opacidade dos algoritmos é associada a arbitrariedade de critérios e de conclusões (PEIXOTO, 2020). Os métodos do XAI serão detalhados na seção 2.2.4 - *Machine Learning* (ML) deste capítulo e serão utilizados como estrutura para auxiliar a descrição dos algoritmos que, conforme a bibliografia levantada, são os mais representativos atualmente.

Nesse contexto, é importante destacar que existe uma grande oportunidade para que o ML ofereça melhorias significativas nos processos que envolvem o mercado imobiliário, conforme será apontado no Capítulo 3 e durante a aplicação das entrevistas com profissionais de mercado. Entretanto, essas tecnologias enfrentam desafios relacionados à qualidade dos dados, conformidade regulatória e adoção por parte dos profissionais do setor. À medida que mais avanços são feitos, tanto na oferta de mão-de-obra especializada quanto no aumento de casos de sucesso, é provável que o ML desempenhe um papel ainda maior na evolução do mercado imobiliário (CONWAY, 2018).

2.2 DEFINIÇÕES E CONCEITOS

Esta seção inicia-se com a conceituação de Empreendimentos de Base Imobiliária (EBIs) e metodologias de valoração que envolvem esses ativos. Posteriormente, serão abordados os conceitos que permeiam o *Machine Learning*, incluindo uma introdução sobre Inteligência Artificial e a exploração de algoritmos que constituem a aplicação do ML. Essa base teórica é fundamental para uma compreensão aprofundada dos estudos de caso apresentados no Capítulo 3.

2.2.1 EMPREENDIMENTOS DE BASE IMOBILIÁRIA (EBIs)

Tomando como base a definição de (LIMA JR., 1994), Empreendimentos de base imobiliária (EBIs) representam investimentos cujo estudo econômico/financeiro está intrinsecamente ligado ao desempenho operacional do imóvel, seja através de aluguéis

² O termo *black box* faz referência aos diferentes níveis do fechamento interno de um sistema, restringindo sua exposição de explicação sobre o design interno, estrutura e implementação ao usuário externo (PEIXOTO, 2020).

pré-determinados ou da exploração de atividades específicas. Exemplos comuns incluem shopping centers, hotéis, centros de distribuição, condomínios industriais, e espaços de escritórios para locação. Esses empreendimentos são geralmente de longo prazo e envolvem ciclos financeiros distintos, levando os agentes do mercado a adotarem abordagens diferenciadas ao longo de cada fase, desde a concepção até a exaustão do empreendimento.

Segundo análises de especialistas como (MONETTI, 1996) e (AMATO, 2001), os EBIs passam por quatro ciclos distintos: o ciclo de formatação, envolvendo o planejamento inicial do empreendimento; o ciclo de implantação, que abrange as etapas de construção e desenvolvimento; o ciclo operacional, onde o empreendimento entra em funcionamento e gera receita; e, por fim, o ciclo de exaustão, marcado pela obsolescência ou mudança de uso do imóvel. Essa dinâmica, ilustrada na Figura 2, é cíclica e influencia diretamente as estratégias adotadas pelos diversos agentes envolvidos, contribuindo para uma abordagem fracionada na gestão e no desenvolvimento desses projetos imobiliários.

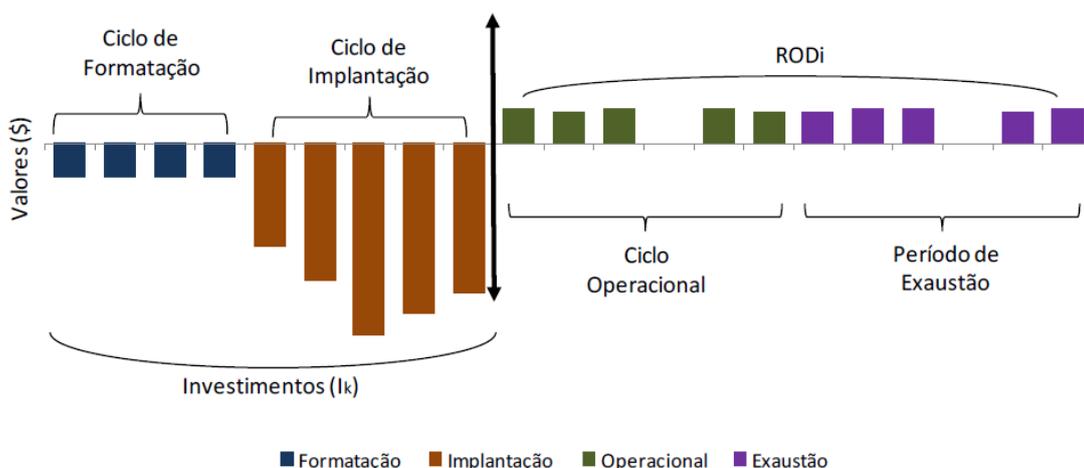


Figura 2 Ciclos característicos e transações financeiras. Fonte: retirado de Florencio de Alencar (2016)

A Figura 2 ilustra de maneira clara as transações financeiras entre os investidores e/ou empreendedores e o ambiente do empreendimento ao longo dos ciclos mencionados, destacando a distinção entre as fases de investimento inicial e obtenção de retorno através de *RODi* (Resultado Operacional Disponível). A primeira fase engloba os ciclos de formatação e implantação, caracterizados pelos investimentos iniciais por parte dos empreendedores, enquanto a segunda fase compreende o ciclo operacional e de exaustão,

marcados pela geração de retorno a partir do desempenho das atividades exploradas no empreendimento.

Impulsionado pela disponibilidade crescente de dados e pelas inovações trazidas pela internet, observa-se um crescimento no segmento de locação residencial (EBIs residenciais). A Pesquisa Nacional por Amostra de Domicílios (Pnad Contínua) indica que a condição de domicílio alugado aumentou, saindo de 17,3% em 2016 para 20,2% em 2022³.

Este estudo tem como recorte o estudo do mercado de EBIs residenciais em São Paulo, visando afunilar as opções para análise de aplicação de tecnologias de *Machine Learning* e para a seleção de empresas a serem entrevistadas. Desta forma, buscando obter resultados mais relevantes dentro deste nicho.

2.2.2 METODOLOGIAS DE VALORAÇÃO PARA EMPREENDIMENTOS DE BASE IMOBILIÁRIA (EBIs)

Diversas abordagens são empregadas na valoração de Empreendimentos de Base Imobiliária (EBIs), cada uma com seus modelos, aplicabilidades e limitações distintas. Conforme (FLORENCIO & ALENCAR, 2016), existem três abordagens clássicas para a valoração de EBIs: [i] comparativa (ou relativa), [ii] custo de reprodução e [iii] renda, resumidas abaixo:

I. Abordagem comparativa

Na abordagem comparativa, a avaliação de um ativo baseia-se na comparação com ativos similares, normalmente padronizados por variáveis como lucros, fluxos de caixa, valores contábeis ou receitas, a fim de determinar seu valor relativo no mercado.

II. Abordagem pelo custo

Segundo Brueggeman e Fisher (2011), extraído de (FLORENCIO & ALENCAR, 2016), na abordagem pelo custo, o valor de uma propriedade é calculado a partir do custo de reprodução, partindo-se do princípio de que nenhum comprador informado estaria disposto a pagar mais por um imóvel

³ Fonte: <https://agenciadenoticias.ibge.gov.br> (acessado em Fevereiro/2024)

do que o custo para construir uma propriedade idêntica em um terreno semelhante.

III. Abordagem pela renda

A abordagem pela renda fundamenta-se no valor presente de todos os fluxos de caixa futuros gerados pelo ativo ao longo de um período, conforme expresso na expressão abaixo, de modelo de fluxo de caixa descontado, destacado por (FLORENCIO & ALENCAR, 2016).

$$V_0 = \sum_{t=1}^n \frac{FC_t}{(1+r)^1} + \frac{FC_2}{(1+r)^2} + \dots + \frac{FC_n}{(1+r)^n}$$

onde:

V_0 = valor do ativo na data zero

FC_t = Fluxo de caixa esperado no final do ano t

r = retorno exigido apropriado (taxa de desconto)

n = período relevante

Tendo em vista que muitos dos conceitos e métodos de análise que são advindos de fluxos de caixa no mercado imobiliário são provenientes de adaptações do mercado financeiro, são necessárias adaptações em face às características particulares do setor da construção civil, no que se refere aos ciclos de análise, depreciação do ativo, procedimentos operacionais, entre outros (AMATO, 2001). Por esse motivo, o método de avaliação com base no valor da oportunidade de investimento (VOI) busca endereçar essas particularidades, configurando uma quarta abordagem.

O conceito do valor da oportunidade de investimento (VOI), segundo (LIMA JR., 1994) consiste como uma abordagem que reflete a disposição de um investidor em aceitar um determinado investimento, considerando sua capacidade de retorno. Nessa perspectiva, o avaliador deve interpretar a postura do investidor, calculando o valor que seria justificável investir ou pagar pelo empreendimento, sob a expectativa de que a operação, se conduzida conforme as condições estabelecidas, resulte em um fluxo de retorno capaz de remunerar o investimento a uma taxa de retorno satisfatória.

Para estimar o VOI, é necessário elaborar fluxos de caixa descontados, projetando os benefícios futuros que podem ser alcançados na operação do empreendimento sob certas

condições, conforme a operação a seguir, extraída de (FLORENCIO & ALENCAR, 2016):

$$VOI_0 = \frac{VOI_n}{(1 + TATco)^{ciclop}} + \sum_{k=1}^{ciclop} \frac{RODi_k}{(1 + TATco)^k}$$

Onde:

- n é extensão do ciclo operacional;
- k é um contador de tempo;
- TATco é a taxa de atratividade arbitrada para o ciclo operacional;
- RODik é o Resultado Operacional Disponível da Operação, a cada período k do ciclo operacional;
- VOIn é o Valor da Oportunidade de Investimento que um investidor universal avesso ao risco pagaria pelo empreendimento, ao final do ciclo operacional, calculado com a expressão extraída de (FLORENCIO & ALENCAR, 2016):

$$VOI_n = \alpha \sum_{k=1}^{exaustão} \frac{RODi_k}{(1 + TATex)^k}$$

Onde:

- Alfa é um fator de ajuste para calibre da dimensão dos investimentos necessários para manutenção da inserção de mercado do empreendimento, a fim de preservar o recebimento de resultados conforme previsto no início do ciclo operacional;
- TATex é a taxa de atratividade arbitrada para o período de exaustão;

Conforme MONETTI (1996) e AMATO (2001) apud (FLORENCIO & ALENCAR, 2016), os Empreendimentos de Base Imobiliária (EBIs) apresentam uma dinâmica de operações de longo prazo, caracterizadas por ciclos orgânicos e temporais que envolvem transações financeiras específicas. Diante dos riscos futuros, a metodologia de avaliação baseada no Valor da Oportunidade de Investimento (VOI) emerge como uma abordagem mais coerente, exigindo uma análise cuidadosa baseada na natureza do ativo, envolvendo uma série de etapas para seu cálculo.

Para chegar ao cálculo do VOI faz-se necessário: (1) a definição de cenários futuros de desempenho, (2) o estabelecimento do horizonte do ciclo operacional, (3) a determinação

de uma taxa de atratividade coerente e (4) a estimativa dos investimentos necessários para a reciclagem do empreendimento ao final do ciclo. Além disso, é crucial arbitrar limites de proteção diante dos riscos de desempenho, considerando distúrbios sistêmicos e conturbações nos ambientes econômico e competitivo (FLORENCIO & ALENCAR, 2016), como a taxa de desconto (tad), utilizada na abordagem VOI e ajustada de acordo com o nível desejado de proteção ao investimento na propriedade.

Adicionalmente aos modelos exemplificados acima, outra abordagem relevante e de grande uso no mercado imobiliário são os modelos estatísticos, também chamados de modelos hedônicos, que utilizam técnicas estatísticas de regressão múltipla para analisar um conjunto de variáveis, consideradas independentes, que influenciam o valor de um imóvel (eg. atributos físicos, de localização, infraestrutura, serviços próximos e outros fatores relevantes) (BAUR, ROSENFELDER, & LUTZ, 2023). Enquanto as metodologias clássicas de valoração são mais tradicionais e baseadas em critérios específicos, os modelos hedônicos conseguem considerar uma ampla gama de variáveis para estimar o valor de um imóvel.

Num modelo hedônico, a importância das variáveis independentes para a determinação da taxa de rentabilidade pode variar conforme a localização do ativo ou seu preço de mercado. No entanto, é essencial garantir que a determinação da importância dessas variáveis seja baseada em relações verdadeiramente significantes, evitando conclusões baseadas em eventos pouco relevantes, fortuitos ou aleatórios. Os modelos de ML têm capacidade, se treinados a partir dos dados disponíveis, de acessar uma compreensão mais profunda das relações não-lineares entre as variáveis, contribuindo para uma avaliação mais precisa da rentabilidade dos ativos EBI (VIANA, VAZ, ALVES, & TORNERI PORTO, 2017).

A fim de ampliar a compreensão sobre o universo que abrange as funcionalidades do ML e visando atender o objetivo secundário deste trabalho, a segunda parte desta seção será focada em conceituar os principais fundamentos e algoritmos a partir da condução de um levantamento abrangente de algoritmos de ML.

2.2.3 INTELIGENCIA ARTIFICIAL (IA)

Desde meados do século passado a IA tem sido impulsionada pelo crescimento da quantidade de dados disponíveis e sua definição é abordada por diversos autores sob perspectivas variadas. Neste trabalho, será adotada a visão de RUSSEL e NORVING (1995), destacando o objetivo conceitual da IA em permitir que as máquinas entendam, aprendam e completem atividades humanas ou, complementarmente, citando John McCarthy, um dos fundadores da IA, "Fazer com que uma máquina se comporte de maneira inteligente, como se um humano estivesse se comportando dessa forma"⁴.

A fim de aprofundar os conceitos de inteligência que compõe a Inteligência Artificial e, posteriormente conectar os conceitos de IA e de ML, será aplicada a subdivisão no campo de estudo de IA, em: (a) raciocínio, (b) aprendizado, (c) reconhecimento de padrões e (d) inferência, que serão especificadas a seguir. Compreender esses conceitos facilitará a transição para a compreensão de algoritmos usados na otimização de processos em *Real Estate*, explicados na seção 2.2.2 de *Machine Learning*.

a) Raciocínio:

Aldo e Gresse Von Wangenheim (2003) descrevem o Raciocínio Baseado em Casos (RBC) como uma abordagem que utiliza experiências passadas para resolver problemas atuais, aplicando soluções já conhecidas, evitando a criação de novas soluções do zero. Conforme a **Figura 3**, um sistema de RBC envolve (1) representação do conhecimento, que, em RBC, se dá principalmente em forma de casos que descrevem experiências concretas, (2) medida de similaridade, a capacidade de encontrar um caso relevante na base de casos existentes para responder a pergunta de um problema atual através de (3) adaptação, que foca em mecanismos para adaptar os casos recuperados completamente apesar de não serem idênticos ao casos atuais.

⁴ Fonte: <http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html>

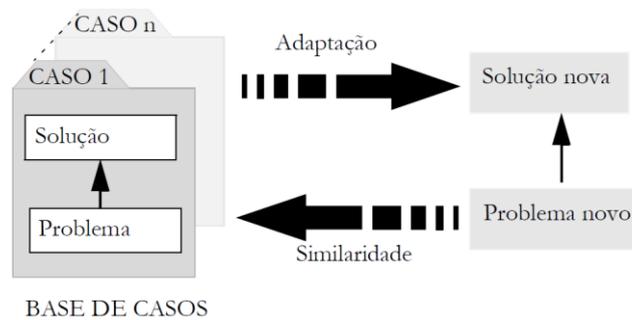


Figura 3 Modelo Básico do Enfoque de CBR – Retirado de WANGENHEIM, 2003

Segundo Aldo e Von Wangenheim (2003), no Ciclo de RBC, proposto por Aamondt e Plaza (1994), tais elementos são resumidos no ciclo dos 4 R's, que são as quatro etapas de execução do RBC: Recuperação, Reuso, Revisão e Retenção, ilustradas na **Figura 4**. Essas etapas mimetizam a forma de solução de problemas da mente humana e são a base para o endereçamento de funções dos algoritmos de ML.

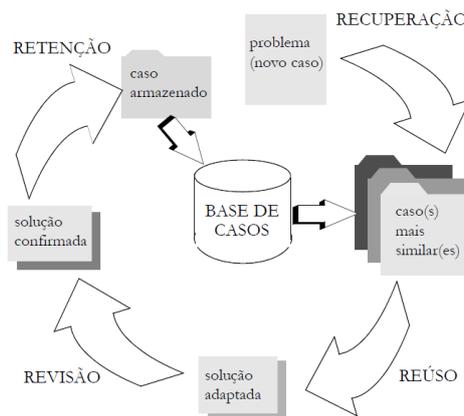


Figura 4 Ciclo de Funcionamento de sistema RBC – Retirado de WANGENHEIM, 2003

b) Aprendizagem:

Dentre as formas de aprendizado que envolvem o ML, são destacados: (1) aprendizado supervisionado, em que um computador, a partir de rótulos fornecidos por humanos, prevê a raça de um cachorro com base em imagens rotuladas de cachorros; (2) aprendizado não supervisionado, que não requer rótulos, às vezes criando suas próprias tarefas de previsão, como tentar prever cada palavra sucessiva em uma frase; e (3) o aprendizado por reforço permite que um agente aprenda seqüências de ações que otimizam suas recompensas totais, como vencer jogos, sem

exemplos explícitos de técnicas boas, possibilitando a autonomia. (MANNING, Artificial Intelligence Definitions, 2020).

c) Reconhecimento de Padrões:

Bishop (2006) define o campo do reconhecimento de padrões como a descoberta automática de regularidades em dados por meio de algoritmos computacionais. Dentro desse processo, ele identifica três fases essenciais que desempenham um papel fundamental na capacidade de um algoritmo em reconhecer padrões e em lidar com diferentes classes de dados de forma eficaz: pré-processamento, treinamento e generalização.

I. Fase de Pré-processamento:

A primeira etapa, conhecida como Fase de Pré-processamento, envolve a transformação das variáveis de entrada originais em um novo espaço de variáveis. Isso é feito com a expectativa de que o problema de reconhecimento de padrões se torne mais fácil de resolver nesse novo espaço. Frequentemente, essa etapa é chamada de extração de características, e pode ser usada para acelerar a computação.

II. Fase de Treinamento / Aprendizado:

Na Fase de Treinamento, um conjunto de dados é utilizado para ajustar os parâmetros de um modelo adaptativo. As categorias dos elementos no conjunto de treinamento são conhecidas com antecedência, geralmente por meio de inspeção e rotulagem manual. Após o treinamento, o modelo pode identificar a categoria de novas amostras que compõem um conjunto de teste. Nessa fase, o conhecimento sobre o domínio a ser trabalhado como, por exemplo, quais qualidades podem ser relevantes para serem rotuladas num processo de valoração de um imóvel, é fundamental.

III. Generalização:

A capacidade de categorizar corretamente novos exemplos que diferem dos utilizados no treinamento é conhecida como generalização. Em aplicações práticas, a variabilidade nos vetores de entrada significa que os dados de treinamento representam apenas uma pequena fração de todos os vetores de entrada possíveis. Portanto, a generalização é um objetivo central no reconhecimento de padrões, permitindo que os modelos se adaptem a uma ampla gama de cenários. Sendo assim, mesmo que o novo problema a ser resolvido não seja exatamente igual ao que existe na base de dados, é possível desenhar um limite de intervalo em que serão reconhecidas as similaridades a partir da generalização de certos aspectos feito com base em inferência, descrita a seguir.

d) Inferência:

A Inferência desempenha um papel de destaque na área de IA sendo um mecanismo fundamental para a tomada de decisões e resolução de problemas complexos. Conforme definido por (ROSS, 2010), o processo de inferência consiste em inferir novos fatos a partir de fatos estabelecidos, ou seja, a inferência envolve a busca por soluções de problemas e a aplicação de ações com base em conhecimento preexistente. Esse processo é cíclico e compreende as etapas de avaliação, onde são selecionadas regras a serem ativadas, e execução, que implica na modificação da base de fatos e regras. A compreensão dessas etapas é essencial para a compreensão do funcionamento da IA na resolução de problemas.

Uma das facetas cruciais da inferência, conforme ressaltada por (BISHOP, 2006, p. 17), é a determinação da probabilidade posterior, calculada a partir de um conjunto de dados de treinamento. Isso é um exemplo prático de inferência e, muitas vezes, um problema bastante desafiador. O problema geral de inferência envolve a determinação da distribuição conjunta de um universo ou, equivalentemente, uma amostra dentro desse universo, fornecendo assim uma descrição probabilística completa da situação em questão.

Nas aplicações práticas, existe frequentemente a necessidade de fazer previsões específicas para um valor "x" ou de tomar ações com base na compreensão dos valores

prováveis de "x". Esse aspecto constitui o cerne da teoria da decisão. Nesse contexto, o problema de classificação é dividido em duas etapas distintas: a etapa de inferência, em que são utilizados dados de treinamento para treinar um modelo de probabilidade, e a etapa subsequente de decisão, em que essas probabilidades posteriores são empregadas para realizar atribuições. É importante ressaltar que existe uma alternativa, que consiste em resolver ambos os problemas de forma integrada, treinando uma função que mapeia diretamente as entradas "x" em decisões, denominada função discriminante.

Em resumo, as abordagens no processo de inferência em IA oferecem opções distintas para a resolução de problemas de decisão. A escolha entre elas depende das necessidades específicas da aplicação e dos recursos disponíveis.

2.2.4 MACHINE LEARNING (ML)

Machine Learning (ML) é a parte da Inteligência Artificial voltada para como computadores podem melhorar sua percepção, conhecimento ou ações com base em experiência ou dados (MANNING, Artificial Intelligence Definitions, 2020). Para isso, o ML se baseia em ciência da computação, estatística, psicologia, neurociência, economia e teoria de controle, um ramo da engenharia que lida com o uso de respostas *ou feedbacks* de um sistema para influenciar o seu comportamento a fim de se autorregular e atingir um objetivo desejado, diminuindo erros (RUSSEL & NORVING, 1995).

Conforme abordado na seção 2.2.3, existem diversas formas de adquirir aprendizado através de dados e, como ponto de partida, é necessário que sejam incorporados rótulos que identificam seu conteúdo para criar uma correlação que dará origem à informação a ser analisada no produto final esperado. Essa rotulação é a referência que possibilita o aprendizado da máquina e pode ser feita de três maneiras⁵: (1) Aprendizado Supervisionado, quando há uma rotulagem prévia realizada por humanos, (2) Aprendizado Não-Supervisionado, quando a máquina é incumbida de identificar e categorizar os dados por conta própria durante o processamento, ou (3) Aprendizado por

⁵ Fonte: <https://developer.ibm.com/articles/cc-models-machine-learning/> (Acessado em: Fevereiro 2024)

Reforço, no qual um sistema é incentivado por meio de recompensas ao alcançar resultados corretos e/ou recebe penalizações diante de insucessos.

Como será exemplificado nesta seção, a aplicação desses tipos de aprendizado se dá a partir da modelagem de algoritmos, que são a base do ML. Um algoritmo lista os passos precisos a serem tomados num conjunto finito de instruções executadas em uma ordem específica para realizar uma tarefa específica (MANNING, 2020).

Retomando a seção 2.1 Contexto de Aplicação de *Machine Learning* no Mercado Imobiliário, o nível de transparência dos algoritmos de ML é um fator importante para que o setor imobiliário absorva tecnologias como o ML. Segundo as entrevistas realizadas por CONWAY (2018), profissionais do mercado imobiliário podem demonstrar relutância em adotar ou confiar na tecnologia, sendo necessário apresentar não só resultados concretos com evidências de retorno sobre investimento para conquistar uma parcela do mercado, como também clareza nos meios em que as decisões foram tomadas.

Sendo assim, a fim de contribuir com técnicas que permitam que o ser humano entenda e, conseqüentemente, confie para efetivamente gerenciar modelos de IA de alta performance com responsabilidade, a “*Explainable Artificial Intelligence*” (XAI) é um campo de estudo que se encaixa como um facilitador no processo de avanços das tecnologias que envolvem IA e ML (GUNNING, 2017).

Qualquer abordagem que busque reduzir a complexidade de um modelo ou simplificar o entendimento de seus resultados pode ser considerado XAI (ARRIETA & et al., 2020). É possível sintetizar os objetivos da XAI em nove conceitos resumidos na **Tabela 1**, que abrangem as principais qualidades de um "modelo explicável". Um modelo explicável compreende as técnicas utilizadas para transformar um modelo não interpretável em um modelo explicável. A explicabilidade pode ser vista como uma característica ativa de um modelo, denotando qualquer ação ou procedimento adotado por um modelo com a intenção de esclarecer ou detalhar suas funções internas.

Objetivos do XAI	Descrição
Credibilidade "Trustworthiness"	Credibilidade se refere ao grau de confiança de que um modelo reagirá conforme o esperado ao enfrentar um problema específico.
Causalidade "Causality"	Causalidade entre variáveis de dados significa encontrar relações de causa e efeito que contribuem para uma compreensão mais completa do modelo.
Transferibilidade "Transferability"	A transferibilidade lida com a descoberta das limitações de modelos para avaliar melhor sua aplicabilidade em outros casos.
Informatividade "Informativeness"	Informatividade está relacionada à distinção entre o problema original de tomada de decisão humana e o problema resolvido por um determinado modelo, incluindo seus mecanismos internos.
Estabilidade "Confidence"	Estabilidade descreve a robustez de um modelo, incluindo a confiança no seu regime de funcionamento
Justiça "Fairness"	Igualdade tenta proibir o uso injusto ou antiético dos resultados e saídas de um modelo por meio de análise ética e esclarecimento dos resultados que afetam as relações.
Acessibilidade "Accessibility"	Acessibilidade refere-se ao envolvimento de usuários finais (não técnicos) no processo de modelagem de IA.
Interatividade "Interactivity"	A interatividade lida com o nível de interação entre os usuários finais e os modelos de XAI para melhorá-los.
Conscientização sobre Privacidade "Privacy Awareness"	A Conscientização sobre Privacidade trata de esclarecer possíveis violações de privacidade, informando os usuários.

Tabela 1 Objetivos da XAI – Baseado em ARRIETA et al., 2019

Tendo como base a metodologia de Arrieta (2020), é possível dividir os modelos em Transparentes e Modelos que precisam de técnicas auxiliares para que atinjam a transparência (Modelos Post-hoc). Modelos transparentes devem satisfazer três níveis essenciais: simulabilidade, fragmentação ou análise das partes, e transparência algorítmica. Algoritmos que não atingem esses níveis devem recorrer a técnicas como Explicações de Texto, Explicações Visuais, Explicações por Segmentos, Explicações por Exemplos/Estudos de Caso e Explicações por Simplificação, bem como Explicações de Relevância. Para uma compreensão mais detalhada desses conceitos, pode-se consultar a **Tabela 2**.

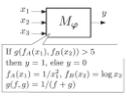
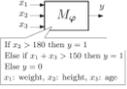
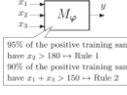
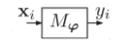
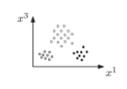
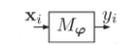
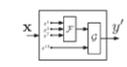
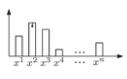
Base: Modelo "Black-Box"	"A natureza de caixa-preta" refere-se à falta de explicabilidade e interpretabilidade dos sistemas baseados em IA.	$x \rightarrow M_{\varphi} \rightarrow y$ $x = (x^1, \dots, x^n)$ $x_i: \text{input}$
Níveis de Transparência em Modelos de ML		
Simulatabilidade "Simulatability"	Refere-se à capacidade de um modelo ser simulado ou pensado estritamente por um ser humano, enfatizando a complexidade. Modelos simples, mas extensos, como sistemas baseados em regras com um grande número de regras, não se encaixam nessa característica. Para um modelo extenso ter simulatabilidade, ele deve ser autocontido o suficiente para que um humano possa pensá-lo como um todo.	
Fragmentação e Análise das Partes "Decomposability"	Refere-se à capacidade de explicar cada parte de um modelo (entrada, parâmetro e cálculo), relacionando-se com a inteligibilidade. Para ser fragmentado, cada entrada deve ser prontamente interpretável, e todas as partes do modelo devem ser compreensíveis sem a necessidade de ferramentas adicionais.	
Transparência Algorítmica "Algorithmic Transparency"	Lida com a capacidade do usuário de entender o processo seguido pelo modelo para produzir qualquer resultado a partir de seus dados de entrada. Modelos algorítmicos transparentes devem ser totalmente exploráveis por meio de análise matemática, como um modelo linear, por exemplo, que permite a compreensão do usuário sobre como o modelo se comportará em diferentes situações.	
Técnicas Auxiliares para Transparência de Modelos Post-hoc de ML		
Texto "Text Explanations"	Trata da geração de explicações em texto para tornar um modelo explicável. Essas explicações podem incluir símbolos que representam o funcionamento do modelo.	
Visualização "Visual Explanations"	Técnicas visuais para explicabilidade pós-hoc visam visualizar o comportamento do modelo, muitas vezes com redução de dimensionalidade para simplificar a visualização. As visualizações podem ser combinadas com outras técnicas para melhorar a compreensão.	
Por Segmentação "Local Explanations"	Abordam a explicabilidade segmentando o espaço de solução e fornecendo explicações para subespaços de solução menos complexos relevantes para o modelo como um todo.	
Por Estudos de Caso, Exemplos "Explanations by example"	Consideram a extração de exemplos de dados relacionados ao resultado gerado por um modelo, para melhor compreensão do próprio modelo.	"Explanatory examples for the model:" - $x_A \rightarrow y_A$ - $x_B \rightarrow y_B$ - $x_C \rightarrow y_C$
Por Simplificação "Explanations by simplification"	Envolve a reconstrução de um novo sistema baseado no modelo treinado para ser explicado, otimizando a semelhança com o modelo original, mas com menor complexidade.	
Por Relevância "Relevance Explanation"	Métodos de explicação de relevância calculam uma pontuação de relevância para as variáveis gerenciadas por um modelo, revelando a importância de cada variável na produção de sua saída. Esses métodos podem ser considerados como uma forma indireta de explicar um modelo.	

Tabela 2 Conceitos para transparência – Baseado em ARRIETA et al., 2019

Mediante a revisão bibliográfica, abarcando aproximadamente 400 contribuições na literatura sobre XAI e conceitos correlatos publicados até o momento, ARRIETA (2019) elaborou um inventário dos principais algoritmos de ML atualmente conhecidos. Com base na distinção entre modelos transparentes e modelos "post-hoc", classificou esses algoritmos de acordo com seus graus de transparência, conforme evidenciado na **Tabela 3**.

Modelo	Modelos Transparentes			Técnicas Post-hoc
	Simulatabilidade "Simulatability"	Fragmentação e Análise das Partes "Decomposability"	Transparência Algorítmica "Algorithmic Transparency"	
LLR "Logistic/ Linear regression"	Sim	Sim	Sim	NA
DTs "Decision Trees"	Sim	Sim	Sim	NA
KNN "K-nearest Neighbors"	Sim	Sim	Sim	NA
RBML "Rule-Base Learners"	Sim	Sim	Sim	NA
GAM "General Additive Models"	Sim	Sim	Sim	NA
BM "Bayesian Models"	Sim	Sim	Sim	NA
RF "Tree ensembles, random forests and multiple classifier systems"	NA	NA	NA	- Por simplificação - Por relevância
SVM "Support Vector Machines"	NA	NA	NA	- Por simplificação - Por segmentação
MLNN "Multi-Layer Neural Networks"	NA	NA	NA	- Por simplificação - Por relevância - Visualização
CNN "Convolutional Neural Networks"	NA	NA	NA	- Por relevância - Visualização
RNN "Recurrent Neural Networks"	NA	NA	NA	- Por relevância

Tabela 3 Classificação de modelos de ML a partir do seu nível de explicabilidade – Baseado em ARRIETA et al., 2019

Tendo como base a listagem de algoritmos selecionado por ARRIETA (2019) como os mais relevantes até o momento, tais modelos serão detalhados abaixo a fim de atingir um

dos objetivos secundários desse trabalho: conduzir um levantamento abrangente de algoritmos de ML aplicados à arbitragem de valor para transações imobiliárias de ativos residenciais, tornando acessível o entendimento de seu funcionamento ao público do mercado imobiliário.

Além de fornecer a base necessária para a leitura do capítulo 3 (Aplicações de Modelos de ML para Arbitragem de Valor em Empreendimentos Residenciais), o entendimento dos algoritmos listados a seguir também foi crucial para a preparação das entrevistas. Isso assegurou uma cobertura mais ampla dos temas abordados nas perguntas específicas e técnicas, além de proporcionar uma compreensão mais profunda do universo de ML, essencial para orientar as entrevistas detalhadas a partir do capítulo 4.

2.2.4.1 " Linear/ Logistic Regression" (LR) - Regressão Linear/ Logística

A regressão linear ou logística (LR) é um modelo de classificação utilizado para prever um valor de uma variável dependente (categoria) a partir do valor de outra variável. Este modelo é um tipo de aprendizado supervisionado que parte da premissa de dependência não-linear (curva “S”) ou linear, respectivamente, entre os atributos e as variáveis previstas, o que limita sua flexibilidade para se ajustar aos dados. Essa rigidez do modelo é o que o mantém como um método transparente (ARRIETA, 2020).

Abaixo, **Figura 5**, ilustra o funcionamento de uma função de regressão linear assim como exemplifica quais são suas qualidades que a classifica dentro dos parâmetros de transparência.

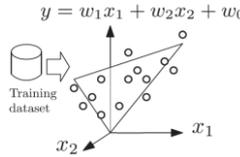
				
w_i : aumenta em y caso x_i aumente w_0 : interceptor, normalizado para testes de variação em y				
Modelo	Simulatabilidade "Simulability"	Fragmentação e Análise das Partes "Decomposability"	Transparência Algorítmica "Algorithmic Transparency"	Técnicas Post-hoc
LLR "Logistic/ Linear regression"	Os atributos são legíveis por humanos e as interações entre eles são mantidas ao mínimo.	As variáveis ainda são legíveis, mas o número de interações e preditores envolvidos nelas aumentou, forçando a fragmentação.	As variáveis e interações são muito complexas para serem analisadas sem ferramentas matemáticas.	NA

Figura 5 Regressão Linear - Parâmetros de transparência e ilustração - imagem retirada de ARRIETA

Importante ressaltar que, apesar da regressão linear estar entre os modelos mais simples de classificação, a fim de manter a capacidade da audiência em simular seu processo e de analisar suas partes, é necessário que o tamanho de sua formulação seja reduzido e que duas variáveis sejam inteligíveis pela audiência. Entende-se que, para aumentar sua complexidade sem perder a transparência, este modelo pode fazer uso de técnicas "post-hoc" que contribuam com soluções mais completas

2.2.4.2 "Decision Trees" (DT) - Árvore de Decisão

As Árvores de Decisão (DT) são estruturas hierárquicas utilizadas para suportar problemas de regressão e classificação. Elas têm sido amplamente utilizadas em contextos de suporte à decisão devido à sua transparência imediata (ARRIETA, 2020). Tais modelos representam uma sequência binária de seleções feitas ao longo de um eixo transversal de uma estrutura ramificada, como uma árvore. Sua flexibilidade é dependente tanto de quais e quantos são os parâmetros escolhidos para quebra do "ramo" quanto quais os intervalos que limitam cada parâmetro, ou seja, a definição dos "thresholds" ou limites que definem quando avançar para o próximo ramo. (BISHOP, 2017, p. 664).

Abaixo, **Figura 6**, ilustra o funcionamento de uma estrutura de árvore de decisão, assim como exemplifica quais são suas qualidades que a classifica dentro dos parâmetros de transparência.

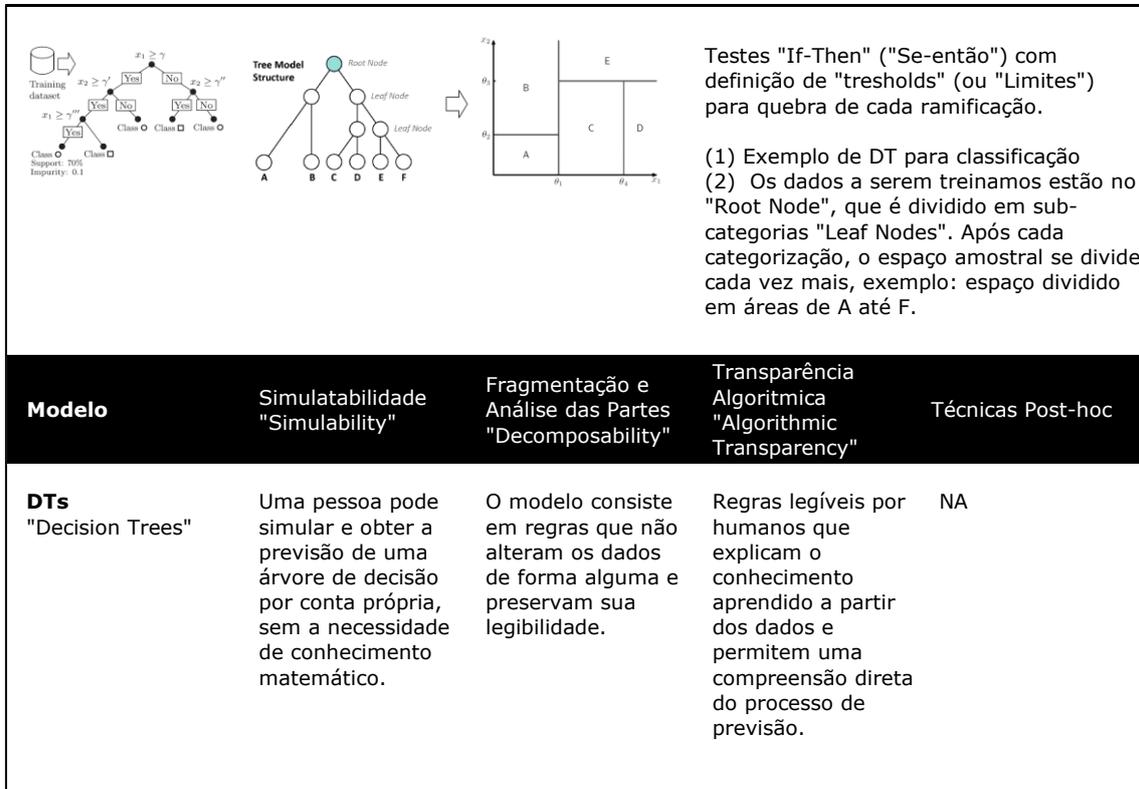


Figura 6 - Árvore de Decisão - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP

Um aspecto das árvores de decisão que pode ser limitador é que seu funcionamento é baseado em divisões rígidas através do uso de *thresholds* a cada "nó" da árvore. Sendo assim, o problema chamado "overfitting" causado na etapa de generalização (explicado em detalhes na seção 2.2.3 Inteligência Artificial), pode acontecer mais facilmente no DT.

Devido a suas propriedades de generalização serem menos eficientes em comparação com outros modelos, isso torna a DT um modelo menos interessante para aplicação em cenários onde o desempenho preditivo é um fator de extrema relevância (ARRIETA, 2023).

2.2.4.3 "K-Nearest Neighbors" (KNN)

O K-Nearest Neighbours (KNN) lida com problemas de classificação de uma maneira metodologicamente simples: ele prevê a classe de uma amostra de teste a partir das classes de seus K-vizinhos mais próximos (através da medida de distância entre amostras vizinhas ou densidade de vizinhos próximos) (ARRIETA & et al., 2020).

A estrutura desse modelo depende da noção de similaridade entre exemplos e pode ser adaptada para cada cenário específico. Tanto variáveis binárias quanto variáveis discretas podem assumir um de K estados possíveis, ou seja, é possível checar a resultados se assemelham à tomada de decisão humana baseada em experiência, que decide com base em casos similares ocorridos no passado.

Abaixo, **Figura 7**, ilustra o funcionamento do modelo KNN, assim como exemplifica quais são as qualidades que classificam esse algoritmo dentro dos parâmetros de transparência.

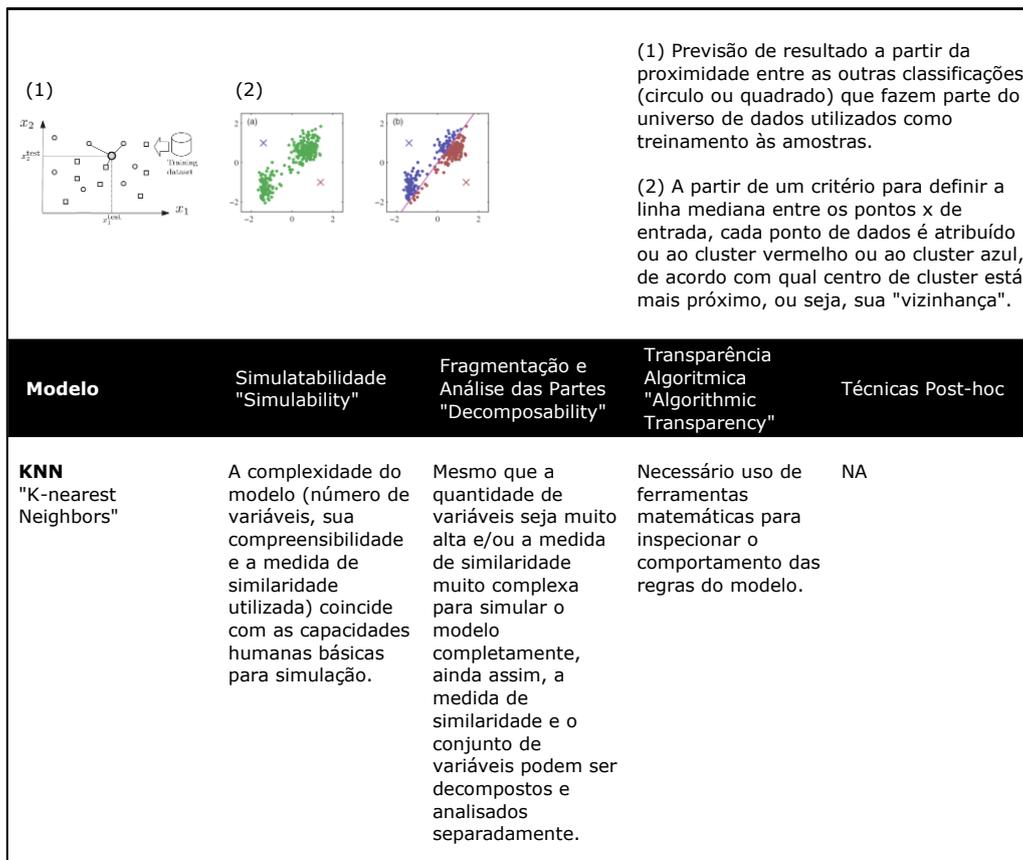


Figura 7 - KNN - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP

2.2.4.4 "Rule Based Learners" (RBL) - Aprendizado Baseado em Regras

O Aprendizado Baseado em Regras refere-se a qualquer modelo que faça uso de regras para caracterizar os dados a partir do "aprendizado" de uma base de dados de treinamento. Essas regras podem ser condicionais do tipo "if-then" (ou "se-então") ou combinações mais complexas, como as "fuzzy rules" (ou modelos difusos) (ARRIETA, 2017).

Os modelos difusos no aprendizado de máquina são um tipo de algoritmo baseado em regras que incorporam o conceito de lógica difusa, que permite a representação da incerteza e imprecisão nos dados e na tomada de decisões, o que pode ser particularmente útil em situações em que os dados não são binários ou definidos com clareza (ROSS, 2010).

Abaixo, **Figura 8**, ilustra o funcionamento do algoritmo RBL, com regras "if-then" e "fuzzy", assim como exemplifica quais são as qualidades que classificam esse algoritmo dentro dos parâmetros de transparência.

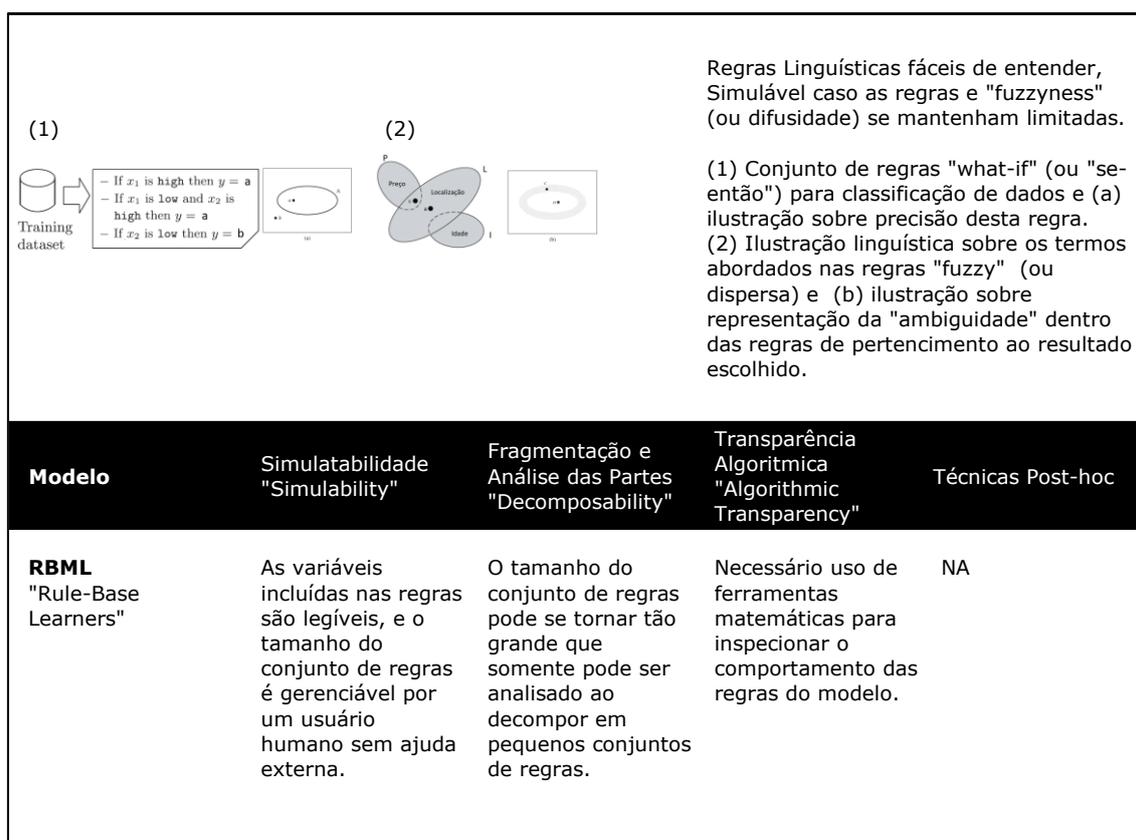


Figura 8 - RBL - Parâmetros de transparência e ilustração - adaptado de ARRIETA e ROSS

Os RBLs são excelentes modelos em termos de interpretabilidade em diversas áreas, devido à sua relação similar ao comportamento humano. Isso os torna adequados para entender e explicar outros modelos, mesmo para um usuário não especializado (ARRIETA,2010).

Todos os modelos são abstrações matemáticas do mundo físico real. Quanto mais suposições são necessárias para tornar o modelo compatível com estruturas matemáticas conhecidas que abordem o problema real, mais incerteza é introduzida no processo de modelagem. Ignorar essa incerteza equivale a ignorar o mundo real e nossa compreensão dele (ROSS, 2010) e os modelos "fuzzy" são uma maneira de embarcar esta complexidade.

2.2.4.5 "General Additive Models" (GMA) - Modelos Aditivos Generalizados

Os *General Additive Models* (GMA) são modelos nos quais o valor da variável a ser prevista é obtido pela agregação de funções suaves desconhecidas definidas para as variáveis preditoras (ARRIETA, 2010), ou seja, a partir de derivadas contínuas em todos os pontos de seu domínio, este modelo ajuda para que não existam descontinuidades bruscas, saltos ou picos em seu "gráfico".

Abaixo, **Figura 9**, ilustra o funcionamento do modelo GMA, assim como exemplifica quais são suas qualidades que classificam esse algoritmo dentro dos parâmetros de transparência.

<p>Função $g(z)$ é fragmentada em funções lineares $f_i(x_i)$ separadas por um intervalo $w [w_1, \dots, w_n]$</p>				
Modelo	Simulatabilidade "Simulability"	Fragmentação e Análise das Partes "Decomposability"	Transparência Algorítmica "Algorithmic Transparency"	Técnicas Post-hoc
GAM "General Additive Models"	As variáveis e a interação entre elas, conforme as funções suaves envolvidas no modelo, devem estar dentro das capacidades humanas para entendimento.	O tamanho do conjunto de regras pode se tornar tão grande que somente pode ser analisado ao decompor em pequenos conjuntos de regras.	Necessário uso de ferramentas matemáticas para inspecionar o comportamento das regras do modelo.	NA

Figura 9 - GMA - Parâmetros de transparência e ilustração – adaptado de ARRIETA

Segundo Arrieta, como o objetivo desse modelo é inferir funções suaves cuja composição agregada se aproxime da variável prevista, seu resultado não atingirá a precisão em si, mas sim a necessidade de compreender o problema subjacente e a relação entre as variáveis envolvidas nos dados. Por isso, os GMAs têm sido aceitos em algumas comunidades como a escolha de modelagem padrão, apesar de seu comportamento reconhecido como menos eficaz em comparação com modelos mais complexos.

2.2.4.6 "Bayesian Models" (BM) - Modelos Bayesianos

"Bayesian Models" (BM) são representados como modelos de grafos probabilísticos, nos quais as conexões entre variáveis refletem suas dependências condicionais (ARRIETA, 2020). Uma das principais vantagens dos modelos Bayesianos é que eles evitam o problema de "overfitting", que representa a adaptação brusca para adaptar os parâmetros aos dados existentes.

Abaixo, **Figura 10**, ilustra o funcionamento do modelo Bayesiano, assim como exemplifica quais são as qualidades que classificam esse algoritmo dentro dos parâmetros de transparência.

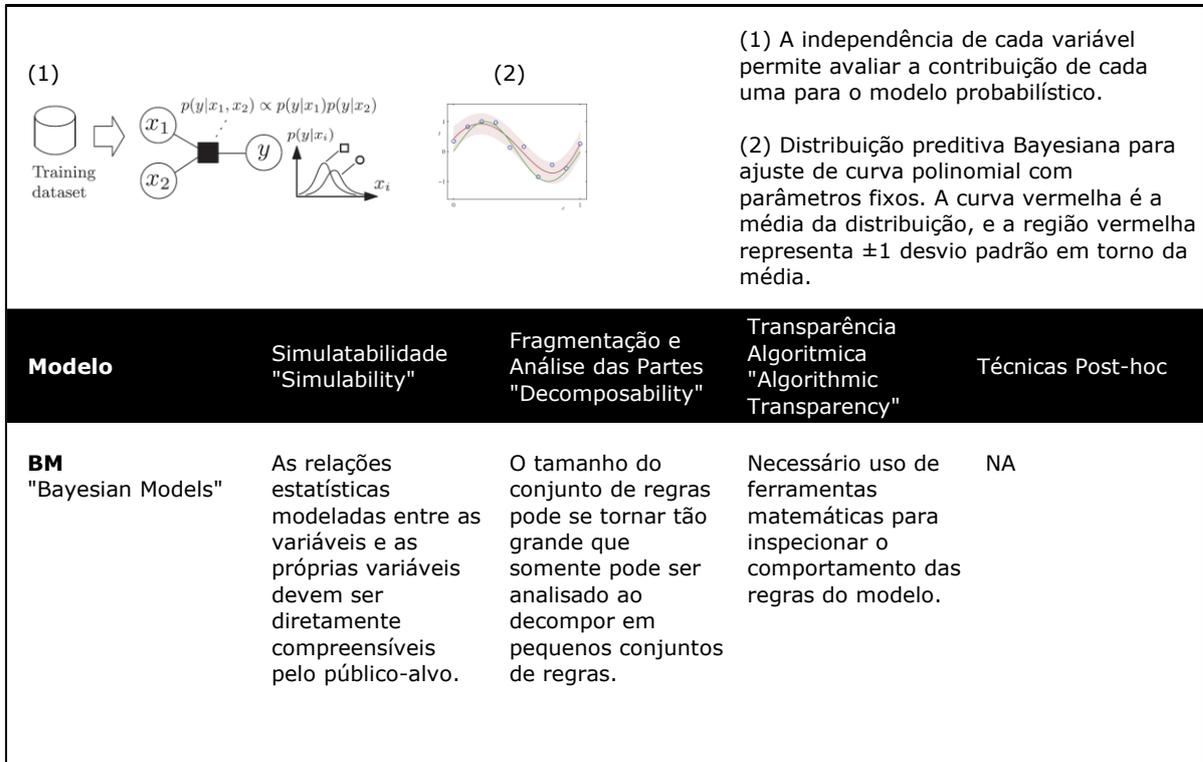


Figura 10 - Modelo Bayesiano - Parâmetros de transparência e ilustração - adaptado de ARRIETA e BISHOP

BMs permitem que o número de parâmetros seja maior do que o número de pontos de dados, adaptando-se automaticamente ao tamanho do conjunto de dados através do uso de probabilidade para representar a incerteza, o que é fundamental para fazer inferências racionais (BISHOP, 2017).

Uma crítica comum ao BM é que a escolha da distribuição que baseia o treinamento de seus dados, que pode ser influenciada por conveniência matemática ou de dados disponíveis. No entanto, avanços em métodos de amostragem tornaram os modelos Bayesianos práticos em diversas áreas de problemas (BISHOP, 2017).

2.2.4.7 Tree ensembles, random forest and multiple classifier systems" (TE, RF, MC)

Os "Multiple Classifier Systems" (MCS), em tradução livre Sistema de Múltiplos Classificadores, envolve a combinação de previsões de múltiplos modelos para melhorar o desempenho preditivo geral. Refere-se a uma abordagem que envolve o uso de vários classificadores individuais (modelos de ML) para realizar tarefas de classificação ou tomada de decisões, reduzindo o espaço de busca para a solução mais viável dentro do espaço amostral, conforme ilustrado na **Figura 11**.

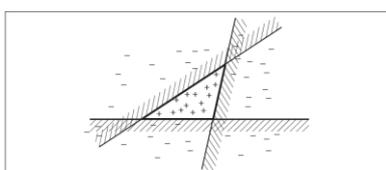


Ilustração do aumento de poder expressivo obtido pelo aprendizado de conjunto. A partir de três hipóteses de limite linear, cada uma das quais classifica como positivo no lado não sombreado, a região triangular é a hipótese resultante.

Figura 11 Ilustração sobre resultado da redução do espaço amostral - Extraído de RUSSEL (2020), p.749.

Os algoritmos de "Tree Ensemble" (TE), ou Agrupamento de Árvores, são um exemplo de MCSs e sua estratégia envolve a combinação de múltiplas hipóteses, ou árvores de decisão, para melhorar o desempenho preditivo e a capacidade de generalização do modelo (RUSSEL, 2010, p. 748), dentre elas, a "Random Forest" (RF), ou Floresta Aleatória, exemplificada mais adiante. A **Figura 12**, exemplifica os parâmetros de transparência desses modelos de aprendizagem:

Modelo	Simulatabilidade "Simulability"	Fragmentação e Análise das Partes "Decomposability"	Transparência Algorítmica "Algorithmic Transparency"	Técnicas Post-hoc
TE & RF "Tree ensembles, random forests and multiple classifier systems"	NA	NA	NA	- Por simplificação - Por relevância

Figura 12 Modelo Tree ensembles, random forest and multiple classifier systems - Parâmetros de transparência - adaptado de ARRIETA

Existem três principais tipos de agrupamentos usados com árvores de decisão: (A) bagging, (B) boosting e (C) stacking, descritos abaixo:

- (A) O “*bagging*” (ou "Agregação"), representado por métodos como o RF, cria várias árvores de decisão independentes ao treiná-las em subconjuntos aleatórios dos dados de treinamento, reduzindo a variância do modelo. As previsões do RF são obtidas ao fazer com que cada árvore vote na classificação ou valor de regressão para um determinado exemplo de teste, e a classe ou valor com mais votos é escolhido como a previsão final. Abaixo, **Figura 13** com ilustração explicativa.

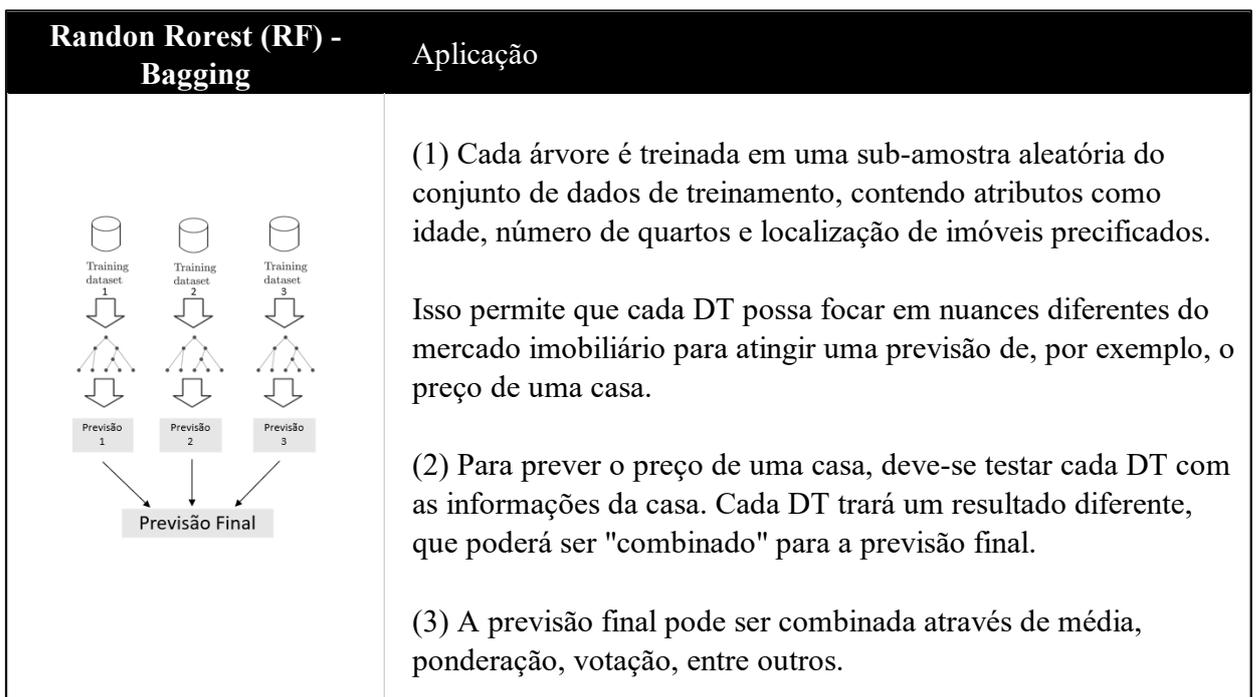


Figura 13 Random Forest - Ilustração e Aplicação - Produção original

- (B) Em contraste, o “*boosting*” (ou "Reforço"), exemplificado por algoritmos como o Gradient Boosting Machine (GBM), citado na seção 2.1, foca em treinar árvores de decisão sequencialmente, dando mais peso, a cada interação, aos exemplos mal classificados em iterações anteriores para melhorar a precisão global (LUCIC, 2019). Abaixo, **Figura 14** com ilustração explicativa.

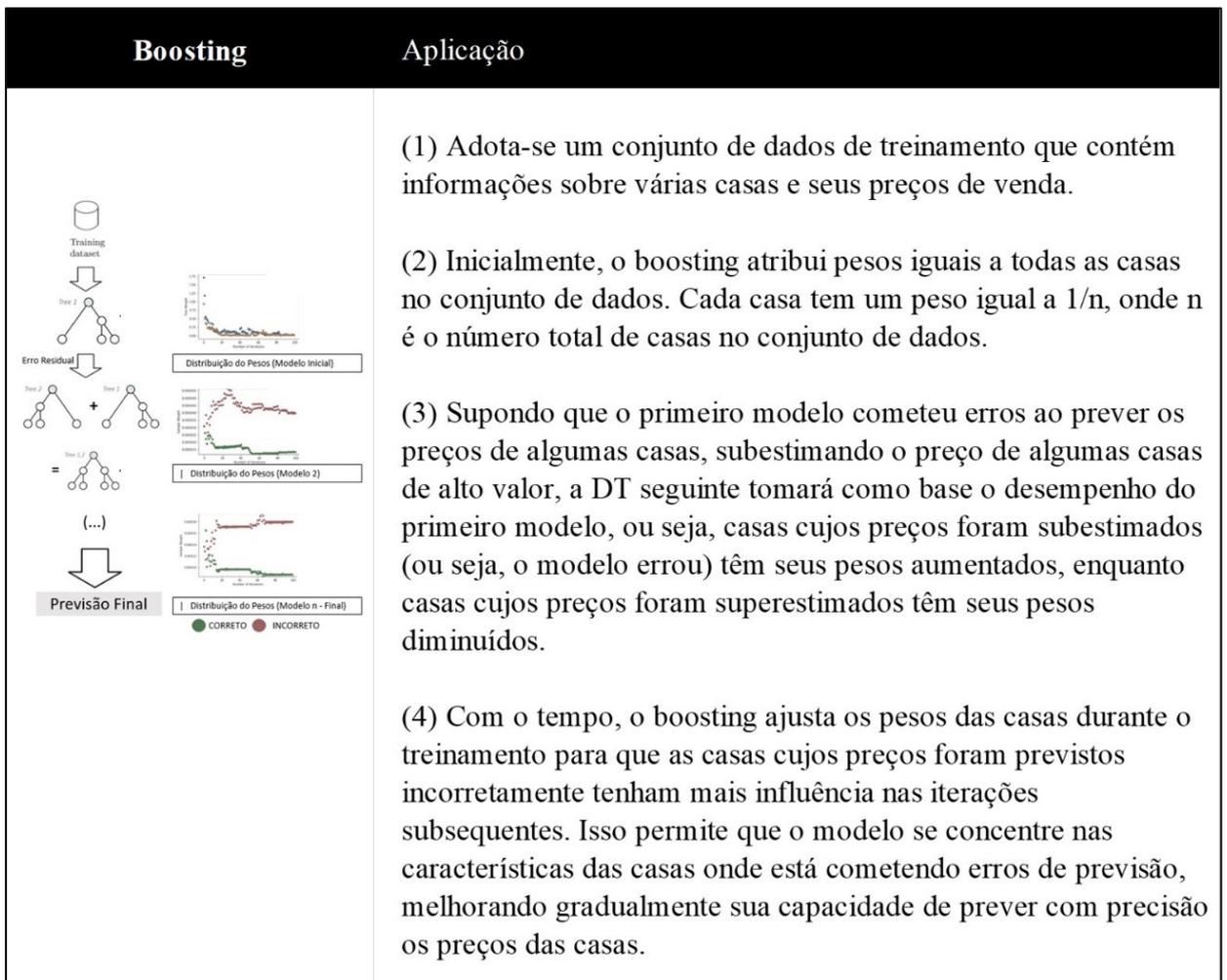


Figura 14 Bossting - - Ilustração e Aplicação – Adaptado de LUCIC

(C) Por fim, o “*stacking*” (ou “Empilhamento”) combina a saída de várias árvores (ou outros modelos) em um único modelo, permitindo que modelos mais simples se beneficiem da especialização de modelos mais complexos. Arrieta cita o “*Stacking with auxiliary features*” como exemplo, nesse método, você treina vários modelos base com os dados de treinamento e, em seguida, usa as previsões desses modelos como características adicionais (auxiliares) para treinar um meta-modelo que fará a previsão final. O objetivo é permitir que o meta-modelo aproveite as informações geradas pelos modelos base para melhorar as previsões.

Abaixo, **Figura 15** com ilustração explicativa.

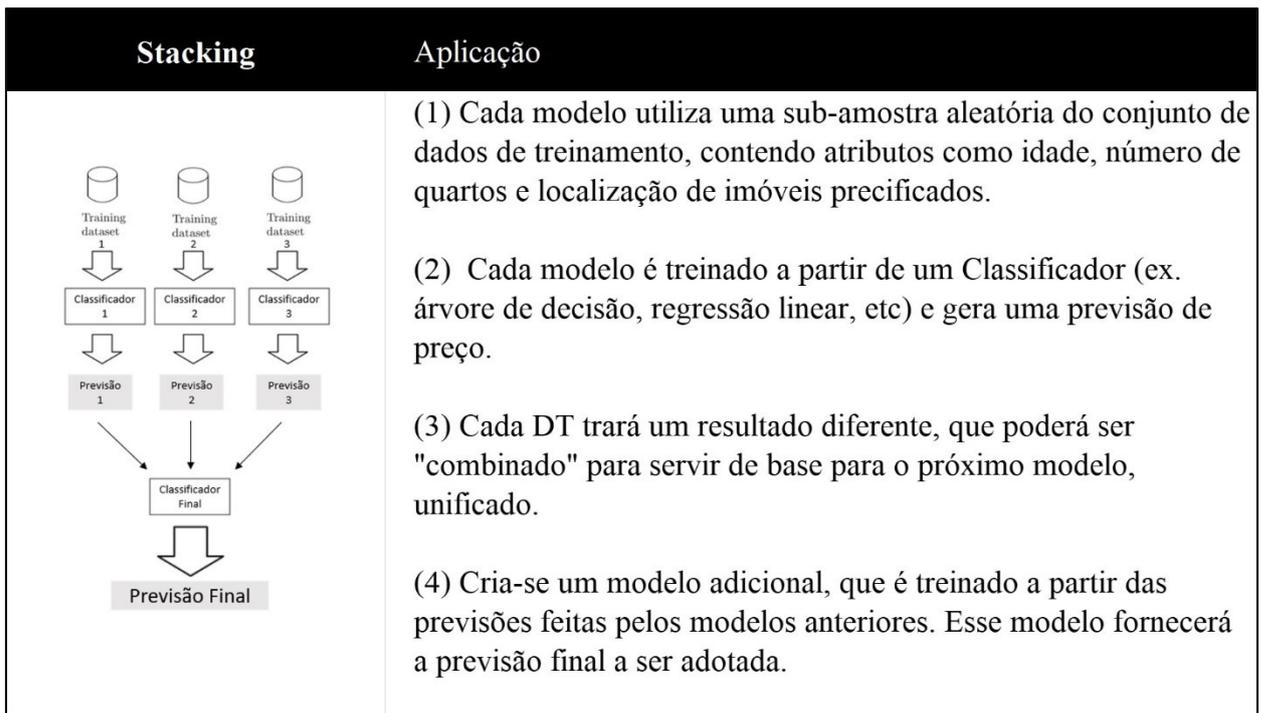


Figura 15 Stacking - Ilustração e Aplicação - ilustração explicativa adaptada de ARRIETA, CONWAY e ROSS

2.2.4.8 "Support Vector Machines" (SVM) - Máquinas de Vetores de Suporte

Os *Support Vector Machines* (SVMs) são uma classe de modelos de aprendizado supervisionado amplamente utilizada em tarefas de classificação, regressão e detecção de dados anômalos em relação a um conjunto de dados existentes (ARRIETA & et al., 2020). O conceito principal por trás das SVMs é identificar um hiperplano que maximize a margem entre os pontos de dados mais próximos de cada classe e dele próprio (CONWAY, 2018). Tais hiperplanos tornam mais simples e eficaz a classificação de dados, que são segregados a partir da separação geradas por esses hiperplanos.

Enquanto os MCs, descritos anteriormente, possuem um conceito que abrange todos os problemas que envolvem a classificação em mais de duas classes distintas, o SVM originalmente lida com classificação binária, mas pode ser estendido para problemas de classificação multiclasse.

Abaixo, a **Figura 16** com ilustração explicativa de SVM.

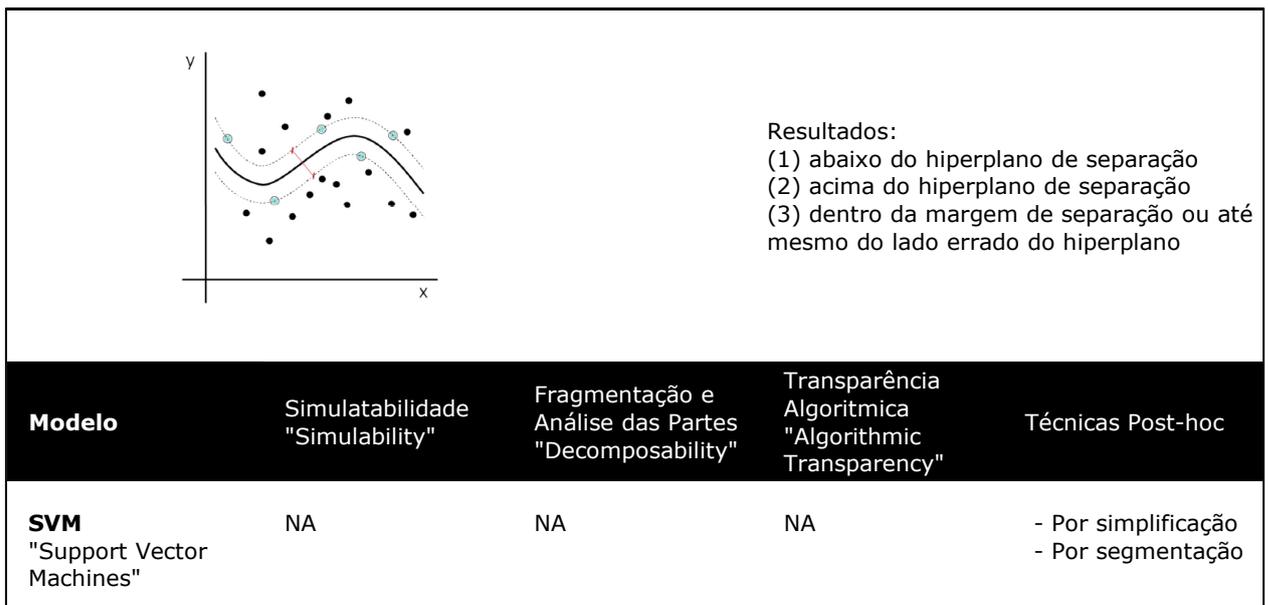


Figura 16 - Modelo SVM - Parâmetros de transparência, ilustração explicativa e exemplo de aplicação no mercado de valorção, em Real Estate - retirado de ARRIETA, BISHOP e CONWAY

Entendendo que os dados podem não ser completamente separáveis, é importante destacar que existe um grau de erro, ou seja, alguns objetos (eg. imóveis) podem estar dentro da margem de separação ou até mesmo do lado errado do hiperplano. Em alguns casos, os objetos podem não ser linearmente separáveis com base em suas características originais e as funções de kernel são utilizadas para mapear as características dentro de um espaço de maior dimensão, onde a separação linear seja possível (e.g. função polinomial de Kernel) (BISHOP, 2006).

No entanto, entendendo que há o risco de “*overfitting*”, ou superajuste, uma ferramenta utilizada é a introdução de “*slack variables*”, ou variáveis de folga. Cada ponto de dados tem sua própria variável de folga associada, utilizadas para formular restrições que permitem algum grau de erro na classificação e, assim, penalizam os pontos que estão do lado errado do hiperplano ou dentro da margem de separação (BISHOP, 2006).

Conforme apontado por BISHOP (2006), permitir que imóveis inadequados fiquem na margem de separação ou do lado errado acontece em benefício de algumas vantagens significativas para a qualidade do modelo como um todo, incluindo: uma tolerância controlada a erros, que torna o modelo mais flexível ao lidar com imóveis que não se encaixam perfeitamente em categorias estritas, e aumento sua robustez em relação a

ruídos e características atípicas nos dados do mercado imobiliário. Além disso, essa abordagem contribui para a redução do risco de “*overfitting*”, garantindo que o modelo não se ajuste excessivamente aos dados de treinamento, tornando-o mais apto a generalizar eficazmente para novos imóveis.

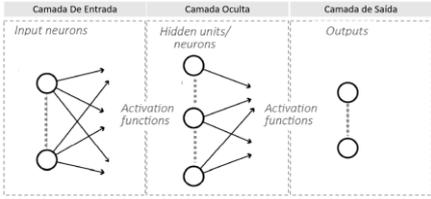
2.2.4.9 “*Neural Networks*” (NN) – Rede Neural

As *Neural Networks* (NNs) são modelos computacionais inspirados no funcionamento do cérebro humano (ROSS, 2010). Ela é descrita como uma série de transformações funcionais dos dados, onde variáveis de entrada são submetidas a “*Activation Functions*”, ou funções de ativação, resultando em unidades denominadas “*Hidden Units Neurons*”, ou unidades ocultas. Essas unidades ocultas são então combinadas para formar as saídas da rede.

A camada oculta consiste em uma série de operações matemáticas (e.g. Combinação ponderada de entradas, adicionar um valor de viés, função de ativação para introduzir não linearidades na rede etc.) para processar os dados de entrada e gerar saídas intermediárias (BISHOP, 2006).

A relação entre o aumento das camadas ocultas e o desempenho de uma MLN não é necessariamente direta e pode ser influenciada por vários fatores desde a necessidade de aumentar as camadas para melhorar a capacidade do algoritmo para atender a complexidade do problema até diminuir as camadas para reduzir o risco de *overfitting*. A definição da quantidade de camadas deve ser feita de forma experimental, avaliando o desempenho do modelo.

Abaixo, **Figura 17**, ilustra o funcionamento de um Neural Network assim como exemplifica quais são as qualidades que classificam esse algoritmo dentro dos parâmetros de transparência.



Modelo	Simulatabilidade "Simulability"	Fragmentação e Análise das Partes "Decomposability"	Transparência Algorítmica "Algorithmic Transparency"	Técnicas Post-hoc
MLNN "Multi-Layer Neural Networks"	NA	NA	NA	- Por simplificação - Por relevância - Visualização
RNN "Recurrent Neural Networks"	NA	NA	NA	- Por relevância
CNN "Convolutional Neural Networks"	NA	NA	NA	- Por relevância - Visualização

Figura 17 Modelo Neural Networks - Parâmetros de transparência, ilustração explicativa adaptada de CONWAY

Segundo Schmidhuber, em uma rede neural padrão, vários processadores simples e interconectados, chamados neurônios, geram sequências de ativações com valores reais. Neurônios de entrada são ativados por sensores que percebem o ambiente, enquanto outros neurônios se ativam por meio de conexões ponderadas com neurônios previamente ativos.

A complexidade desses comportamentos pode exigir longas cadeias de estágios de cálculo, onde cada estágio transforma a ativação agregada da rede de maneira não linear. Neste contexto, o “*Deep Learning*”, ou Aprendizado Profundo, busca contribuição de cada parte da rede para a saída final (ARRIETA & et al., 2020).

A escolha sobre as funções de ativação desempenha um papel importante na otimização do algoritmo e a experimentação é fundamental para determinar a mais adequada para cada contexto. Alguns exemplos possíveis no mercado de arbitragem de valor de imóveis seriam:

- A função de ativação linear, adequada para a camada de saída, permite prever os preços dos imóveis de forma contínua (Conway, 2023)
- A função ReLU (“*Rectified Linear Unit*“, ou Unidade de Retificação Linear), que é amplamente empregada nas camadas ocultas, permite a captura de relações não lineares entre as características dos imóveis e seus valores de mercado (Schmidhuber, 2023).
- A função *Leaky ReLU* (“*Leaky Rectified Linear Unit*“, ou Unidade de Retificação Linear “com vazamento”) é uma variação da função ReLU para ajudar a rede a capturar relações mais complexas entre as características dos imóveis, por exemplo, que podem ser beneficiadas pela modelagem de relações não lineares cujos outputs podem ser negativos.

Com a *Leaky ReLU*, mesmo quando os neurônios geram valores negativos, eles não se tornam completamente inativos, como acontece com a função ReLU. Em vez disso, eles produzem uma pequena contribuição, que multiplica o dado de saída por um fator constante chamado “*leakage rate*”, ou taxa de vazamento. Isso impede que esses neurônios se tornem inativos e permite que eles continuem sendo treinados (GOODFELLOW, BENGIO, & COURVILLE, 2016).

No contexto de Real Estate, isso pode ajudar a lidar com características que têm relações negativas com o preço do imóvel, como a distância a locais indesejados. A *Leaky ReLU* permite que a rede capture essas relações negativas, onde a localização desfavorável pode diminuir o preço, sem fazer com que os neurônios se tornem completamente inativos.

- As funções Sigmoides ou “Tanh” (tangente hiperbólica) podem ser utilizadas em camadas ocultas e são aplicadas para normalizar características e que as características tenham um efeito balanceado no modelo, independentemente de suas escalas originais. Essas funções permitem que as ativações permaneçam em intervalos específicos, como entre 0 e 1 ou -1 e 1, por exemplo (BISHOP, 2006).

Em seu estudo, Arrieta (2020) comenta sobre a diversidade de arquiteturas de redes neurais, destacando como as principais: MLN, as “*Convolutional Networks*” (CN), ou redes convolucionais, e “*Recurrent Networks*” (RN), ou redes recorrentes. Suas

características e aplicações dentro da área de Valoração, em *Real Estate*, serão descritas a seguir:

2.2.4.9.1 “Multi-Layer Networks” (MLN) – Redes de Múltiplas Camadas

Essas redes consistem em várias camadas de “neurônios interconectados” em que cada camada processa informações e passa os resultados para a próxima camada. São amplamente utilizadas em tarefas de aprendizado supervisionado, como classificação e regressão e funcionam bem para dados estruturados e não sequenciais (EVERITT & HUTTER, 2018).

A fim de tornar o entendimento sobre MLN mais ilustrativo, a rotina abaixo, para ativação e treinamento da rede neural de Bishop (2006), foi somada a aplicação do algoritmo no campo de arbitragem de valor de imóveis:

- (1) Coleta de Dados, Pré-Processamento e Divisão de dados: Após a seleção de um conjunto de dados que contenha informações sobre propriedades imobiliárias, como tamanho, localização, número de quartos, banheiros, preço de venda histórico e outros fatores relevantes, é iniciada a etapa de limpeza e normalização dos dados que, posteriormente, serão separados em conjuntos de treinamento a fim de testar e avaliar o desempenho do modelo.
- (2) Construção de Rede Neural:
 - a. Definição do número de neurônios por camada:
 - i. Camada de Entrada (e.g. Área, número de quartos, localização, idade);
 - ii. Camada Oculta (A depender conforme teste de desempenho);
 - iii. Camada de Saída (e.g. Preço do imóvel).
 - b. Escolha da Função de Ativação:
 - i. Camada Oculta (e.g. Função ReLU para relacionar preço e tamanho do imóvel, Função Leaky ReLU para relacionar o preço do imóvel e sua distância para uma avenida, que é um ponto negativo, de muito ruído, que impacta negativamente o preço)
 - ii. Camada de Saída (e.g. Função linear para prever valores contínuos como preço por metro quadrado);

- c. Conexões Ponderadas: Cada neurônio em cada camada (exceto a de entrada) está conectado a todos os neurônios na camada anterior e recebe valores de entrada que são ponderados por conexões. Durante o treinamento, esses pesos são ajustados para minimizar o erro de previsão. Caso entenda-se que o preço do imóvel é mais afetado por sua localização do que por seu tamanho, as conexões com a variável localização terão peso maior do que aquelas conectadas ao tamanho;
- (3) Ativação e Treinamento da Rede Neural, Avaliação e Ajustes: O desempenho do modelo pode ser avaliado usando o conjunto de teste, calculando métricas como erro médio absoluto (“*Mean Absolute Error*” – MAE). Durante o treinamento, os pesos da rede são ajustados iterativamente para minimizar o erro entre as previsões da rede e os preços reais das propriedades.

2.2.4.9.2 "Recurrent Neural Networks" (RNN) - Redes Recorrentes

No mercado de valoração em *Real Estate*, muitas vezes é útil considerar uma sequência de dados ao longo do tempo como, por exemplo, o histórico de preços de uma propriedade ao longo dos anos. Com o "*Recurrent Neural Networks*" (RNN), é possível processar dados sequenciais através de “*loops*” que permitem que as informações anteriores afetem as futuras, capturando dependências temporais (ARRIETA & et al., 2020).

Seguindo a estrutura passo-a-passo utilizada na explicação de aplicação para os modelos MLN (1- Coleta de Dados, Pré-Processamento e Divisão de dados, 2 – Construção de Rede Neural e 3 - Ativação e Treinamento da Rede Neural, Avaliação e Ajustes), é possível replicar as mesmas etapas. Entretanto, o RNN possui algumas diferenças pontuais a começar pela arquitetura do modelo. Enquanto as redes neurais “*feedforward*” tem as informações fluindo apenas em uma direção, da camada de entrada para a camada de saída, a RNN possui conexões recorrentes, com redes de “*backpropagation*”, ou retropropagação, que permitem que informações sejam realimentadas para camadas anteriores (GOODFELLOW, BENGIO, & COURVILLE, 2016).

A flexibilidade de cada modelo também é um fator de diferenciação. Como os modelos de “*feedforward*” são mais flexível em sua arquitetura, é possível ajustar o número de camadas ocultas, neurônios por camada e funções de ativação com maior facilidade, já a RNN possui uma arquitetura específica com realimentação de informações, o que a torna

mais especializada em lidar com séries temporais, mas menos flexível na quantidade de construção de redes neurais (CIABURRO & IANNACE, 2021). Abaixo, a Figura 18 ilustra o desenho de cada modelo.

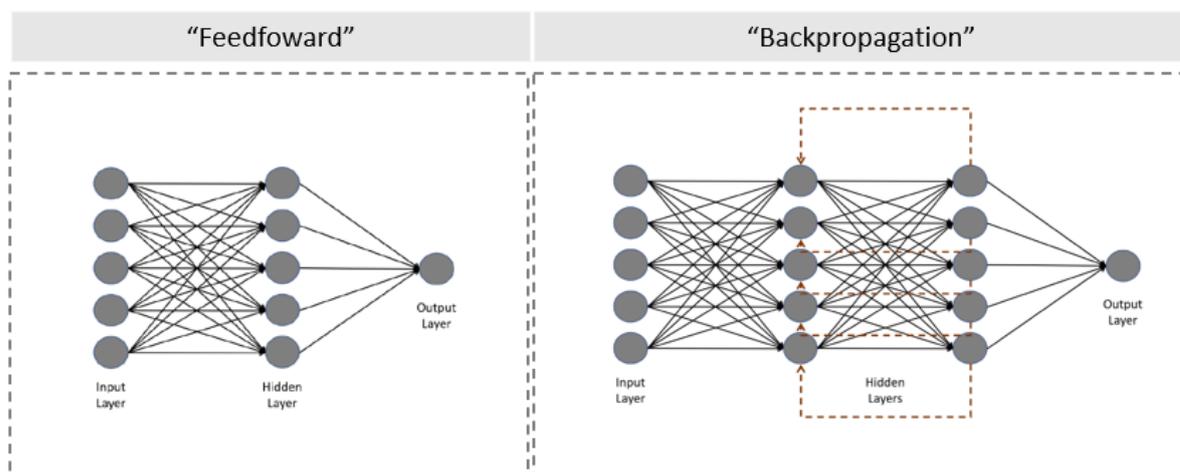


Figura 18 Ilustração da Estrutura de "Feedforward" e "Backpropagation" - Diagramas extraídos de CIABURRO, 2021

Outra diferença importante é a presença de unidades recorrentes nos modelos RNN, que têm a capacidade de manter estados internos e atualizá-los com base nas entradas atuais e nos estados anteriores, funcionando como uma memória interna (GOODFELLOW, BENGIO, & COURVILLE, 2016). Por isso, cada modelo é mais eficiente para casos diferentes, o MLP para prever o preço de imóveis com base em características estáticas (e.g. número de quartos, tamanho do terreno, localização etc.), o RNN consegue prever o preço de imóveis considerando a base de dados em séries temporais (e.g. histórico de preços ao longo do tempo).

2.2.4.9.3 "Convolutional Neural Networks" (CNN)

Essa arquitetura de rede tem como foco processar dados não-estruturados, como imagens. Utilizam camadas de convolução para extrair características das imagens, como bordas e texturas. São amplamente aplicadas em tarefas de visão computacional, como classificação de imagens e detecção de objetos (ARRIETA & et al., 2020).

Embora o uso deste algoritmo possa beneficiar o mercado de valoração (e.g. análise de fotografias realizadas em vistorias para definir estado de conservação do imóvel, classificação de propriedades a partir de atributos visuais, entre outros), este trabalho não

aprofundará suas aplicações e funcionamento visto que grande parte do levantamento bibliográfico para estudos de caso não aplica este modelo.

3 APLICAÇÕES DE MODELOS DE *MACHINE LEARNING* PARA ARBITRAGEM DE VALOR EM EMPREENDIMENTOS RESIDENCIAIS

A valoração de imóveis desempenha um papel importante tanto no mercado imobiliário quanto na economia em geral, sua implementação ocorre em várias áreas, como decisões de empréstimos, impostos e investimentos imobiliários (FLORENCIO & ALENCAR, 2016). A aplicação de ML na valoração imobiliária tem gerado resultados significativos em diversas pesquisas recentes ao redor do mundo, que serão detalhadas a seguir e exploram diferentes abordagens e técnicas de ML para, cada vez mais, arbitrar preços de venda ou de aluguel de imóveis que são coerentes com os valores que eles serão transacionados, ou seja, ao *fair tradable value* – valor para uma transação sem pressões.

Segundo Rocha Lima Júnior (1998), é necessário identificar transações de bens semelhantes ao empreendimento em avaliação dentro do mercado e construir uma amostra representativa suficiente para que seja possível inferir que o bem avaliado possa ser objeto de uma transação com valor comparável aos da amostra. Os modelos de ML seguem essa lógica e dependem de dados para serem treinados a identificar padrões (detalhados na seção 2.2.3 Inteligência Artificial). Entretanto, a evolução dos algoritmos tem permitido não só identificar padrões cada vez mais complexos como encontrar padrões similares mesmo em ativos que não são semelhantes como, por exemplo, em casas com diferentes características físicas, como será exemplificado nos estudos a seguir.

Abaixo, serão resumidos os estudos de caso selecionados a partir uma ampla bibliografia a fim de ilustrar alguns exemplos de aplicação dos algoritmos citados no capítulo anterior. Conforme Tabela 4, esses estudos de caso estão divididos em dois grupos (1) Avaliação de desempenho de modelos de ML para definição de valor de venda de um ativo residencial (BAUR, YASDANI e BAO), selecionando os algoritmos mais eficazes para os resultados, e (2) Avaliação de modelo para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais (LINCHUANG, POTRAWA e VIANA et al. Esses estudos de caso foram selecionados tanto pela diversidade quanto às fontes de dados e localização geográfica de cada exemplo como também pela variedade de algoritmos selecionados.

Baur (2023)	Yazdani (2021)	Gao (2022)	Linchuang (2021)	Potrawa (2022)	Viana et al. (2017)
BERLIM - ALEMANHA	COLORADO - EUA	SIDNEY - AUSTRALIA	XIAMEN - CHINA	ROTTERDÃ - HOLANDA	SÃO PAULO - BRASIL
<p>Objetivo</p> <p>Seleção de modelo mais eficiente</p>	<p>Seleção de modelo mais eficiente</p>	<p>Seleção de modelo mais eficiente</p>	<p>Teste de modelo para para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais</p>	<p>Teste de modelo para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais</p>	<p>Teste de modelo para para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais</p>
<p>Metodologia</p> <p>Diversos modelos de ML foram treinados a partir da inclusão de conjuntos de dados de ofertas imobiliárias para apartamentos de aluguel em Berlim.</p>	<p>O autor do estudo explorou e comparou várias técnicas para prever o preço de imóveis em Boulder, Colorado, nos Estados Unidos.</p>	<p>Analisou um conjunto de dados de imóveis na região de Sydney, Austrália, com o objetivo de avaliar o desempenho de diferentes técnicas de ML na previsão de valores de imóveis.</p>	<p>Utilizou o algoritmo de aprendizado de máquina Gradient Boosting Decision Tree (GBDT) para avaliar os efeitos não-lineares do Bus Rapid Transit (BRT), ou corredor de ônibus, nos preços de imóveis em Xiamen, China.</p>	<p>Melhorar a metodologia hedônica tradicional, a fim de compreender a percepção dos investidores sobre o valor das características dos imóveis em Roterdã</p>	<p>Utilizou métodos ML para análise da taxa de rentabilidade de empreendimentos imobiliários residenciais para locação de curto prazo.</p>
<p>Base de Dados</p> <p>Características estruturadas da propriedade, como a idade do prédio e o número de quartos, bem como dados não estruturados, como a descrição textual da propriedade.</p>	<p>Dados do MLS ("Multiple Listing Service"), classificações de escolas públicas, taxas de criminalidade (Crime Reports), informações estatísticas do Colorado, classificação com notas para localizações com base na capacidade de fazer múltiplas tarefas a uma distância a pé (WalkScore), e o Censo (US Census Bureau). O conjunto de dados final compreendeu 1.018 imóveis residenciais vendidos em Boulder em 2019.</p>	<p>Características dos imóveis, como número de quartos, banheiros, garagens, tipo de imóvel e tamanho da área, bem como características locais, como distância para o centro da cidade, distância para a costa, distância para a universidade e distância para o hospital.</p>	<p>Coleta de dados de imóveis em áreas dentro de 1,5 km do corredor BRT e a aplicação do algoritmo GBDT para examinar a relação não-linear entre o BRT e os preços dos imóveis.</p>	<p>Informações sobre as características das propriedades (como o preço, a área de estar, o número de quartos, o tipo de casa, a presença de garagem, banheiro, elevador, entre outros) bem como informações sobre o tempo de viagem a pé, de bicicleta e o transporte público até o centro da cidade. Além disso, também foram extraídas imagens das propriedades, coletadas de sites de aluguel e do Google Maps. Essas imagens foram usadas para extrair variáveis adicionais, como a presença de uma vista da cidade, a proximidade de áreas verdes ou corpos d'água, e a qualidade da insolação.</p>	<p>Os dados utilizados foram oriundos do portal de classificados online VivaReal e do site de serviços de hospitalidade Airbnb, totalizando 2.177 imóveis analisados.</p>
<p>Algoritmos</p> <p>Random Forest, Support Vector, Gradient Boosting.</p>	<p>"K-Nearest Neighbors" (KNN), "Rule Based Learners" (RBL), "Random Forest (RF)", "Support Vector Machines" (SVM), "Neural Networks" (NN)</p>	<p>Linear/ Logistic Regression" (LR), "Decision Trees" (DT), Random Forest (RF), "Support Vector Machines" (SVM), "Neural Networks" (NN)</p>	<p>" Linear/ Logistic Regression" (LR), "Gradient Boosting Decision Tree" (GBDT)</p>	<p>" Linear/ Logistic Regression" (LR), "Random Forest" (RF), "Convolutional Neural Networks" (CNN)</p>	<p>"Decision Trees" (DT), Random Forest (RF)</p>

Tabela 4 Resumo dos estudos de caso

i. Baur (2023)

Em um estudo realizado por Baur (2023), diversos modelos de ML foram treinados a partir da inclusão de conjuntos de dados de ofertas imobiliárias para apartamentos de aluguel em Berlim. Os dados incluem características estruturadas da propriedade, como a idade do prédio e o número de quartos, bem como dados não estruturados, como a descrição textual da propriedade.

Para que essas informações pudessem ser usadas na arbitragem do preço de venda dos imóveis em modelos de ML, as descrições foram transformadas em representações numéricas. Alguns exemplos apresentados em seu artigo para essa transformação são o método de “*bag-of-words*”, que conta o número de ocorrências de uma palavra dentro de um texto ou modelo, selecionando as palavras mais influentes na descrição de uma propriedade e as inclui como parte de um modelo de regressão linear.

Os modelos de ML, tais como *Random Forest*, *Support Vector*, *Gradient Boosting* (Explicados detalhadamente na Seção 2.2.4 *Machine Learning*), entre outros, foram treinados para arbitrar o valor de aluguel dos imóveis com base nas características estruturadas e nas representações numéricas das descrições das propriedades. Como resultado, o modelo de maior desempenho foi o *Gradient Boosting*, que se beneficiou da inclusão de descrições textuais ao lidar com dados ausentes.

Conforme explicado na seção 2.2.4.7, Figura 14, o funcionamento de ajuste de pesos a partir do treinamento dos dados acaba melhorando a precisão do modelo e, dentre os resultados apontados por Baur, reduz significativamente o erro de previsão, havendo redução dos erros médios absolutos (MAE), em até 17,09%, em comparação aos modelos tradicionais, para ofertas de aluguel de apartamentos em Berlim.

ii. Yazdani (2021)

Outro estudo, conduzido por Yazdani (2021), indica como modelos de ML podem ser extremamente úteis em diversas áreas, incluindo a arbitragem de valor de aluguel ou venda de imóveis. O autor do estudo explorou e comparou várias técnicas para prever o preço de imóveis em Boulder, Colorado, nos Estados Unidos. Os dados para o estudo foram coletados de várias fontes, incluindo bancos de dados do MLS (“*Multiple Listing*

Service)⁶, classificações de escolas públicas, taxas de criminalidade (*Crime Reports*) e informações estatísticas do Colorado, classificação com notas para localizações com base na capacidade de fazer múltiplas tarefas a uma distância a pé (*WalkScore*), e o Censo (*US Census Bureau*). O conjunto de dados final compreendeu 1.018 imóveis residenciais vendidos em Boulder em 2019.

Os resultados indicaram uma associação não-linear entre as características das propriedades e seus preços, por exemplo, ao invés de uma relação linear direta (mais quartos resulta em preço mais alto), o modelo pode indicar um número ótimo de quartos que maximiza o valor percebido, ou seja, o preço que será pago por tais atributos. Por exemplo, em uma determinada área, imóveis com três quartos podem ser mais valorizados do que imóveis com dois ou quatro quartos.

A partir da comparação entre o modelo de precificação hedônica, *Neural Networks* (NN), *Random Forest* (RF), método *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) e *Rule Based Learner* (RBL com aplicação de *fuzzy rules*) M o estudo busca arbitrar o valor de venda de casas na cidade de Boulder, chegando num valor próximo ao que as casas foram de fato vendidas. Ou seja, é demonstrada uma associação não linear entre as características das moradias e seus preços, o que sugere que esses modelos podem ser mais precisos e eficientes do que o modelo de precificação hedônica para prever os preços do mercado imobiliário.

iii. Gao (2022)

Seguindo a mesma linha de raciocínio da Yazdani, a pesquisa de Gao (2022) analisou um conjunto de dados de imóveis na região de Sydney, Austrália, com o objetivo de avaliar o desempenho de diferentes técnicas de ML na previsão de valores de imóveis. Os dados incluíam características dos imóveis, como número de quartos, banheiros, garagens, tipo de imóvel e tamanho da área, bem como características locais, como distância para o centro da cidade, distância para a costa, distância para a universidade e distância para o hospital.

⁶ “Um serviço de listagem múltipla (MLS) é um banco de dados estabelecido por corretores imobiliários cooperantes para fornecer dados sobre propriedades à venda. Um MLS permite que os corretores vejam as listagens de propriedades à venda uns dos outros com o objetivo de conectar compradores de casas a vendedores” Fonte: <https://realtyna.com/blog/list-mls-us/> (acessado em Fevereiro/2024)

Para o estudo, os modelos foram aplicados em diferentes níveis geográficos (áreas estatísticas definidas pelo Bureau de Estatísticas da Austrália) e foram aplicados diversos métodos, incluindo preço hedônico (HPM) com regressão linear, *Decision Trees*, *Random Forest*, *Gradient Boosting*, *Support Vector Machine* e *Neural Networks*.

Os resultados mostraram que os algoritmos de aprendizado de máquina e a regressão geograficamente ponderada (GWR) superaram os métodos tradicionais de preços hedônicos. Em particular, o *Gradient Boosting* mostrou o melhor desempenho em termos de precisão da previsão. Além disso, a pesquisa revelou que a localização geográfica e as características do bairro desempenham um papel significativo na precisão dos algoritmos e arbitragem dos valores dos imóveis.

iv. Linchuang (2021)

Complementarmente, Linchuang (2021) utilizou o algoritmo de aprendizado de máquina *Gradient Boosting Decision Tree* (GBDT), quando comparado aos modelos de regressão linear, para avaliar os efeitos não-lineares do *Bus Rapid Transit* (BRT), ou corredor de ônibus, nos preços de imóveis em Xiamen, China. A metodologia envolveu a coleta de dados de imóveis em áreas dentro de 1,5 km do corredor BRT e a aplicação do algoritmo GBDT para examinar a relação não-linear entre o BRT e os preços dos imóveis. Os resultados indicaram uma associação positiva entre a acessibilidade às estações de BRT e os preços dos imóveis, e uma associação negativa entre a proximidade ao corredor do BRT e os preços dos imóveis. Sendo assim, a partir de determinada distância aos BRTs a relação de proximidade muda a qualidade da associação, sugerindo uma relação não linear entre BRT e preços de casas.

v. Potrawa (2022)

Num estudo mais conservador, Potrawa (2022) focou seus esforços em modernizar a metodologia hedônica tradicional, com seus modelos simples de regressão linear, a fim de compreender a percepção dos investidores sobre o valor das características dos imóveis em Roterdã, utilizando o modelo *Random Forest* para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais.

Os dados utilizados nesse estudo foram coletados do *Journal of Business Research*, que incluíam informações sobre as características das propriedades (como o preço, a área de

estar, o número de quartos, o tipo de casa, a presença de garagem, banheiro, elevador, entre outros) e descrições das propriedades, bem como informações sobre o tempo de viagem a pé, de bicicleta e o transporte público até o centro da cidade. Além disso, também foram extraídas imagens das propriedades, coletadas de sites de aluguel e do Google Maps. Essas imagens foram usadas para extrair variáveis adicionais, como a presença de uma vista da cidade, a proximidade de áreas verdes ou corpos d'água, e a qualidade da insolação. Essas variáveis foram extraídas usando técnicas de reconhecimento de imagem, como *Convolutional Neural Networks*.

Os resultados mostraram que a aplicação de técnicas de ML aumentou em 25% a precisão dos modelos de preços hedônicos. Identificando que o preço marginal da área habitável para propriedades alugadas é apenas linear por partes, sendo que cada metro quadrado adicional entre alguns grupos de faixa de metragem representa um aumento ou diminuição do valor percebido, ou seja, daquilo que pode ser convertido em preço pago pelo ativo.

Adicionalmente, com o uso de covariáveis extraídas de imagens de satélite foi descoberto que o valor dos atributos de habitação depende não apenas de sua própria quantidade e qualidade, mas também da quantidade de outros atributos. O valor da vista panorâmica de Roterdã foi estimado em mais de 100 euros por mês, para propriedades grandes e bem localizadas.

i. Viana et al. (2017)

Finalmente, utilizando um estudo de caso brasileiro, Viana et al. (2017) utilizou métodos ML para análise da taxa de rentabilidade de empreendimentos imobiliários residenciais para locação de curto prazo. Os dados utilizados foram oriundos do portal de classificados online VivaReal e do site de serviços de hospitalidade Airbnb, totalizando 2.177 imóveis analisados. A partir da primeira classificação sobre qual a taxa de rentabilidade nominal dos ativos (maior ou menor que 7%), foram utilizados os algoritmos *Neural Networks* e *Decision Trees* para avaliar quais características em comum os empreendimentos com TR acima de 7% possuem.

Como resultado, detectou-se qual a importância das variáveis independentes em cada algoritmo, identificando Preço do Imóvel, Diária Média, Localização e Área (tamanho do

imóvel) como os principais fatores, seguidos de Vagas de Garagem, Valor do Condomínio e Número de Dormitórios, nesta ordem.

Conforme ressaltado pelos autores, rentabilidade passada não significa rentabilidade futura e as informações obtidas no estudo podem ser utilizadas como referência para uma segunda etapa de criação de cenários a partir dos quais a análise de qualidade do investimento poderá ser feita entre fronteiras de comportamento. Essa observação se aplica a todos os estudos apresentados anteriormente.

Outra observação relevante apontada pelos autores foi sobre a utilização de preços transacionados em substituição aos preços anunciados, que podem significar informações mais adequadas. Ainda assim, tomando como base a superioridade da análise através dos algoritmos de ML em comparação com a regressão logística, tanto devido à diminuição dos erros de estimação quanto pela melhor adequação para incluir variáveis não-lineares no modelo, a utilização dos algoritmos continua sendo útil para a identificação de fatores de influência no valor transacionado ou da taxa de rentabilidade.

Em suma, comparando os exemplos acima, os avanços no uso de ML na valoração imobiliária têm revelado sua eficácia na previsão de preços, mas também ressaltam a importância de equilibrar precisão com interpretabilidade, conforme estudos de (ARRIETA & et al., 2020) sobre XAI (“*Explainable Artificial Intelligence*”), detalhados na seção 2.2.4. À medida que a tecnologia continua a desempenhar um papel fundamental na avaliação imobiliária, é essencial considerar a evolução das abordagens e suas implicações na tomada de decisões no mercado imobiliário.

4 METODOLOGIA DE PESQUISA

Este capítulo apresenta a metodologia empregada na elaboração do questionário, juntamente com a fundamentação teórica que deu origem às perguntas para as entrevistas. Adicionalmente, esclarece-se o critério de seleção das empresas entrevistadas e a forma de aplicação e apresentação dos dados.

4.1 METODOLOGIA

Visando atender aos objetivos deste trabalho, a metodologia de pesquisa aplicada nesse trabalho é a SURVEY, conforme explicitado no item 1.3.

Para isso a autora necessitou fazer: a) Revisão bibliográfica sobre o tema (capítulo 02); b) Definição de questionário; c) Seleção das empresas para, finalmente; d) Aplicação do questionário; e, e) Análise dos dados. Todas as etapas, a partir do item b, serão detalhadas a seguir.

4.2 FORMULAÇÃO DO QUESTIONÁRIO

Este estudo, em consonância com a dissertação de CONWAY (2018), busca responder a duas questões centrais, adaptadas a seguir para atender não somente ao setor imobiliário em geral, mas também para o nicho em que essa pesquisa busca focar (Empreendimentos Residenciais de Base Imobiliária):

- A. Como as tecnologias de IA e ML estão sendo aplicadas no setor residencial de base imobiliária?
- B. Quais oportunidades concretas se apresentam para a adoção de IA e ML no setor residencial de base imobiliária?

Para abordar essas questões, este trabalho seguiu 6 das 7 perguntas de CONWAY (2018), fazendo algumas adaptações e adicionando perguntas pertinentes, apontadas na **tabela 4 e 5**, na coluna “Adaptação”, da seguinte forma:

- I. (*): Questão mantida exatamente como na pesquisa de CONWAY (2018).
- II. (**): Questão adaptada (eg. Afunilamento para o nicho de mercado imobiliário de valoração para Empreendimentos de Base Imobiliária, adição de alternativas que ou se encaixam nos grupos definidos por CONWAY, de modo a facilitar a comparação entre os trabalhos, ou definem atividades específicas do ramo de valoração imobiliária, a fim de especificar as respostas para o tema central deste trabalho, ou específicas para detalhar tipos ou fontes de dados, por exemplo.
- III. (***) : Questão diferente do questionário original, ou removida por falta de coerência com o trabalho ou adicionada para atingir os objetivos do trabalho.

Com o objetivo de assegurar a clareza das indagações e respostas, bem como garantir a qualidade das entrevistas, implementou-se melhorias relacionadas à lista de respostas (em amarelo na **tabela 4 e 5**). Foi introduzida uma seleção de alternativas de múltipla escolha para cada pergunta, visando abranger amplamente as opções possíveis. Essa abordagem garantiu que os entrevistados fossem sempre capazes de selecionar uma resposta dentro de um intervalo definido, ao mesmo tempo em que se reservava espaço para contribuições adicionais, tanto por meio de comentários quanto através de uma opção de resposta dissertativa, a fim de englobar qualquer possibilidade não prevista nas alternativas predefinidas.

Adicionalmente, a organização das perguntas também foi ajustada para que a entrevista pudesse ser dividida em dois blocos: 1) Aplicação na Empresa e 2) Opinião Geral de Mercado. Deste modo, buscou-se, atingir uma dinâmica em que seria possível detectar o nível de aderência às tecnologias de ML e IA, se já estava presente nas empresas ou se estava num estágio inicial, em que existe a ciência sobre sua existência e potencial, porém sem abertura para aplicação no ambiente profissional da empresa entrevistada.

O questionário completo pode ser encontrado no Apêndice 1, na parte final deste material.

#	CONWAY (EN)	CONWAY (PT)	Questões Aplicadas no Questionário	Comentários sobre as adaptações das questões	Adaptação
1	What area of technology/real estate are you/your company focused on?	Em qual área de tecnologia/mercado imobiliário você/sua empresa estão focados?	Não Aplicável	Foco apenas em seleção de empresas focadas em residenciais de base imobiliária.	(***)
SEÇÃO 1: FOCO: APLICAÇÃO NA EMPRESA					
2	Many articles and media outlets are talking about machine learning and AI, but there may be inconsistencies in what is considered machine learning or AI. How would you define machine learning and artificial intelligence?	Muitos artigos e veículos de mídia falam sobre aprendizado de máquina e IA, mas pode haver inconsistências no que é considerado aprendizado de máquina ou IA. Como você definiria aprendizado de máquina e inteligência artificial?	Muitos artigos e veículos de mídia falam sobre aprendizado de máquina e IA, mas pode haver inconsistências no que é considerado aprendizado de máquina ou IA. Como você definiria aprendizado de máquina e inteligência artificial?	Questão dissertativa mantida conforme CONWAY.	(*)
3	Data is important to ML and AI modeling and tools. What are the key data sources in your business? What types of data are you using/collecting to be used in your product or service?	Dados são importantes para modelagem e ferramentas de ML e IA. Quais são as principais fontes de dados em seu negócio? Que tipos de dados você está utilizando/coletando para serem usados em seu produto ou serviço?	Dados são importantes para modelagem e ferramentas de ML e IA. Quais são as principais fontes de dados em seu negócio e quais os principais tipos de dados coletados?	Adição de alternativas de múltipla escolha para auxiliar o entrevistado na especificação de possíveis fontes de dados. Adição de alternativas para auxiliar o entrevistado na especificação de possíveis tipos de dados.	(**)
4	I am interested in understanding how machine learning is being used in real estate today. Is this something that you/your company are using? If so, please describe how the techniques are being applied.	Estou interessado em entender como a aprendizagem de máquina está sendo utilizada no mercado imobiliário atualmente. Isso é algo que você/sua empresa estão utilizando? Se sim, por favor, descreva como as técnicas estão sendo aplicadas.	Estou interessado em entender como o ML está sendo utilizado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente. Isso é algo que você/sua empresa estão utilizando? Se sim, por favor, descreva como as técnicas estão sendo aplicadas.	Adaptação da questão, filtrando o nicho de mercado imobiliário de valoração para Empreendimentos de Base Imobiliária. Adição de alternativas que se encaixam nos grupos definidos por CONWAY.	(**)
	Não Aplicável	Não Aplicável	Caso a pergunta 3 tenha sido respondida com "Sim", poderia informar quais algoritmos de ML estão sendo utilizados?	Inclusão da questão para especificar, caso exista aplicação de tecnologias de ML na empresa entrevistada, quais os algoritmos que estão sendo utilizados (alternativas) e como (dissertativa).	(***)
	Não Aplicável	Não Aplicável	Caso a pergunta 3 tenha sido respondida com "Sim", poderia informar em quais áreas específicas tais tecnologias tem sido aplicadas na sua empresa?	Inclusão da questão para especificar, caso exista aplicação de tecnologias de ML na empresa entrevistada, quais áreas específicas, dentro de valoração imobiliária, que estão sendo aplicadas (alternativas) e como (dissertativa).	(***)

Tabela 5 Lista de Questões Aplicadas no Questionário (Seção 1: Foco em Aplicação na Empresa): Versão Original, Traduzida e Adaptada - com comentários sobre os ajustes realizados

#	CONWAY (EN)	CONWAY (PT)	Questões Aplicadas no Questionário	Comentários sobre as adaptações das questões	Adaptação
SEÇÃO 2: FOCO: OPINIÃO GERAL DE MERCADO					
5	Are you aware of any other ways that ML is being applied in real estate or related fields today?	Você está ciente de outras maneiras pelas quais ML está sendo aplicado no mercado imobiliário ou em outras áreas relacionadas ao mercado imobiliário atualmente?	Você está ciente de outras maneiras pelas quais o ML está sendo aplicado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente?	Adaptação da questão, filtrando o nicho de mercado imobiliário de valoração para Empreendimentos de Base Imobiliária. Adição de alternativas que se encaixam nos grupos definidos por CONWAY.	(**)
6	Are there any areas of the real estate business that you think would be good opportunities for applying ML or AI techniques in the future?	Existem áreas em negócios do mercado imobiliário que você acredita serem boas oportunidades para aplicar técnicas de ML ou IA no futuro?	Existem áreas específicas no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária que você acredita serem boas oportunidades para aplicar técnicas de aprendizado de máquina ou inteligência artificial no futuro?	Adaptação da questão, filtrando o nicho de mercado imobiliário de valoração para Empreendimentos de Base Imobiliária. Adição de alternativas que se encaixam em atividades específicas do ramo de valoração imobiliária.	(**)
7	What types of data might be missing or useful to enhance analysis or ML and AI applications in real estate?	Que tipos de dados podem estar faltando ou ser úteis para aprimorar a análise ou aplicações de ML e IA no mercado imobiliário ?	Que tipos de dados podem estar faltando ou ser úteis para aprimorar a análise ou aplicações de ML e IA no mercado de Empreendimentos Residenciais de Base Imobiliária ?	Adaptação da questão, filtrando o nicho de mercado imobiliário de valoração para Empreendimentos de Base Imobiliária. Adição de alternativas para auxiliar o entrevistado na especificação de possíveis tipos de dados.	(**)

Tabela 6 Lista de Questões Aplicadas no Questionário (Seção 1: Foco na Opinião Geral de Mercado): Versão Original, Traduzida e Adaptada - com comentários sobre os ajustes realizados

4.3 SELEÇÃO DAS EMPRESAS

Esse trabalho partiu da seleção de empresas relacionadas ao mercado de empreendimentos de base imobiliária residencial, com foco em valoração, que atuam na cidade de São Paulo. Tal triagem foi realizada com base em fatores como reconhecimento de mercado, presença global, e percepção geral do porte de cada empresa no setor imobiliário, ou seja, uma simplificação generalizada não baseada em métricas financeiras específicas, mas que fosse possível criar um grupo heterogêneo, tentando capturar uma amostra representativa do mercado de valoração.

Não houve pré-seleção de empresas que já aplicam as tecnologias discutidas neste trabalho, entretanto, com o intuito de obter uma compreensão abrangente do panorama do mercado de valoração de ativos imobiliários e facilitar a compreensão sobre as empresas entrevistadas ainda que preservando suas identidades, a seleção de empresas para entrevistas foi dividida em três categorias distintas, somando 10 empresas, que serão ilustradas na Figura 19 Quadro resumo de agrupamento das empresas selecionadas para entrevista. Fonte: Elaboração própria. Figura 19 e explicadas a seguir: 4 focadas na comercialização de ativos residenciais para renda (Grupo C), 5 direcionadas ao desenvolvimento, administração e operação de residenciais de locação (“*Long-Stay*”⁷) (Grupo LS) e, finalmente, 1 empresa focada no fornecimento de inteligência com base em dados para esse setor (Grupo BD) – O porte das empresas foi classificado conforme número de funcionários de acordo com o SEBRAE⁸.

⁷ “empreendimentos residenciais voltados para renda” (Fonte: <https://cte.com.br/blog/gerenciamento/long-stay/>). Acessado em Maio/2024.

⁸ https://sebrae.com.br/Sebrae/Portal%20Sebrae/UFs/SP/Pesquisas/MPE_conceito_empregados.pdf

Grupo	#	Empresa	Porte
Comercialização	1	C1	MÉDIO
	2	C2	GRANDE
	3	C3	GRANDE
	4	C4	GRANDE
Long Stay	5	LS1	MÉDIO
	6	LS2	MÉDIO
	7	LS3	GRANDE
	8	LS4	MÉDIO
	9	LS5	MÉDIO
Base de Dados	10	BD1	PEQUENO

Figura 19 Quadro resumo de agrupamento das empresas selecionadas para entrevista. Fonte: Elaboração própria.

As distintas categorias das empresas selecionadas apresentam motivações específicas e potenciais aplicações distintas entre si para a implementação de *Machine Learning* (ML) e *Inteligência Artificial* (AI) no nicho em estudo, desse trabalho. Entende-se que a diversidade desses segmentos pode proporcionar maior visualização sobre como essas tecnologias estão sendo incorporadas, compreendidas e exploradas no setor residencial para renda, contribuindo assim para uma análise holística da evolução do mercado imobiliário no contexto tecnológico.

Através de uma rede de contatos profissionais e acadêmicos, foi possível estabelecer contato com todas as 10 empresas que foram pré-selecionadas para este estudo. Entre essas empresas 8 apresentaram disponibilidade para realização da entrevista, 4 no grupo “C”, 3 do grupo “LS” e 1 do grupo “BD”. Todas as empresas serão detalhadas a seguir:

I. Grupo C: Empresas focadas na comercialização de ativos residenciais para renda

a. Empresa C1
(Empresa entrevistada)

A empresa atua em São Paulo e começou como incorporadora, voltada à venda de imóveis, adicionando locação de imóveis e administração de condomínios, com aproximadamente 3.500 condomínios sob sua gestão, no seu *core-business*. Como parte da inteligência gerado em seu negócio, a empresa realiza análises intra, inter e extra condominiais a partir do armazenamento e geolocalização de dados advindos de seus condomínios.

Tamanho da empresa: 501-1.000 funcionários

Fundada em: 1954

b. Empresa C2
(Empresa entrevistada)

Plataforma de aluguel residencial com atuação tanto no Brasil quanto no México, com soluções tecnológicas e financeiras para contribuir com a melhora de processos e atividades de imobiliárias (eg. Gestão de anúncios, Fiança de aluguel, Financiamento, Seguro etc.), abrangendo 8 mil imobiliárias e, aproximadamente, 180 mil imóveis.

Tamanho da empresa: 1.001-5.000 funcionários

Fundada em: 2018

c. Empresa C3

(Empresa entrevistada)

Empresa, com atuação nacional, especialista em na compra e venda de imóveis, locação e administração, e lançamentos imobiliários, prestando assessoria a incorporadores, compradores e vendedores de imóveis novos e usados.

Tamanho da empresa: 501-1.000 funcionários

Fundada em: 1935

d. Empresa C4

(Empresa entrevistada)

Plataforma de aluguel residencial, com sede no Brasil e atuação em outros países da América Latina, que oferece soluções para todas os *stakeholders* (imobiliárias, corretores, síndicos, condôminos, compradores, vendedores, proprietários e inquilinos) envolvidos nas etapas do aluguel da moradia. Além da comercialização, também possui banco de dados que auxilia na criação de indicadores e relatórios de inteligência de mercado.

Tamanho da empresa: 1.001-5.000 funcionários

Fundada em: 2012

II. Grupo “LS”: Empresas focadas no desenvolvimento, administração e operação de residenciais de locação (“*Long-Stay*”)

a. Empresa LS1

(Empresa entrevistada)

Empresa brasileira que desenvolve, administra e opera residências para locação em São Paulo. Criada a partir de um fundo de investimento que investe no mercado imobiliário em diversos países, a empresa

atingiu um faturamento de R\$ 17 milhões em 2023, com um crescimento de 317% em comparação ao ano anterior.

Tamanho da empresa: 51-200 funcionários

Fundada em: 2017

b. Empresa LS2

(Empresa entrevistada)

*Proptech*⁹, com atuação nacional, que oferece soluções no setor residencial para renda, ajudando inquilinos a pivotar entre unidade da franquia de maneira flexível a depender de suas necessidades, incorporadores a potencializarem suas vendas e investidores a maximizarem seus lucros.

Tamanho da empresa: 51-200 funcionários

Fundada em: 2019

c. Empresa LS3

(Empresa entrevistada)

Empresa Multinacional, focada no mercado de imóveis para aluguel e especializadas em todos os setores residenciais (eg. residências estudantis, multifamiliares, residenciais para idosos etc.). Com gerenciamento de mais de US\$ 41 bilhões em ativos e operação de cerca de 740.000 unidades habitacionais para aluguel em todo o mundo.

Tamanho da empresa: + de 10.001 funcionários

Fundada em: 1993

⁹ “(do inglês, property technology, ou tecnologia imobiliária) é o uso de software (ou qualquer outra forma de tecnologia da informação) para otimizar o processo de gestão de propriedades ou imóveis.” (Fonte: <https://www.jotform.com/pt/blog/o-que-e-proptech/>) Acessado em: Maio, 2024

d. Empresa LS4

(Empresa não conseguiu agenda para entrevista)

Proptech, com atuação na cidade de São Paulo, que oferece soluções para *stakeholders* do setor residencial para renda, com mais de 1000 unidades e estando presente de ponta a ponta no processo de locação, desde o processo de curadoria dos investimentos, com sistema de precificação inteligente, até a produção das unidades e a comercialização delas em plataforma de locação e suporte aos inquilinos. A empresa já captou mais de R\$200 milhões através de aportes de fundos de Venture Capital e de retornos do seu Fundo de Investimento Imobiliário.

Tamanho da empresa: 51-200 funcionários

Fundada em: 2019

e. Empresa LS5

(Empresa não conseguiu aprovação para responder dentro do prazo devido a informações confidenciais)

Empresa focada no mercado imobiliário de aluguel de imóveis de alto padrão com mais de 600 unidades residenciais para locação.

Tamanho da empresa: 51-200 funcionários

Fundada em: 2015

III. Grupo “BD”: Empresas focadas no oferecimento de inteligência com base em dados para o setor residencial para renda

a. Empresa BD1

(Empresa entrevistada)

Plataforma com dados territoriais para mercado imobiliário (legal, fiscal, ambiental, infraestrutura, preço de imóveis etc.) e consultoria

para assessorar governos e empresas em análises de mercado e decisões de políticas públicas (precificação de imóveis, identificação de oportunidades de vendas, ciclos de mercado etc.).

Tamanho da empresa: 2-10 funcionários

Fundada em: 2018

Dentre as empresas entrevistadas, foram escolhidos profissionais ocupando cargos de liderança (Fundadores, CEOs, Gerentes etc.) ou funcionários indicados por líderes nesses cargos. Entende-se que essa seleção é crucial para captar pessoas estratégicas que podem oferecer uma visão ampla, compreendendo não apenas a atual implementação de tecnologias como ML, mas também vislumbrando oportunidades para aprimorar sistemas e processos internos. Essa escolha visa não apenas compreender o estado atual, mas também antecipar e explorar as potenciais melhorias e inovações que essas lideranças enxergam para o futuro de suas organizações.

4.4 APLICAÇÃO DO QUESTIONÁRIOS

Foram oferecidas duas formas de preenchimento do questionário, entrevista por videochamada e preenchimento de maneira independente, a partir do envio do formulário. Para ambos os casos, o questionário contou, tanto com alternativas de múltipla escolha como também espaços no formulário dedicados para explanação e exemplos com o intuito de ilustrar e justificar a escolha das alternativas. Todos os entrevistados, com exceção da empresa “LS1” preferiram preencher o questionário de forma independente.

A primeira questão, “Muitos artigos e veículos de mídia falam sobre aprendizado de máquina e IA, mas pode haver inconsistências no que é considerado aprendizado de máquina ou IA. Como você definiria aprendizado de máquina e inteligência artificial?”, foi colocada no início da entrevista a fim de avaliar o nível de conhecimento dos entrevistados antes de adentrar nas seções de foco.

Após essa avaliação inicial a fim de assegurar que as respostas nas seções subsequentes estivessem fundamentadas em um entendimento abrangente dessas tecnologias, o questionário seguiu duas seções explicadas a seguir.

I. SEÇÃO 1: FOCO EM APLICAÇÕES NA EMPRESA

Nesta seção, o objetivo é identificar se há implementações de ML e IA nas empresas representadas pelos entrevistados. Compostas por cinco perguntas, a seção abrange (a) especificação sobre quais tipos de dados e fontes de dados a empresa utiliza, (b) exemplos de maneiras que a empresa utiliza esses dados e (c) caso exista uso de ML, como esse uso está sendo aplicado (quais algoritmos e quais as áreas em valoração).

Conforme mencionado na seção 4.3 (Seleção de empresas entrevistadas), não houve filtragem prévia para selecionar as empresas que utilizam ML e AI de alguma maneira, possibilitando, assim, respostas afirmativas ou negativas à pergunta sobre implementação dessas tecnologias nas empresas. Apesar disso, a continuidade da seção, ao indagar sobre os tipos de dados utilizados, busca-se extrair respostas que podem direcionar a *insights* para a identificação de potenciais benefícios associados ao emprego dessas tecnologias.

II. SEÇÃO 2: FOCO EM OPINIÃO GERAL DE MERCADO E IDENTIFICAÇÃO DE OPORTUNIDADES

Esta seção, composta por três perguntas, visa compreender a visão dos entrevistados em relação às oportunidades que podem aprimorar as práticas no mercado imobiliário. Inicialmente, o entrevistado é questionado sobre se está ciente de outras maneiras pelas quais o ML está sendo aplicado no nicho em questão e, em seguida, mais especificamente, sobre quais as áreas, dentro do mercado de valoração desses ativos, eles acreditam que existe a oportunidade de aplicar as técnicas de ML. Finalmente, de maneira mais genérica, a fim de captar respostas de entrevistados menos familiarizados com

a tecnologia, mas que podem estar num estágio inicial de utilização de dados, pergunta-se quais tipos de dados eles acreditam estarem faltando ou serem úteis para que enfim o ML possa ser aplicado.

4.5 APRESENTAÇÃO DOS DADOS

Pergunta 1: Muitos artigos e veículos de mídia falam sobre *Machine Learning* (ML) e Inteligência Artificial (IA), mas pode haver inconsistências no que é considerado ML ou IA. Como você definiria ML e IA?

Conforme apresentado na seção 4.4 (Aplicação do questionário), a primeira pergunta buscou assegurar que as respostas nas seções subsequentes estivessem fundamentadas em um entendimento sobre o que é ML e IA.

Comparando as respostas com os conceitos apresentados na seção 2.2.3 (Inteligência Artificial) e 2.2.4 (*Machine Learning*), todos os entrevistados, com exceção da empresa C2, chegaram a uma descrição bastante próxima das definições apresentadas neste trabalho. A empresa C2, conforme será evidenciada a seguir, possui um entendimento menos teórico e mais de mercado, entretanto, isso não é indicativo de que o entrevistado não conhece a tecnologia. Pelo contrário, ele possui uma opinião com base em experiências profissionais.

As justificativas apresentadas abaixo indicam que essas tecnologias não representam um tema desconhecido para os entrevistados. Ideias-chave como raciocinar, aprendizado (supervisionado, não-supervisionado e por reforço), reconhecer padrões, ações evolutivas e a ausência de interferência humana no processo de funcionamento do algoritmo foram citadas nas respostas, que serão expostas a seguir:

IA é um termo designado para aplicações automatizadas que executam tarefas complexas que antes exigiam interação humana. Machine Learning é um subcampo da IA, é um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. O machine learning é centrado na criação de sistemas que aprendam ou melhorem seu desempenho com base nos dados que eles consomem. Embora todo machine learning seja IA, nem toda IA é machine learning. (Empresa C1)

Contemporaneamente, caracterizaria ML como conjuntos de algoritmos e técnicas com maior apelo para dados tabulares e estruturados. Restringiria IA para visão computacional e LLMs. Mas é muito difícil separar as definições. Elas são mais fortemente influenciadas pelo mercado do que pela semântica ou pelas técnicas per se. (Empresa C2)

A capacidade de computadores, com base informações fornecidas, em executar ações sozinhos e de forma evolutiva, ou seja, suas decisões vão ficando cada vez mais exatas à medida em que vão aprendendo com mais informações. (Empresa C3)

ML é um subconjunto que está contido em AI. ML envolve o desenvolvimento de modelos matemáticos que permitem que computadores aprendam a partir de dados e façam previsões ou tomem decisões sem serem explicitamente programados para cada tarefa. IA é uma área mais geral que busca desenvolver algoritmos e modelos para realizar tarefas que normalmente requerem inteligência humana. (Empresa C4)

IA trata da criação de sistemas capazes de realizar tarefas que normalmente requerem inteligência humana, como raciocinar, aprender e reconhecer padrões. ML é um subcampo da IA focado no desenvolvimento de algoritmos que identificam padrões e inferem regras a partir dos dados, aprimorando sua performance conforme são expostos a mais informações. (Empresa LS1)

Inteligência Artificial (IA) é um campo amplo que engloba qualquer sistema ou máquina que pode realizar tarefas que normalmente exigem inteligência humana, como entender linguagem natural, reconhecer padrões, tomar decisões e resolver problemas. Machine Learning (ML) é uma subárea da IA que se concentra em ensinar as máquinas a aprender com dados. Em vez de serem programadas especificamente para realizar uma tarefa, as máquinas em ML aprendem a partir de exemplos e experiências. (Empresa LS2)

ML é uma aplicação de IA. O conceito de IA é amplo, sendo qualquer modelo, técnica ou ferramenta que seja capaz de aprender com dados e gerar resultados de maneira automática, sem interferência ou instruções humanas específicas (a não ser no viés da base utilizada como aprendizado). Já ML é um conjunto de modelos específicos de aprendizado de máquina. (Empresa BD1)

Além de endereçarem a resposta, também houve contribuições com relação a opinião profissional dos entrevistados a respeito da nomenclatura utilizada:

Atualmente chamar algo de IA tem um apelo maior do que ML para empresas e investidores. Cerca de 7 ou 10 anos atrás chamar de ML era muito mais apelativo do que de modelagem estatística. Nessa época responderia que o que separava ML de modelagem estatística é justamente um apelo mais computacional (visão computacional e modelos para dados não estruturados) para ML e mais para dados tabulares para modelagem estatística. Parece um ciclo. O conjunto de

técnicas são menos relevantes para a nomenclatura do que o impacto no mercado. (Empresa C2)

Após essa avaliação inicial a fim de assegurar que as respostas nas seções subsequentes estivessem fundamentadas em um entendimento abrangente dessas tecnologias, foi possível avançar para as seções seguintes.

Para apresentar os dados adquiridos a partir das respostas dos entrevistados, as Figuras a seguir seguem a seguinte legenda: Cada coluna representa uma empresa, as células assinaladas com “x” indicam que foi uma alternativa selecionada pela empresa e aquelas com “(*)” indicam que houve uma dissertação para endereçar as perguntas que não eram de múltipla escolha.

I. SEÇÃO 1: FOCO EM APLICAÇÕES NA EMPRESA

Pergunta 2: Dados são importantes para modelagem e ferramentas de ML e IA. Quais são as principais fontes de dados em seu negócio e quais os principais tipos de dados coletados?

Conforme Figura 20, todas as empresas entrevistadas utilizam dados coletados por pelo menos duas fontes distintas. As respostas indicam que dados públicos ou dados disponíveis *online* são uma fonte comum a todas as empresas, sendo que todas, com exceção da BD1, coletam dados internamente e todas, com exceção da C3, coletam dados de plataformas externas. Adicionalmente, nenhuma empresa adicionou novos tipos de dados que estariam faltando nas alternativas.

Dentre os tipos de dados coletados, foram selecionados em todas as respostas tanto dados estruturados, como Planilhas e Bancos de Dados Relacionais, quanto dados não-estruturados, como Texto da Web e documentos de texto não formatados (eg. word). Adicionalmente, 4 das empresas entrevistadas apontaram uso de dados não estruturados como Áudio e Imagens nas atividades de suas empresas. A fim de justificar ou exemplificar o uso de tais dados, foram obtidas as seguintes respostas:

Depende muito [do] objetivo das análises/modelos criados. Nossas principais aplicações são para: modelo de precificação de imóveis e crédito (dados tabulares obtidos interna e externamente), atendimento ao cliente (dados de áudio e textos não formatados) (Empresa C2)

Hoje os dados que utilizamos vem majoritariamente de fontes intrínsecas a nossa operação, contudo, temos uma infraestrutura para coleta de dados internos incluindo web scrapers e duas plataformas terceiras (Geoimovel e Geofusion). (Empresa LS1)

Para pesquisa de mercado, análises de viabilidade de empreendimentos, terrenos e rentabilidade, utilizamos base de dados internas e coletamos dados em bases públicas e de parceiros. (Ex: construção do estudo de preço de locação, utilizamos bases de dados [da nossa empresa] e monitoramos a oscilação de mercado para pareamento dos preços). (Empresa LS2)

Para a coleta de dados, utilizamos as seguintes fontes, que são provenientes de diversos lugares, como: [1] Entrata: Gestão de propriedade, valoração, e tudo que permeia o pré e pós-vendas. [2] Yardi: Controle financeiro. [2] Stripe: Gerenciadora de pagamentos. Exemplos de fontes de dados públicas incluem o site da Receita Federal para consulta de CPF, verificações de antecedentes com empresas privadas e afins. [Além disso, coletamos] anúncios de imóveis, fotos de satélite, legislação urbanística, entre outros. (Empresa LS3)

Coletamos anúncios de imóveis, fotos de satélite, legislação urbanística, entre outros. (Empresa BD1)

Dados são importantes para modelagem e ferramentas de ML e IA. Quais são as principais fontes de dados em seu negócio e quais os principais tipos de dados coletados?								
2.1 Fonte de Coleta dos Dados	C1	C2	C3	C4	LS1	LS2	LS3	BD1
Dados Coletados Internamente	X	X	X	X	X	X	X	
Dados Contratados de Plataformas Externas	X	X		X	X	X	X	X
Dados Públicos / Disponíveis Online	X	X	X	X	X	X	X	X
2.2 Tipos de Dados Coletados	C1	C2	C3	C4	LS1	LS2	LS3	BD1
Planilhas: Dados organizados em formato de planilha, como Microsoft Excel ou Google Sheets.	X	X	X		X	X	X	X
Bancos de Dados Relacionais: Fontes de dados organizadas em tabelas com esquemas predefinidos, como MySQL, PostgreSQL, SQL Server.	X	X	X	X	X	X	X	X
Texto da Web: Texto em páginas da web que não segue um formato específico.	X	X	X	X	X	X		X
2 Documentos de Texto Não Formatados: Texto em formato livre, como documentos Word não estruturados.	X			X				
Mídias Sociais: Dados de redes sociais, incluindo postagens, comentários, vídeos e imagens.	X							
E-mails: Conteúdo de e-mails, incluindo corpo do e-mail e anexos.								
Vídeos: Dados de vídeo não estruturados, como vídeos no YouTube ou Vimeo.								
Áudio: Gravações de áudio, como arquivos de podcast ou chamadas telefônicas.		X						
Imagens: Imagens não estruturadas, como fotos digitais ou imagens escaneadas.				X		X		X
Caso as fontes listadas acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações.								
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.		(*)			(*)	(*)	(*)	(*)

Figura 20 Pergunta 2 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.

Pergunta 3: Estou interessado em entender como o ML está sendo utilizado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente. Isso é algo que você/sua empresa estão utilizando? Se sim, por favor, descreva como as técnicas estão sendo aplicadas.

Conforme Figura 21, apenas a empresa C1 e LS3 indicam que não usam ML e todas as demais utilizam ML para, pelo menos, análise de dados. As respostas indicam que a maioria das empresas que responderam positivamente a esta pergunta também usam de ML para Coleta e Fornecimento de Dados e Análise Geo-espacial.

Adicionalmente, nenhuma das empresas sentiu a necessidade de adicionar novas maneiras que essa tecnologia estaria sendo utilizada, o que indica que as alternativas foram suficientes para contemplar as possíveis respostas.

A fim de justificar ou exemplificar as formas de aplicação do ML em cada empresa, foram obtidas as seguintes respostas:

[Para coleta e fornecimento de dados,] compramos dados de anúncios de fornecedores. Como geralmente se encontra mais de um anúncio para o mesmo imóvel, precisamos aplicar algumas técnicas para identificar essas duplicadas e remover da base. [Para análise de dados, avaliação de risco e análise geo-espacial,] um de nossos principais modelos de ML na companhia é nosso AVM (modelo de valoração automatizado). Ele é usado direta e indiretamente em diversas análises na companhia e é peça central em algumas decisões de alocação de recursos como marketing de performance, por exemplo. (Empresa C2)

Temos três exemplos de utilização de ML/IA na empresa hoje: [1] modelo de precificação de imóveis de acordo com as características dele, [2] modelo de indicação de imóveis similares com base em um outro imóvel e [3] modelo de identificação de imóveis “réplica” para evitar que pareçam ter ofertas duplicadas no nosso site. (Empresa C3)

Atualmente, usamos técnicas de Machine Learning para tentar precificar ativos, mas ainda não alcançamos um nível de precisão que permita a utilização dessas técnicas no dia a dia, permanecendo assim em fase experimental. (Empresa LS1)

Utilizamos ML para captura e análise dos dados desestruturamos de oferta e venda de imóveis, análise de *features* que os compradores mais buscam e previsão de preço de compra/venda. (Empresa LS2)

Atualmente, no Brasil, não utilizamos ML para a valoração de empreendimentos residenciais. Em vez disso, nossa metodologia baseia-se em diversas bases de dados, tanto públicas quanto privadas, mas a valoração é realizada através de regras fixas implementadas em planilhas Excel, o que não configura o uso de ML. Embora existam iniciativas globais e locais para incorporar diversas Inteligências Artificiais, como Copilot e Chat GPT, no processo de análise de dados, por enquanto, essas tecnologias serão utilizadas principalmente como ferramentas de suporte e análise de dados. Essa abordagem visa garantir precisão e confiabilidade enquanto exploramos as possibilidades futuras de integrar ML de forma mais abrangente em nossas operações. (Empresa LS3)

Utilizamos estes dados para construção de modelos de demanda, preços de imóveis e identificação de objetos em imagens/fotos de satélite. (Empresa BD1)

Estou interessado em entender como o ML está sendo utilizado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente.								
Isso é algo que você/sua empresa estão utilizando? Se sim, por favor, descreva como as técnicas estão sendo aplicadas.								
3.1 SIM, COMO?	C1	C2	C3	C4	LS1	LS2	LS3	BD1
COLETA E FORNECIMENTO DE DADOS		X	X	X		X		X
ANÁLISE DE DADOS		X	X	X	X	X		X
AVALIAÇÃO DE RISCO		X		X				X
PROCESSOS DA EMPRESA				X				
3 PROCESSAMENTO E GERAÇÃO DE "LINGUAGEM NATURAL"				X				
VISÃO COMPUTACIONAL				X				X
PLANEJAMENTO 3D/DE ESPAÇO								
ANÁLISE GEO-ESPACIAL		X		X	X	X		X
INTERNET DAS COISAS								
OUTROS								
Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.								
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.		(*)	(*)		(*)	(*)	(*)	(*)
3.2 NÃO	X						X	

Figura 21 Pergunta 3 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.

Pergunta 4: Caso a pergunta 3 tenha sido respondida com “Sim”, poderia informar quais algoritmos de ML estão sendo utilizados?

Conforme Figura 22, dentre as 7 empresas que utilizam algoritmos de ML, com exceção da empresa C3, que indicou, no espaço destinado a justificativa das respostas, não ter conhecimento específico sobre qual algoritmo é utilizado, os algoritmos mais utilizados são o LR, com 4 seleções, seguido de DT e (TE, RF, MC), ambos com 3 seleções e, finalmente, KNN, com duas seleções. Os algoritmos GMA e SVM tiveram apenas uma seleção, pela mesma empresa, BD1.

Adicionalmente, sobre as alternativas oferecidas, apenas a empresa C4 sentiu a necessidade de adicionar um novo algoritmo à listagem: “Utilizamos também Large Language Models (LLMs)¹⁰”, que não foi um algoritmo mapeado dentre os mais

¹⁰ “Os Modelos de Linguagem de Grande Escala (Large Language Models ou LLMs) são um tipo de modelo de Inteligência Artificial criado para entender e gerar texto.” (Fonte: <https://blog.dsacademy.com.br/o-que-sao-large-language-models-llms/>). Acessado em: Maio/2024

relevantes no nicho de valoração imobiliária, entretanto, foi listado nas perguntas 3 e 6 na alternativa “processamento e geração de linguagem natural”.

A fim de justificar ou exemplificar quais os algoritmos utilizados na aplicação do ML em cada empresa, foram obtidas as seguintes respostas:

Dividimos nosso modelo de precificação de imóveis em duas etapas. É muito importante que os preços fornecidos também sejam altamente justificáveis e fazemos isso através de imóveis e transações comparáveis ao imóvel precificado. Utilizamos métodos baseados em árvore para encontrar imóveis / transações similares e aplicamos uma regressão quantílica para chegar em uma curva de preços para os imóveis ao invés de uma estimativa pontual. (Empresa C2)

Atendimento ao cliente, busca e recomendação, análise de risco do inquilino. (Empresa C4)

LR é utilizada para modelos de precificação hedônica de imóveis residenciais para locação; DT são aplicadas em modelos de risco de inadimplência de clientes; e KNN é usado para mapear regiões com maior Development Yield na cidade de São Paulo. (Empresa LS1)

A utilização de uma biblioteca de algoritmos é bem particular para cada tipo de solução, a grande massa está dentro da regressão bem ajustada. (Empresa LS2)

Caso a pergunta 3 tenha sido respondida com "Sim", Poderia informar quais algoritmos de ML estão sendo utilizados ?								
	C1	C2	C3	C4	LS1	LS2	LS3	BD1
" Linear/ Logistic Regression" (LR) - Regressão Linear/ Logística		X		X	X	X		X
"Decision Trees" (DT) - Árvore de Decisão					X	X		X
"K-Nearest Neighbors" (KNN)					X			X
"Rule Based Learners" (RBL) - Aprendizado Baseado em Regras								
"General Additive Models" (GMA) - Modelos Aditivos Generalizados								X
"Bayesian Models" (BM) - Modelos Bayesianos								
4 Tree ensembles, random forest and multiple classifier systems" (TE, RF, MC)		X		X		X		
"Support Vector Machines" (SVM) - Máquinas de Vetores de Suporte								X
"Neural Networks" (NN) - Rede Neural				X				
"Multi-Layer Networks" (MLN) - Redes de Múltiplas Camadas				X				
"Recurrent Neural Networks" (RNN) - Redes Recorrentes				X				
"Convolutional Neural Networks" (CNN)				X				
Caso as fontes listadas acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações e destaque quais técnicas acima são aplicadas dentro do campo de valoração.				(*)				
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.		(*)	(*)	(*)	(*)	(*)		

Figura 22 Pergunta 4 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.

Pergunta 5: Caso a pergunta 3 tenha sido respondida com “Sim”, poderia informar quais algoritmos de ML estão sendo utilizados?

Conforme Figura 23, dentre as 7 empresas que utilizam algoritmos de ML, 6 aplicam nas áreas de análise de imóveis e 5 utilizam para detectar condições de mercado e análise de valor do imóvel. Apenas a empresa BD1 indica utilizar os algoritmos para tarefas relacionadas a legislação urbana e titulação de dispêndios.

A fim de justificar ou exemplificar como os algoritmos estão sendo utilizados em áreas específicas para aplicação do ML em cada empresa, obtivemos a resposta abaixo. Para as empresas C2 e LS1, os entrevistados julgaram que essa resposta já foi endereçada na pergunta anterior.

Aplicamos nosso AVM em diversas áreas da empresa: criação de índices para o mercado, alocação de recursos (e.g. investimento de marketing, ordenação da vitrine de imóveis) e trazendo transparência para o mercado (criamos relatórios de preços que podem ser usados por imobiliárias para facilitar disputas entre compradores e vendedores). (Empresa C2)

Caso a pergunta 3 tenha sido respondida com "Sim", poderia informar em quais áreas específicas tais tecnologias tem sido aplicadas na sua empresa?								
5.1 SIM. QUAIS?	C1	C2	C3	C4	LS1	LS2	LS3	BD1
Análise do imóvel (Localização, Vizinhança, Infra, Especificações Técnicas, etc)		X	X	X	X	X		X
Legislação Urbana (Zoneamento, etc)								X
Titulação e Dispêndios (Matrícula, Escritura, IPTU/ITR, Habite-se, etc)								X
5 Condições de Mercado (Oferta, Demanda, Preços Praticados, Tendências, etc)		X		X	X	X		X
Análise de Valor (Posicionamento do imóvel perante o mercado, Aplicações de Metodologias, Conclusão de Valor e etc.)		X		X	X	X		X
Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.								
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.		(*)						
5.2 NÃO	X						X	

Figura 23 Pergunta 5 – indicação de respostas respondidas por entrevistados. Fonte: Elaboração própria.

II. SEÇÃO 2: FOCO EM OPINIÃO GERAL DE MERCADO E IDENTIFICAÇÃO DE OPORTUNIDADES

Pergunta 6: Você está ciente de outras maneiras pelas quais o ML está sendo aplicado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente?

Conforme Figura 24, existe uma variação entre as respostas da pergunta 6 e 3, que possuem as mesmas alternativas. Enquanto na pergunta 3 questiona sobre maneiras em que o ML pode ser aplicado dentro da empresa, na pergunta 6 o foco é voltado para identificação de outras formas de utilização de ML, fora das empresas.

As respostas da pergunta 6 não apresentam um padrão de preferência entre os entrevistados quanto às áreas de aplicação: 4 dentre as 9 alternativas foram assinaladas por 3 empresas: “coleta e fornecimento de dados”, “análise de dados”, “avaliação de risco”, “análise geo-espacial” e “internet das coisas, seguida de “planejamento 3d/espço”, assinalada 2 vezes e “processamento de linguagem natural e visão computacional, selecionada apenas 1 vez.

Apesar de, na pergunta 3, as empresas C2 e LS1 indicarem que fazem uso de algoritmos dentro da empresa para atividades relacionadas a valoração de EBIs residenciais, na resposta à pergunta 6, elas mostram não ter ciência de outras maneiras pelas quais o ML está sendo aplicado. Todas as demais empresas indicam enxergar potencial de uso dessa tecnologia de outras maneiras, conforme listagem das alternativas de múltipla escolha.

A empresa C1, que, junto a empresa LS3, são as únicas entrevistadas que apontaram não utilizar ML em sua empresa, conforme resposta à pergunta 2, indica, por meio das alternativas assinaladas na pergunta 6, que consegue perceber diversos usos em que essa tecnologia poderia ser utilizada, entretanto, não foram oferecidos exemplos que pudessem justificar sua percepção.

A empresa C3, a fim de justificar ou exemplificar como os algoritmos estão sendo utilizados em áreas específicas para aplicação do ML, aponta um uso fora da esfera de valoração, que também é reconhecido pela empresa C4, na pergunta 4, quando menciona sobre o algoritmo LLM par atendimento ao cliente:

Processamento e geração de “linguagem natural”: uso de chats com IA generativa para atendimento ao cliente.

Internet das coisas: já soube de empresas que fornecem este serviço para o mercado imobiliário – elas possuem máquinas que identificam quando uma pessoa está passando por perto (através do aparelho celular) e utilizam isso como gatilho para disparar uma mensagem de SMS. Isso é utilizado no processo de venda em alguns lançamentos imobiliários. (Empresa C3)

A empresa C4 reforça a “visão computacional” como exemplo de outras maneiras em que o ML está sendo aplicado no nicho em questão, que também é uma alternativa assinalada na pergunta 3, sobre as maneiras que a empresa vem aplicando essa tecnologia dentro da própria empresa. Além dessa alternativa, a empresa C4 também indica conhecer exemplos de aplicação em “internet das coisas”. Entretanto, não foram oferecidos exemplos que pudessem justificar sua percepção.

A empresa LS2 reforça três maneiras em que o ML está sendo aplicado no nicho em questão, assinalando, tanto em resposta à pergunta 2 quanto 3: “coleta e fornecimento de dados”, “análise de dados” e “análise geo-espacial”. Além disso, a empresa LS2 também indica conhecer exemplos de aplicação em “planejamento 3d/de espaço”. Entretanto, não foram aprofundados exemplos em que pudessem justificar sua percepção, conforme indicado a seguir:

Vejo muitas empresas e soluções aplicando tecnologia em análises geo espaciais para monitorar projetos, pesquisas e traçar ICP. (Empresa LS2)

A empresa LS3, uma multinacional, apesar de informar que o ML não é aplicado em nenhum dos processos da empresa no Brasil, em resposta à pergunta 6, indica enxergar oportunidades em 5 dentre as 9 áreas listadas: “coleta e fornecimento de dados”, “análise de dados”, “avaliação de risco”, “processos da empresa” e “internet das coisas”. Em sua justificativa, que será apresentada a seguir, o entrevistado indica alguns exemplos de outras empresas que foram capazes de aplicar tecnologias de ML para a valoração de empreendimentos residenciais.

Alguns exemplos sobre como o ML está sendo integrado ao mercado imobiliário para aprimorar a precisão das avaliações, melhorar a eficiência das transações e fornecer insights mais profundos para investidores, compradores e vendedores:

[1] Modelos de Avaliação Automatizada (AVMs): [a] Zillow (Zestimate): Utiliza ML para analisar grandes quantidades de dados,

incluindo preços de venda anteriores, características do imóvel, localização, e tendências do mercado, para estimar o valor de uma propriedade. [b] Redfin Estimate: Semelhante ao Zestimate, o Redfin Estimate usa algoritmos de ML para fornecer estimativas de valor residencial baseadas em dados de vendas e características do imóvel.

[2] Análise de Sentimentos e Redes Sociais: [a] Reali: Analisa dados de redes sociais, avaliações de bairro, e comentários de usuários para ajustar as estimativas de valor dos imóveis com base na percepção pública e nas tendências sociais.

[3] Análise Preditiva para Investimentos: [a] HouseCanary: Utiliza ML para prever valores futuros de propriedades, analisando dados históricos, econômicos, demográficos e de mercado. Isso auxilia investidores a tomarem decisões mais informadas sobre compra e venda de imóveis.

[4] Plataformas de Gestão Imobiliária: [a] Compass: Usa ML para recomendar preços de listagem e prever a rapidez com que uma propriedade pode ser vendida, baseando-se em dados de mercado, características do imóvel e comportamento do comprador.

[5] Personalização de Ofertas: [a] Opendoor: Aplica ML para oferecer avaliações instantâneas e personalizadas de imóveis, ajustando automaticamente as ofertas de compra com base em fatores como condição do imóvel, localização e tendências de mercado.

[6] Análise Geoespacial: [a] CityBldr: Usa ML para avaliar o potencial de desenvolvimento de terrenos e propriedades, combinando dados geoespaciais com informações de zoneamento e infraestrutura.

[7] Plataformas de Crowdfunding Imobiliário: [a] Fundrise: Utiliza ML para avaliar riscos e retornos potenciais de investimentos em propriedades, ajudando a determinar quais empreendimentos são mais viáveis para financiamento coletivo.

[8] Predição de Preços de Aluguel: [a] RentHop: Usa ML para prever preços de aluguel, considerando fatores como localização, proximidade de transporte público, e características do imóvel. (Empresa LS3)

A empresa BD1 reforça três maneiras em que o ML está sendo aplicado no nicho em questão, assinalando, tanto em resposta à pergunta 2 quanto 3: “coleta e fornecimento de dados”, “análise de dados” e “análise geo-espacial”. Além disso, a empresa BD1 também indica conhecer exemplos de aplicação em “planejamento 3d/de espaço”. Entretanto, não foram aprofundados exemplos em que pudessem justificar sua percepção.

Você está ciente de outras maneiras pelas quais o ML está sendo aplicado no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária atualmente?																		
6.1 SIM, QUAIS?	C1	C2	C3	C4	LS1	LS2	LS3	BD1		C1	C2	C3	C4	LS1	LS2	LS3	BD1	
COLETA E FORNECIMENTO DE DADOS	X					X	X	X	6		X	X	X		X		X	
ANÁLISE DE DADOS	X					X	X	X			X	X	X	X	X	X		X
AValiação DE RISCO	X							X			X		X					X
PROCESSOS DA EMPRESA	X							X					X					
PROCESSAMENTO E GERAÇÃO DE "LINGUAGEM NATURAL"	X		X										X					
VISÃO COMPUTACIONAL	X			X									X					X
PLANEJAMENTO 3D/DE ESPAÇO						X		X										
ANÁLISE GEO-ESPACIAL				X		X		X			X		X	X	X			X
INTERNET DAS COISAS	X		X					X										
OUTROS																		
Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.																		
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.			(*)			(*)	(*)			(*)	(*)		(*)	(*)	(*)	(*)		
6.2 NÃO		X			X					X							X	

Figura 24 Pergunta 6 e Pergunta 3, respectivamente – indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.

Pergunta 7: Existem áreas específicas no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária que você acredita serem boas oportunidades para aplicar técnicas de ML ou IA?

Conforme Figura 25, existe uma variação entre as respostas da pergunta 7 e 5, que possuem as mesmas alternativas. Enquanto na pergunta 5 o foco é voltado para as formas de utilização de ML dentro das empresas, a pergunta 7 explora o cenário de quais outras oportunidades os entrevistados conhecem e que podem ser usos de ML dentro do mercado residencial de valoração de EBIs.

Focando nas respostas da pergunta 7, dentre as 8 empresas entrevistadas, todas assinalaram “análise de valor” e “condições de mercado”. “Análise de imóvel” foi também uma das áreas mais assinaladas, com 7 votos, e, em seguida, “titulação e dispêndios” teve 3 votos. Apesar de BD1 ter sido a única empresa que indicou aplicar ML em “legislação urbana”, apenas a empresa C1 reconheceu as oportunidades para aplicação de ML nessa área.

A empresa C1, apesar de informar que o ML não é aplicado em nenhum dos processos da empresa, na resposta à pergunta 7, indica enxergar diversos usos em potencial para o M, entretanto, não foram oferecidos exemplos que pudessem justificar sua percepção.

A empresa C2 sinalizou as mesmas alternativas tanto na pergunta 5 quanto na 7, entretanto, indicou, em justificativa, que enxerga potencial nas demais opções,

“legislação urbana” e “titulação e dispêndios”, entretanto, não soube especificar em exemplos como os algoritmos poderiam ser usados nesse contexto:

Não conheço muito sobre Legislação Urbana e Titulação e Dispêndios, mas acredito que no contexto que usamos na [empresa] faz bastante sentido. (Empresa C2)

A empresa C3 reforçou uma das maneiras com que o ML pode ser aplicado, marcando o item “análise de imóvel” em ambas as perguntas, 5 e 7. Além disso, também selecionou “condições de mercado” e “análise de valor” como áreas em potencial para aplicação de ML. A fim de ilustrar sua resposta, foram trazidos os exemplos a seguir:

Acho que tem muito potencial no seguinte contexto: a análise de dados relacionados a demanda para sugestão de compra de terreno, definição de configuração de produto, sugestão de público ideal. (Empresa C3)

A empresa C4 reforçou três das maneiras com que o ML pode ser aplicado, marcando o item “análise de imóvel”, “condições de mercado” e “análise de valor” em ambas as perguntas, 5 e 7. Além disso, também selecionou “titulação e dispêndios” como uma área em potencial para aplicação de ML, entretanto, não foram oferecidos exemplos que pudessem justificar sua percepção.

As empresas LS1 reforçou as mesmas áreas tanto na pergunta 7 quanto na 5: “análise do imóvel”, “análise de valor” e “condições de mercado” e, em sua justificativa, destacou a precificação de ativos como forma de ilustrar sua escolha. A seguir, justificativa compartilhada pela empresa:

Na minha opinião, a precificação de ativos é a principal oportunidade para a aplicação de técnicas de Machine Learning e Inteligência Artificial. (Empresa LS1)

A empresa LS2 sinalizou as mesmas alternativas tanto na pergunta 5 quanto na 7 e, em sua justificativa, não apontou maneiras específicas de aplicação nessas áreas, apenas destacando a dificuldade de disseminação dessas tecnologias no mercado de Real Estate em geral:

A Aplicação de AI e ML dentro do setor ainda é pouco disseminada no mercado de Real Estate, na [nossa empresa], tentamos mudar isso e dar acesso às incorporadoras a esse tipo de tecnologia (que tem mão de obra e custo computacional elevados). (Empresa LS2)

A empresa LS3, apesar de informar que o ML não é aplicado em nenhum dos processos da empresa, na resposta à pergunta 7, indica enxergar oportunidades nas áreas “análise de imóvel”, “condições de mercado” e “análise de valor”. Em sua justificativa, que será apresentada a seguir, o entrevistado indica possuir conhecimento amplo sobre o tema, dando diversos exemplos. Conforme foi indicado na pergunta 5, a empresa não aplica ML em suas unidades no Brasil, entretanto, dá a entender que em outras filiais essa tecnologia já é utilizada.

Existem várias áreas específicas no mercado imobiliário de valoração de empreendimentos residenciais onde as técnicas de ML ou IA podem ser especialmente benéficas. Aqui estão três dos melhores exemplos:

[1] **Análise Preditiva de Preços Imobiliários:** Prever com precisão os preços futuros dos imóveis com base em uma vasta gama de dados históricos e atuais. Aplicação de ML/IA: Utilizar algoritmos de ML para analisar dados como preços de venda anteriores, características do imóvel, localização, condições econômicas e demográficas. Técnicas como regressão linear, árvores de decisão e redes neurais podem ser empregadas para criar modelos preditivos robustos. Benefícios: Aumenta a precisão das avaliações, auxilia investidores e compradores a tomarem decisões mais informadas, e melhora a capacidade das empresas de ajustar suas estratégias de mercado.

[2] **Análise de Sentimentos e Dados Sociais:** Incorporar dados de redes sociais, avaliações online e outras fontes de feedback do usuário para ajustar as avaliações de imóveis. Aplicação de ML/IA: Usar processamento de linguagem natural (NLP) e análise de sentimentos para interpretar dados textuais de redes sociais, fóruns, e avaliações online. Algoritmos de clustering e classificação podem ajudar a identificar tendências e sentimentos predominantes sobre diferentes bairros e propriedades. Benefícios: Proporciona uma compreensão mais holística do valor percebido dos imóveis, ajudando a capturar fatores intangíveis que influenciam a demanda e o valor das propriedades.

[3] **Otimização de Portfólios de Investimento Imobiliário:** Maximizar os retornos de investimentos em portfólios imobiliários através da identificação de padrões e oportunidades de mercado. Aplicação de ML/IA: Implementar modelos de ML para analisar grandes volumes de dados financeiros, de mercado e de propriedade para identificar as melhores oportunidades de investimento. Técnicas como aprendizado por reforço podem ser usadas para otimizar estratégias de investimento de forma dinâmica. Benefícios: Melhora a alocação de ativos, reduz riscos, e aumenta os retornos sobre investimentos imobiliários através de decisões baseadas em dados.

[Em resumo,] essas áreas apresentam excelentes oportunidades para a aplicação de ML e IA permitindo avanços significativos na precisão das avaliações, na compreensão das dinâmicas de mercado e na otimização de investimentos no setor imobiliário. (Empresa LS3)

A empresa BD1 escolheu reforçar três áreas em que enxergam oportunidades para aplicação de técnicas de ML no mercado de valoração de EBIs residenciais, assinalando, tanto na pergunta 7 quanto na 3, as seguintes áreas: “titulação e dispêndios”, “condições de mercado” e “análise de valor”. Apesar de terem indicado que as áreas “análise de imóvel” e “legislação urbana” tem aplicação de ML dentro da empresa, a empresa LS1 não identificou maiores oportunidades nessas duas áreas. Não foram oferecidos exemplos que pudessem justificar sua percepção.

Existem áreas específicas no mercado imobiliário de valoração de Empreendimentos Residenciais de Base Imobiliária que você acredita serem boas oportunidades para aplicar técnicas de ML ou IA ?																		
7.1 SIM. QUAIS?	C1	C2	C3	C4	LS1	LS2	LS3	BD1		C1	C2	C3	C4	LS1	LS2	LS3	BD1	
Análise do imóvel (Localização, Vizinhança, Infra, Especificações Técnicas, etc)	X	X	X	X	X	X	X		7		X	X	X	X	X		X	
Legislação Urbana (Zoneamento, etc)	X																	X
Titulação e Dispêndios (Matrícula, Escritura, IPTU/ITR, Habite-se, etc)	X			X				X										X
Condições de Mercado (Oferta, Demanda, Preços Praticados, Tendências, etc)	X	X	X	X	X	X	X	X		5		X		X	X	X		X
Análise de Valor (Posicionamento do imóvel perante o mercado, Aplicações de Metodologias, Conclusão de Valor e etc.)	X	X	X	X	X	X	X	X				X		X	X	X		
Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.																		
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.			(*)	(*)		(*)	(*)	(*)				(*)						
7.2 NÃO										X							X	

Figura 25 Pergunta 7 e Pergunta 5, respectivamente – indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.

Pergunta 8: Que tipos de dados podem estar faltando ou ser úteis para aprimorar a análise ou aplicações de aprendizado de máquina e IA no mercado de valoração de Empreendimentos Residenciais de Base Imobiliária?

Conforme Figura 26, existe uma variação entre as respostas da pergunta 8 e 2, que possuem as mesmas alternativas. Enquanto na pergunta 2.2 o foco é voltado para os tipos de dados coletados dentro das empresas, a pergunta 8 explora cenário em que os entrevistados devem sinalizar, dentro desse mesmo universo de opções, quais dados podem estar faltando ou ser úteis para aprimorar o uso de ML dentro do mercado residencial de valoração de EBIs.

Os tipos de dados mais destacados na pergunta 8 foram: Bancos de dados relacionais, com 6 votos, seguido por planilhas, com 4 votos e documentos de texto e mídias sociais, com 2 votos cada.

Na resposta à pergunta 8, a empresa C1 expande as alternativas e, conforme apontado na pergunta 2, apesar de fazer uso tanto de dados estruturados, como Planilhas e Bancos de Dados Relacionais, quanto dados não-estruturados, como dados não-estruturados, como Texto da Web, documentos de texto não formatados e mídias sociais; na pergunta 8 são acrescentados os demais tipos de dados, tendo todas as alternativas assinaladas e indicando a visão da empresa C1 sobre potencialmente fazer uso dos mais diversos tipos de dados. Entretanto, não foram oferecidos exemplos que pudessem justificar sua percepção.

A empresa C2 reforçou, na pergunta 8, o uso de banco de dados relacionais, que também foi assinalado na pergunta 2, juntamente com outros dois tipos de dados: planilhas e texto na web. Para justificar sua escolha, o entrevistado faz menção aos dados de valores de transação, que ainda não são disponibilizados pela prefeitura, conforme indicado a seguir:

Temos um mercado ainda com muita assimetria de informação. Todo mundo se beneficiaria com a digitalização dos cartórios e disponibilização de dados de transação pelas prefeituras. Algumas prefeituras já fazem esse trabalho, outras demandam processos onerosos e só fornecem dados a partir de pedidos (por LAI) e outras julgam esses pedidos improcedentes. A federalização desse processo poderia ser uma boa saída. (Empresa C2)

A empresa C3 indica, conforme justificativa exposta a seguir, que não enxerga a necessidade de mais dados além dos que ela já tem conhecimento e utiliza – de acordo com respostas à pergunta 2, a empresa utiliza dados de planilhas, bancos relacionais e texto na web.

Não sinto que faltam dados. (Empresa C3)

A empresa C4 é a única que reconheceu dados de imagem como um tipo de dado que pode estar faltando para aprimorar as análises ou aplicações de ML no mercado em questão, entretanto, em sua justificativa, não exemplificou o porquê de sua escolha, mas sim, apontou que sente deficiências importantes no acesso a dados rotulados, que se encaixariam nas alternativas “planilha” e “banco de dado relacional”.

Para mim, o principal problema não é somente a falta de dados, mas sim a falta de dados rotulados. Para precificação de imóveis, é muito difícil de se obter dados de compra ou aluguel de residências. (Empresa C4)

A empresa LS1 seguiu a mesma linha de raciocínio da empresa C2, tanto nas alternativas – reforçando, na pergunta 8, o uso de banco de dados relacionais, que também foi

assinalado na resposta à pergunta 2, quanto na justificativa, apontando a necessidade de haver um avanço na disponibilização de dados transacionados e, no caso da empresa LS1, sendo bastante clara sobre os dados de locação e não apenas de venda, conforme é indicado a seguir.

Atualmente, não temos dados estruturados sobre valores fechados de locação, como temos no mercado de venda de imóveis. Esta é a informação que mais falta atualmente. No entanto, não devemos esperar uma mudança nesse cenário a curto prazo, pois o mercado de locação é muito fragmentado. (Empresa LS1)

A empresa LS2 reforçou, na pergunta 8, o uso de planilhas, banco de dados relacionais, que também foram assinalados na pergunta 2, além de incluir mídias sociais como tipos de dado que podem estar faltando para aprimorar as análises ou aplicações de ML no mercado em questão. Para justificar sua escolha, o entrevistado chamou atenção não somente para a informação que o dado está relacionado (oferta de venda e locação), como também a localização que esses dados fazem menção (a maioria no eixo SP/RJ):

Dados de oferta de venda e locação estruturados, pesquisas com maior capilaridade regional sobre o mercado imobiliário - a maior parte dos dados disponíveis estão concentrados no eixo SP/RJ. (Empresa LS2)

A empresa LS3 reforçou, na pergunta 8, o uso de planilhas, banco de dados relacionais, que também foram assinalados na pergunta 2, além de incluir texto da web e mídias sociais como tipos de dado que podem estar faltando para aprimorar as análises ou aplicações de ML no mercado em questão. Em sua justificativa, o entrevistado explica que a inclusão de diferentes tipos dados influencia na melhoria da precisão das avaliações, entretanto, acaba focando mais nas informações adquiridas pelos dados e não nos tipos de dado em si e como extraí-los:

Dados Detalhados sobre o Imóvel: Informações detalhadas sobre o estado de conservação, idade do imóvel, renovações e reformas realizadas (Dados sobre características específicas, como eficiência energética, sistemas de aquecimento e refrigeração, e materiais de construção utilizados);

Dados de Infraestrutura e Serviços Locais: Localização e distância de escolas, hospitais, parques, e centros comunitários (Dados sobre proximidade e qualidade do transporte público, acessos a estradas principais e trânsito local);

Dados Econômicos e Demográficos: Dados sobre emprego, renda média, taxas de desemprego, e crescimento econômico da região (Informações

sobre a composição demográfica, incluindo idade, nível de educação, e tamanho médio das famílias);

Dados de Mercado Imobiliário: Registros detalhados de transações imobiliárias anteriores, incluindo preços de venda, datas de transação, e comparativos de vendas (Dados sobre inventário de imóveis, tempo médio no mercado, e flutuações sazonais nos preços dos imóveis);

Dados de Sentimentos e Avaliações Públicas: Análises de opiniões de residentes e compradores sobre o bairro e os imóveis, coletadas de redes sociais e sites de avaliações. (Dados sobre índices de criminalidade e percepção de segurança na área);

Dados de Condições Ambientais: Informações sobre a qualidade do ar, níveis de poluição, e presença de áreas verdes (Dados sobre a vulnerabilidade a desastres naturais, como enchentes, terremotos e incêndios florestais);

Dados de Políticas Públicas e Regulamentações: Informações sobre leis de zoneamento, permissões de construção, e regulamentações locais que podem afetar o valor da propriedade (Dados sobre incentivos fiscais, subsídios e outros benefícios governamentais que podem impactar o mercado imobiliário)

Dados de Análise de Uso do Solo: Dados sobre a utilização atual do solo, como áreas residenciais, comerciais, industriais e agrícolas (Informações sobre planos futuros de desenvolvimento urbano e expansão de infraestrutura). (Empresa LS3)

A empresa BD1 reforçou, na pergunta 8, o uso de planilhas e banco de dados relacionais, que também foram assinalados na pergunta 2, como tipos de dados que podem estar faltando para aprimorar as análises ou aplicações de ML no mercado em questão. Para justificar sua escolha: Para justificar sua escolha, assim como as empresas C1 e LS1, mais uma vez a falta de acesso aos dados de transações de imóvel foi apontado:

Informações de vendas reais de imóveis, com preços reais transacionados. Informações do ITBI são insuficientes, pois carecem de atributos importantes dos imóveis, tais como área privativa, padrão, número de vagas etc. (Empresa BD1)

Que tipos de dados podem estar faltando ou ser úteis para aprimorar a análise ou aplicações de aprendizado de máquina e IA no mercado de valoração de Empreendimentos Residenciais de Base Imobiliária ?																
	C1	C2	C3	C4	LS1	LS2	LS3	BD1	C1	C2	C3	C4	LS1	LS2	LS3	BD1
Planilhas: Dados organizados em formato de planilha, como Microsoft Excel ou Google Sheets.	X					X	X	X	X	X	X		X	X	X	X
Bancos de Dados Relacionais: Fontes de dados organizadas em tabelas com esquemas predefinidos, como MySQL, PostgreSQL, SQL Server.	X	X			X	X	X	X	X	X	X	X	X	X	X	X
Texto da Web: Texto em páginas da web que não segue um formato específico.	X						X		X	X	X	X	X	X		X
Documentos de Texto Não Formatados: Texto em formato livre, como documentos Word não estruturados.	X								X			X				
8 Mídias Sociais: Dados de redes sociais, incluindo postagens, comentários, vídeos e imagens.	X					X	X		X							
E-mails: Conteúdo de e-mails, incluindo corpo do e-mail e anexos.	X															
Vídeos: Dados de vídeo não estruturados, como vídeos no YouTube ou Vimeo.	X															
Áudio: Gravações de áudio, como arquivos de podcast ou chamadas telefônicas.	X								X							
Imagens: Imagens não estruturadas, como fotos digitais ou imagens escaneadas.	X			X							X		X		X	
Caso os tipos de dados listados acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações.		(*)	(*)	(*)	(*)	(*)	(*)	(*)								

Figura 26 Pergunta 8 e Pergunta 2, respectivamente - indicação de respostas respondidas por entrevistados e comparação entre respostas. Fonte: Elaboração própria.

5 ANÁLISE DOS DADOS

Antes de analisar os dados sob a perspectiva dos objetivos primários deste trabalho, é importante destacar as conclusões sobre qual o nível de aderência do uso de tecnologias de ML na amostragem de empresas selecionadas. Para isso, criou-se, para esse trabalho, uma divisão em 4 níveis:

Nível 1. Empresas que já utilizam dados para processos da empresa em geral;

Nível 2. Empresas que conhecem o conceito de IA e ML;

Nível 3. Empresas que veem potenciais de aplicação de tecnologias de IA e ML no seu nicho de EBI residencial ou em outros nichos dentro de Real Estate, sendo capazes de enxergar oportunidades;

Nível 4. Empresas que já aplicam ML.

Dentro da amostra, tivemos os seguintes resultados:

Nível 1. Empresas que já utilizam dados para processos da empresa em geral

(8 das 8 empresas atingiram o Nível 1)

Conforme as respostas à pergunta 2, todas as empresas entrevistadas utilizam dados em suas operações, indicando a importância tanto de [1] dados estruturados (por exemplo, planilhas e bancos de dados relacionais), nos quais ambos os tipos de dados foram selecionados majoritariamente - com a exceção da empresa C4, que não utiliza planilhas - quanto de [2] dados não estruturados (por exemplo, texto na web, documentos de texto não formatados, mídias sociais, áudios e imagens), em que “texto na web” foi o mais selecionado, com 7 dentre 8 votos, seguido de “imagens”, com 3 votos e “documentos não formatados” com 2 votos.

Esse resultado indica que existe grande potencial para o aproveitamento de dados não-estruturados, que ainda são pouco utilizados.

Portanto, entendendo que o uso de dados é o primeiro passo para avaliar como processar esses dados por meio de tecnologias como ML, por exemplo, todas as empresas atingiram o Nível 1 de aderência.

Nível 2. Empresas que conhecem o conceito de IA e ML

(8 das 8 empresas atingiram o Nível 2)

Com base nas respostas à pergunta 1, todas as empresas entrevistadas demonstraram um entendimento preciso de ML e IA, alinhado com as definições deste estudo. Isso sugere que esses conceitos não são estranhos a eles, o que pode indicar a possibilidade para que, futuramente, oportunidades de implementação dessas tecnologias sejam identificadas em seus respectivos setores.

Nível 3. Empresas que veem potenciais de aplicação de tecnologias de IA e ML no seu nicho de EBI residencial ou em outros nichos dentro de Real Estate, sendo capazes de enxergar oportunidades

(6 das 8 empresas atingiram o Nível 3)

Conforme respostas às perguntas 6, 7 e 8, todas as empresas assinalaram alternativas que indicavam ou terem conhecimento de outras aplicações dentro no nicho de valoração de EBIs residenciais que estivessem fora da empresa em que trabalham ou serem capazes de identificar oportunidades para que tecnologias de ML pudessem ser aplicadas em seu nicho, tanto em relação a tipos de dados (pergunta 8) quanto a áreas específicas no mercado de valoração de EBIs residenciais (pergunta 6 e 7). Isso indica que o conceito de ML tem seu valor percebido entre as empresas.

Entretanto, conforme Figura 27, somente 6 das 8 empresas foram capazes de dissertar sobre suas escolhas ao menos em uma das alternativas, indicando maior conhecimento sobre como efetivamente o ML poderia contribuir para seu nicho, conforme legenda apresentada a seguir: (*) Não conseguiu justificar sua resposta; (**) Tentou justificar sua resposta, apontando alguns exemplos, sem muitos detalhes, ou não apontou maneiras específicas de aplicação, mas destacou pontos relevantes à área em estudo (eg. desafios, deficiências etc.); (***) conseguiu justificar sua resposta com detalhes.

Cód. Empresa	Pergunta 6	Pergunta 7	Pergunta 8
C1	(*)	(*)	(*)
C2	(*)	(*)	(***)
C3	(*)	(**)	(*)
C4	(*)	(*)	(***)
LS1	(*)	(***)	(***)
LS2	(*)	(**)	(***)
LS3	(***)	(***)	(**)
BD1	(*)	(*)	(***)

Figura 27 Classificação quanto a qualidade das respostas às perguntas 6, 7 e 8. Fonte: Elaboração própria.

Detalhando a identificação de oportunidades, para essas 6 empresas, da pergunta mais genérica para a mais específica:

a. Pergunta 8: Tipos de dados podem estar faltando ou serem úteis para aprimorar a análise ou aplicações de ML e IA no mercado de valoração de EBIs Residenciais:

As respostas de múltipla escolha dessa essa pergunta se concentraram na alternativa “Bancos de dados relacionais”, com apenas uma exceção, que foi a empresa C2, indicando não sentir falta de mais dados para ter esse aprimoramento.

Essa escolha indica que existem grandes oportunidades para que essas empresas consigam visualizar os benefícios do uso de dados não estruturados para aprimorar os seus modelos, conforme mostrado no Capítulo 3 nas pesquisas: (1) de Baur (2023), com a inclusão da descrição textual das propriedades junto aos dados com as características dessas propriedades de modo a auxiliar na arbitragem do preço de venda a partir do uso de alguns modelos de ML, que foram testados para medir qual era o mais eficiente, e (2) de Potrawa (2022), conseguindo aumentar a precisão de modelos hedônicos adicionando, além de dados estruturados, imagens das propriedades, coletadas de sites de aluguel e do Google Maps, para extrair variáveis adicionais, como a presença de uma vista da

cidade, a proximidade de áreas verdes ou corpos d'água, e a qualidade da insolação, por exemplo.

Ainda assim, entende-se que a escolha por dados neste primeiro nível, estruturado, se dá devido aos desafios destacados nas justificativas dessas empresas, que apontam a barreira documental que existe no Brasil, conforme apontado pelas empresas C2, C4, LS1 e LS2. Dentre os depoimentos, destaca-se a dificuldade de acesso a dados de transação para venda e locação.

No primeiro caso, sobre as informações de venda, é apontado como fator determinante o processo de digitalização dos cartórios e, portanto, acesso a dados de transação como valores reais transacionados e atributos tais como área privativa, número de vagas etc. No segundo caso, sobre as informações de locação, o acesso é ainda pior quando comparada às transações de venda e um dos motivos é a característica desse mercado no Brasil, que é bastante fragmentado e, conforme apontado pela empresa LS1, isso dificulta a estruturação de uma base de dados.

Dentre todas as empresas entrevistadas, a LS3 foi a única capaz de listar diversos tipos de dados que pudessem influenciar na melhoria da precisão das avaliações, sendo a empresa que se mostrou mais consolidada no Nível 3 (Empresas que veem potenciais de aplicação de tecnologias de IA e ML no seu nicho de EBI residencial ou em outros nichos dentro de Real Estate, sendo capazes de enxergar oportunidades).

b. Pergunta 6: Maneiras pelas quais o ML está sendo aplicado no mercado de valoração de EBIs Residenciais atualmente:

Somente a empresa LS3 forneceu exemplos que pudessem justificar suas respostas, apresentando diversos estudos de caso em que o ML estava sendo aplicado em outras empresas, sendo a empresa que se mostrou mais consolidada no Nível 3 (Empresas que veem potenciais de aplicação de tecnologias de IA e ML no seu nicho de EBI residencial ou em outros nichos dentro de Real Estate, sendo capazes de enxergar oportunidades).

As respostas de múltipla escolha dessa pergunta foram bastante variadas e não apresentam um padrão de preferência entre os entrevistados quanto às áreas de aplicação. Dentre as 6 empresas classificadas no Nível 3, 2 empresas (C2 e LS1) indicaram não estarem cientes de outras maneiras de aplicação. As alternativas mais selecionadas foram, com 3 votos: “coleta e fornecimento de dados”, “análise de dados” e “análise geo-espacial”, seguida de “planejamento 3d/espaco”, assinalada 2 vezes e “avaliação de risco”, “processo da empresa” e “internet das coisas”, selecionada apenas 1 vez.

c. Pergunta 7: Área específicas dentro do mercado de valoração de EBIs Residenciais que são oportunidades para aplicar técnicas de ML ou IA:

Dentre as áreas selecionadas pelas alternativas de múltipla escolha, os destaques foram “condições de mercado” e “análise de valor”, com todos os votos, “análise do imóvel”, com a única exceção sendo a empresa BD1, que não selecionou essa alternativa, e “titulação e dispêndios” com 2 votos.

As justificativas variaram consideravelmente entre as empresas, sendo que duas empresas foram capazes de explicar de maneira mais detalhada as suas escolhas (LS1 e LS3) e duas empresas tiveram respostas que conseguiram tocar em pontos críticos relacionados ao tema, porém sem tantos detalhes (C3 e LS2).

A empresa LS1 foi enfática sobre entender que “a precificação de ativos é a principal oportunidade para a aplicação de técnicas de ML e IA” e a empresa LS3 foi capaz de apresentar exemplos para melhorar a precisão das avaliações em algumas áreas do mercado como precificação de ativos e otimização de portfólios de investimento imobiliário.

Enquanto isso, a empresa C3 indicou aplicações mais genéricas que são aplicáveis a processos de empresas em geral, como “análise de demanda para sugestão de compra de terreno, definição de configuração de produto

[e] sugestão de público ideal”, e a empresa LS2 não apontou maneiras específicas de aplicação nessas áreas, porém indicou algumas dificuldades para a disseminação dessas tecnologias.

Sendo assim, mais uma vez, fica claro que existe uma grande oportunidade para que o mercado se aprimore para fazer mais relações entre os algoritmos e as áreas específicas do mercado de valoração de EBIs residenciais, conforme mostrado no Capítulo 3 nas pesquisas evidenciadas a seguir, que testam modelos para descobrir relações complexas que não poderiam ser capturadas por modelos convencionais: (1) Linchuang (2021), a partir da coleta de dados de imóveis em áreas dentro de 1,5 km do corredor BRT e a aplicação do algoritmo GBDT conseguiu examinar a relação não-linear entre o BRT e os preços dos imóveis, e (2) Viana et al. (2017), que utilizou métodos de ML para análise da taxa de rentabilidade de empreendimentos imobiliários residenciais para locação de curto prazo a partir da detecção da importância das variáveis independentes em cada algoritmo (eg. preço do imóvel, diária Média, localização, área (tamanho do imóvel) etc.).

Nível 4. Empresas que já aplicam ML

(5 das 8 empresas atingiram o Nível 4)

Conforme respostas às perguntas 3, 4 e 5, apesar de 6 dos 8 entrevistados terem indicado aplicar ML dentro de suas empresas, a empresa C3 não conseguiu indicar quais os algoritmos utilizados. Dentre as empresas, a C4 é a que utiliza maior quantidade de algoritmos, com 7 algoritmos, 6 deles listados na alternativa de múltipla escolha e 1 (Large Language Models (LLMs)) adicionado pelo entrevistado. BD1 é a segunda empresa com maior uso de algoritmos, com 5, seguida pela LS1 e LS2, com 3 algoritmos, e, finalmente, pela C2, com 2 algoritmos.

Com 5 votos, o algoritmo mais utilizado foi o de regressão linear (“LR”), seguido pela árvore de decisão (“DT”) e (TE, RF, MC), com 3 votos e o K-vizinhos mais próximos (“K-Nearest Neighbors” - KNN), com 2 votos.

Dentre as aplicações de tais algoritmos, foram destacados a coleta de dados, inclusive com a utilização de web scrapers, valoração automatizada, indicação de imóveis similares com base imóvel modelo e identificação de objetos em imagens/fotos de satélite. Isso indica que parte importante das contribuições da aplicação de ML identificadas nessa pesquisa estão sendo aplicadas nas empresas.

Adicionalmente, seguindo outra perspectiva para análise de resultados, a fim de comparar os algoritmos levantados na bibliografia (capítulo 2) como os mais relevantes para esse nicho, com [a] aqueles utilizados nos estudos de caso (capítulo 3) e [b] os aplicados nas empresas entrevistadas, a Figura 28 foi criada para identificar se existia algum padrão.

Algoritmos (Capítulo 2)	[a]	[b]
"Linear/ Logistic Regression" (LR) - Regressão Linear/ Logística	3	5
"Decision Trees" (DT) - Árvore de Decisão	3	2
"K-Nearest Neighbors" (KNN)	1	2
"Rule Based Learners" (RBL) - Aprendizado Baseado em Regras	1	0
"General Additive Models" (GMA) - Modelos Aditivos Generalizados	0	1
"Bayesian Models" (BM) - Modelos Bayesianos	0	0
Tree ensembles, random forest and multiple classifier systems" (TE, RF, MC)	4	3
"Support Vector Machines" (SVM) - Máquinas de Vetores de Suporte	4	1
"Neural Networks" (NN) – Rede Neural	3	1
"Multi-Layer Networks" (MLN) – Redes de Múltiplas Camadas	0	1
"Recurrent Neural Networks" (RNN) - Redes Recorrentes	0	1
"Convolutional Neural Networks" (CNN)	1	1

Figura 28 Indicação de quantidade algoritmos utilizados nos estudos de caso citados no capítulo 3 e citados pelas empresas entrevistadas no questionário aplicado, respectivamente. Fonte: Elaboração própria.

Analisando a Figura 28, somente o algoritmo "General Additive Models" (GMA) não foi quantificado nas aplicações levantadas, o que indica que a maioria dos algoritmos exemplificados no capítulo 2 foram relevantes para compreender os estudos de caso e respostas apontadas em entrevista. Ainda assim, como um dos objetivos secundários deste trabalho é o de conduzir um levantamento abrangente de algoritmos de ML aplicados à arbitragem de valor para transações imobiliárias de ativos residenciais para o público do

mercado imobiliário, o levantamento e explicação do algoritmo GMA teve fins didáticos e de disseminar mais conhecimentos sobre o tema.

Ainda que não tenha ficado evidente um padrão entre os algoritmos mais utilizados nos estudos e caso (capítulo 3) e nos citados pelas empresas entrevistadas, é possível detectar que os algoritmos mais relevantes quando quantificados entre [a] e [b] são: LN, (TE, RF, MC) e DT.

Finalmente, cumprindo outro objetivo secundário desta pesquisa a fim de comparar os resultados obtidos em entrevistas com profissionais de mercado do nicho de Valoração com atuação em São Paulo, com aqueles, também do nicho de valoração, encontrados na pesquisa de mestrado de Conway (2018), a Figura 29 consegue evidenciar um padrão entre as respostas a partir da indicação de [a] quantidade de votos em cada alternativa e [b] proporção entre quantidade de votos e número de entrevistados.

Dentre ambas as pesquisas, os principais campos de atuação são “coleta de dados”, “análise de dados” e “análise geo-espacial”. Nesse trabalho foi possível englobar maior número de entrevistados dentro da área de valoração e, conforme apontam dos resultados na Figura 29, o mercado em São Paulo em 2024 representa uma evolução quando comparada ao mercado nos Estados Unidos (conforme amostragem de Conway), com indicação de aderência de ML em mais campos de aplicação no mercado imobiliário, com destaque para “avaliação de risco” e “visão computacional”.

CAMPOS DE APLICAÇÃO DE TÉCNICAS DE ML	Conway (2017)		Questionário	
	[a]	[b]	[a]	[b]
COLETA E FORNECIMENTO DE DADOS <i>DATA GATHERING/DISTRIBUTION</i>	3	100%	5	63%
ANÁLISE DE DADOS <i>ANALYTICS</i>	3	100%	6	75%
AVALIAÇÃO DE RISCO <i>RISK ASSESSMENT</i>	0	0%	3	38%
PROCESSOS DA EMPRESA <i>BUSINESS PROCESS</i>	0	0%	1	13%
PROCESSAMENTO E GERAÇÃO DE "LINGUAGEM NATURAL" <i>NATURAL LANGUAGE PROCESSING/GENERATION</i>	0	0%	1	13%
VISÃO COMPUTACIONAL <i>COMPUTER VISION</i>	0	0%	2	25%
PLANEJAMENTO 3D/DE ESPAÇO <i>3D/SPACE PLANNING</i>	0	0%	0	0%
ANÁLISE GEO-ESPACIAL <i>GEOSPATIAL ANALYTICS</i>	2	67%	5	63%
INTERNET DAS COISAS <i>INTERNET OF THINGS (IOT)</i>	0	0%	0	0%
TOTAL DE ENTREVISTADOS	3		8	

Figura 29 Comparação entre respostas adquiridas nas entrevistas de Conway (2018) e a autora. Fonte: Elaboração Própria.

6 CONCLUSÕES E RECOMENDAÇÕES PARA FUTUROS TRABALHOS.

Com o objetivo de entender o panorama de aderência do uso de tecnologias de ML no setor imobiliário, mais especificamente, em ativos residenciais de base imobiliária voltados a valoração. A autora, a partir da exploração de estudos de caso com implementação de algoritmos previamente fundamentados no levantamento bibliográfico, identificou melhorias como [1] Baur (2023): redução de erros médios absolutos (MAE) na precisão de valores de aluguel de apartamentos em Berlim em até 17,09% em comparação aos modelos tradicionais e [2] Potrawa (2022): aplicação de técnicas de ML causando 25% de aumento na precisão dos modelos de preços hedônicos. Além disso, também foi possível identificar diversas relações complexas que não poderiam ser capturadas por modelos convencionais [3] Linchuang (2021): os efeitos não-lineares do Bus Rapid Transit (BRT), ou corredor de ônibus, nos preços de imóveis em Xiamen, China; [4] Viana et al. (2017) a influência de características dos imóveis (eg. preço, diária, número de vagas de garagem etc.) na sua taxa de rentabilidade nominal.

A fim de identificar o estágio de adoção que o nicho de valoração de ativos residenciais de base imobiliária está, foram entrevistadas 8 dentre as 10 empresas pré-selecionadas, que fazem parte desse mercado, sendo possível atingir uma amostra representativa do mercado e cumprir o propósito fundamental deste trabalho: medir o nível de aderência de tecnologias de ML nesse setor.

Com o intuito de facilitar a compreensão sobre as empresas entrevistadas ainda que preservando suas identidades, as empresas entrevistadas foram classificadas em três categorias distintas: (Grupo C), com 4 empresas focadas na comercialização de ativos residenciais para renda, (Grupo LS), com 3 empresas direcionadas ao desenvolvimento, administração e operação de residenciais de locação (“*Long-Stay*”¹¹) e, finalmente, o (Grupo BD), com 1 empresa focada no fornecimento de inteligência com base em dados para esse setor. Complementarmente, para analisar os dados, foi possível dividir o nível de aderência em 4 categorias: [Nível 1] Empresas que já utilizam dados para processos da empresa em geral; [Nível 2] Empresas que conhecem o conceito de IA e ML; [Nível 3] Empresas que veem potenciais de aplicação de tecnologias de IA e ML no seu nicho

¹¹ “empreendimentos residenciais voltados para renda”
(Fonte:<https://cte.com.br/blog/gerenciamento/long-stay/>). Acessado em maio/2024.

de EBI residencial ou em outros nichos dentro de Real Estate, sendo capazes de enxergar oportunidades; e, finalmente, [Nível 4] Empresas que já aplicam ML.

Os resultados mostram que 100% da amostra atingiu o Nível 1 de aderência, indicando que todas as empresas utilizam dados em sua operação e, portanto, tem potencial para eventualmente iniciar a aplicação de tecnologias como ML em seu negócio; 100% atingiram o Nível 2, sugerindo que, uma vez que os conceitos de ML e IA não são estranhos às empresas, isso pode indicar a possibilidade de que, futuramente, oportunidades de implementação dessas tecnologias sejam identificadas e engajadas em seus respectivos setores; 75% atingiu o Nível 3 (2 dentre 4 do Grupo C, 3 dentre 3 do Grupo LS e a empresa do Grupo BD), confirmando que ainda existe espaço para divulgação de mais oportunidades de aplicação de tecnologias de ML que possam beneficiar as atividades do nicho de valoração em EBIs residenciais e, finalmente, 62% atingiu o Nível 4 (3 dentre 4 do Grupo C, 2 dentre 3 do Grupo LS e a empresa do Grupo BD), que representam as empresas que já aplicam ML e conseguem apontar quais os algoritmos que estão sendo adotados e para qual propósito eles são usados.

Dentre as principais aplicações apontadas pelas empresas no setor residencial de base imobiliária no segmento de valoração estão as áreas: “condição de mercado”, “análise de valor” e “análise do imóvel”, com uma abrangência bastante variada entre as aplicações, entre as principais: “coleta e fornecimento de dados”, “análise de dados” e “análise geo-espacial”.

As principais oportunidades e desafios identificados pelas empresas nesse nicho foram, [1] a digitalização das documentações em cartório, que tornará possível acessar dados de transação para compra e venda de imóveis e, conseqüentemente, influenciará na qualidade dos dados que treinam os algoritmos tonando as estimativas mais próximas ao *Market Tradable Value* e [2] a busca por uma solução que unifique dados de locação diante de um mercado que foi identificado por uma das empresas como “fragmentado”.

Cumprindo um dos objetivos secundários desse trabalho, a comparação dos resultados deste estudo com o de Conway (2018) para o Instituto de Tecnologia de Massachusetts (MIT) - "Artificial Intelligence and Machine Learning: Current Applications in Real Estate", revela uma evolução do mercado de São Paulo em 2024 em relação ao mercado dos EUA (segundo Conway) em 2018. Ambas as pesquisas destacam "coleta de dados",

"análise de dados" e "análise geo-espacial" como principais áreas de atuação. Adicionalmente, São Paulo apresenta maior aderência de Machine Learning em mais campos, sobretudo em "avaliação de risco" e "visão computacional".

Em conclusão, com esse trabalho, foi possível destacar, conforme identificado tanto nas referências bibliográficas quanto nas entrevistas, que a aplicação efetiva de ML e IA pode representar um ganho estratégico e um diferencial competitivo para as empresas do setor imobiliário. Entretanto, focando no setor de valoração de EBIs residenciais, apenas 62% da amostragem se encontra no nível mais avançado de aderência (Nível 4), em que já existe aplicação de ML dentro da empresa. Entende-se que, a fim de tornar este número ainda mais representativo, é preciso consolidar um Nível 3 com mais exemplos que sejam capazes de ampliar o conhecimento sobre o funcionamento desses algoritmos e de exemplificar formas de utilizá-los dentro do nicho de valoração de EBIs residenciais para que assim mais empresas possam avançar para o Nível 4.

6.1 RECOMENDAÇÕES PARA FUTUROS TRABALHOS

Considerando a rápida evolução do ML e as vantagens competitivas que tais tecnologias oferecem, é essencial que os profissionais do setor imobiliário estejam atualizados através de programas de treinamento e desenvolvimento contínuo para garantir que os *stakeholders* possam aproveitar ao máximo as inovações tecnológicas. Sendo assim, trabalhos que contribuam para a construção de uma forma de pensar que leve em consideração como e quais dados armazenar a partir do entendimento dos potenciais resultados que poder ser alcançados em larga escala serão fundamentais para a evolução da aderência de tecnologias de ML no setor imobiliário.

Além disso, recomenda-se uma consideração cuidadosa dos aspectos éticos relacionados à implementação de ML no mercado imobiliário. Isso inclui a transparência nas decisões dos modelos, a equidade no tratamento de diferentes tipos de propriedades e a mitigação de possíveis vieses nos algoritmos a fim de endereçar quaisquer dúvidas por parte de profissionais que poderiam advogar pelo uso eficiente dessas tecnologias em prol de maior qualidade na leitura de dados para tomada de decisão.

Ao abordar essas recomendações, os futuros trabalhos podem contribuir significativamente para a evolução positiva da aplicação de ML no mercado imobiliário, especialmente na área de valoração, promovendo maior eficiência, precisão e ética na tomada de decisões. Para isso, inicialmente recomenda-se ampliar a amostragem por meio da utilização da mesma estrutura de perguntas, abrangendo todas as áreas consideradas no estudo de Conway (2018). Este trabalho concentrou-se exclusivamente no enfoque de valoração e uma abordagem mais ampla poderia oferecer uma visão mais completa.

Além disso, uma análise mais aprofundada do impacto prático e da eficiência operacional da implementação de soluções baseadas em ML na valoração imobiliária também ampliaria o universo dessa discussão. Isso abrange avaliações de tempo, recursos e custos associados à integração dessas tecnologias nos processos existentes do setor. Ao realizar estudos de viabilidade econômica para avaliar o retorno sobre o investimento na implementação de soluções baseadas em IA na valoração imobiliária e ao compreender os benefícios financeiros a longo prazo é possível que seja ainda mais claro o incentivo para a adoção mais ampla dessas tecnologias.

REFERÊNCIAS

- AMATO, F. (2001). Arbitragem de Valor: uma rotina de análise para empreendimentos de base imobiliária. *Dissertação (Mestrado)*.
- ANDERS, H., & et al. (338 - 364 de Maio de 2022). House price prediction with gradient boosted trees under different loss functions. *Journal of Property Research*.
- ARRIETA, A., & et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58, 82 - 115.
- BAUM, A. (2017). *Proptech 3.0: The Future of Real Estate*. Said Business School and University of Oxford.
- BAUR, K., ROSENFELDER, M., & LUTZ, B. (2023). Automated real estate valuation with machine learning models using property descriptions. *Expert Systems With Applications*.
- BISHOP, C. (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.
- CHAN, F., SCHULZ, R., & ZHANG, Z. (2023). An application of machine learning in real estate economics: what extra benefits could machine learning techniques provide? *25th International Congress on Modelling and Simulation*, (pp. 123 - 129). Darwin, NT, Australia.
- CIABURRO, G., & IANNACE, G. (Maio de 2021). Machine Learning-Based Algorithms to Knowledge Extraction from Time Series Data: A Review. *Data*, pp. 1 - 30.
- CONWAY, J. (2018). *Artificial Intelligence and Machine Learning: Current Applications in Real Estate*. Dissertação de Mestrado, Massachusetts Institute of Technology, Massachusetts.
- DAMORAN, A. (2012). *Investment Valuation: tools and techniques for determining the value of any asset*. New York: John Wiley & Sons.

- ENCICLOPÉDIA BRITÂNICA. (s.d.). *Definição: machine learning, 2009*. Acesso em 2023, disponível em <https://global.britannica.com/technology/machine-learning>
- EVERITT, T., & HUTTER, M. (2018). *Universal Artificial Intelligence*. Springer.
- FLORENCIO, L. d., & ALENCAR, C. T. (Agosto de 2016). Valuation de Empreendimentos de Base Imobiliária para Garantia: Uma abordagem pelo Valor da Oportunidade de Investimento e Taxa de Desconto Ajustada ao Risco. *Revista de Finanças Aplicadas*, p. 30.
- GAO, Q., & et al. (Dezembro de 2022). Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia. *Land Use Policy*, pp. 1 - 18.
- GOODFELLOW, I., BENGIO, Y., & COURVILLE, A. (2016). *Deep Learning*. MIT Press.
- GUNNING, D. (2017). *Explainable Artificial Intelligence (XAI)*. DARPA.
- HARASIM, L. (2015). Educação Online e as Implicações da Inteligência Artificial. *Revista da FAEEBA – Educação e Contemporaneidade*, pp. 25 - 39.
- HOLDEFER, D., YENDO, G., & ALVES, D. (2022). Inteligência Artificial no Registro de Imóveis Brasileiro: Desafios e Possibilidades, à Luz dos Princípios da Administração Pública. *Revista de Direito, Governança e Novas Tecnologias*, 128 - 145.
- KNOLIKA, P., & et al. (Junho de 2020). Review on the Application of Artificial Neural Networks in Real Estate Valuation. *International Journal of Advanced Trends in Computer Science and Engineering*, pp. 2918 - 2925.
- KOK, N., KOPONEN, E.-L., & MARTINEZ, C. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *The Journal of Portfolio Management*, 202 - 211.
- LIMA JR., J. D. (1994). *Arbitragem de Valor de Portfólios de Base Imobiliária*. Boletim Técnico da Escola Politécnica da USP, Departamento de Engenharia de Construção Civil, São Paulo.

- LIMA JR., J. D., MONETTI, E., & DE ALENCAR, C. T. (2011). *REAL ESTATE. Fundamentos para Análise de Investimentos*. São Paulo: Elsevier Editora LTDA.
- LIU, J., WANG, J., & WANG, S. (Janeiro de 2006). Pattern recognition: An overview. *International Journal of Computer Science and Network Security*, pp. 56 - 57.
- MANNING, C. (2020). *Artificial Intelligence Definitions*. Palo Alto: Stanford university. HAI (Human-Centered Artificial Intelligence).
- MANNING, C. (2020). *Artificial Intelligence Definitions*. Stanford University HAI (Human-Centered Artificial Intelligence).
- MCCARTHY, J., ROCHESTER, N., & SHANNON, C. (1995). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *Nova impressão: AI Magazine Volume 27 Number 4 (2006)*, 12 - 14.
- MCKINSEY & COMPANY. (2023). Technology Trends Outlook 2023. *Mckinsey Digital*.
- MEHTA, R. (2017). *Big Data Analytics with Java*. Birmingham: Packt.
- MONARD, M., & BARANAUSKAS, J. (2003). Capítulo 4: Conceitos sobre aprendizado de máquina. Em S. O. Rezende, *Sistemas Inteligentes, Fundamentos e Aplicações* (pp. 39 - 56). Manoele Ltda.
- MONETTI, E. (1996). Análise de Riscos do Investimento em Shopping Centers. *Tese (Doutorado)*.
- NAIR, R., & NARAYANAN, A. (2012). *Benefitting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage*. Chicago: Bozz & Company.
- ODUNFA, V., FATEYE, T., & ADEWUSI, A. (2021). Application of Artificial Intelligence (AI) Approach to African Real Estate Market Analysis Opportunities and Challenges. *Proceedings of the Accra Bespoke Multidisciplinary Innovations Conference*, (pp. 121 - 132). Ghana.

- PEIXOTO, F. (2020). *Direito e Inteligência Artificial*. Coleção Inteligência Artificial e Jurisdição. Volume 2. Brasília, DF, Brasil. doi:ISBN nº 978-65-00-08585-3
- POTRAWA, T., & TETEREVA, A. (Maio de 2022). How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, pp. 50 - 65.
- ROSS, T. J. (2010). *Fuzzy Logic with Engineering Applications*. Novo México: Wiley.
- RUSSEL, S., & NORVING, P. (1995). *Artificial Intelligence - A Modern Approach* (Vol. 3º Edição). Pearson.
- SCHMIDHUBER, J. (Maio de 2014). Deep Learning in Neural Networks: An overview. *Neural Networks*, pp. 85 - 117.
- SCHWAB, K. (2016). *A quarta Revolução Industrial*. Geneva: Edipro.
- TAKAOKA, M. (2009). *Ativos para a geração de renda mensal a longo prazo: fatores preponderantes para a decisão de investimento, expectativas dos investidores e ferramentas de avaliação*. Escola Politécnica da Universidade de São Paulo, Engenharia Civil. São Paulo: SP.
- VIANA, J., VAZ, R., ALVES, G., & TORNERI PORTO, P. (Setembro de 2017). Análise da Taxa de Rentabilidade de Imóveis Residenciais para Locação de Curto Prazo: Uma Abordagem Utilizando Big Data e Inteligência Artificial. *17ª Conferência Internacional da LARES*.
- VON WANGENHEIM, A., & VON WANGENHEIM, C. (2003). *Raciocínio Baseado em Casos*. Santa Catarina: Manole.
- WEBER, P., CARL, K., & HINZ, O. (2023). *Applications of Explainable Artificial Intelligence in Finance - A systematic review of Finance, Information Systems, and Computer Science literature*. Frankfurt: Springer.
- YAZDANI, M. (Outubro de 2021). Machine Learning, Deep Learning, and Hedonic: Methods for Real Estate Price Prediction. pp. 1 - 33.

APÊNDICE 1

Questionário Aplicado – Versão Completa

SEÇÃO 1: FOCO EM APLICAÇÕES NA EMPRESA			
<p>2</p> <p>Dados são importantes para modelagem e ferramentas de ML e IA.</p> <p>Quais são as principais fontes de dados em seu negócio e quais os principais tipos de dados coletados?</p>	2.1 Fonte de Coleta dos Dados	Dados Coletados Internamente	
		Dados Contratados de Plataformas Externas	
		Dados Públicos / Disponíveis Online	
	2.2 Tipos de Dados Coletados	Planilhas: Dados organizados em formato de planilha, como Microsoft Excel ou Google Sheets.	
		Bancos de Dados Relacionais: Fontes de dados organizadas em tabelas com esquemas predefinidos, como MySQL, PostgreSQL, SQL Server.	
		Texto da Web: Texto em páginas da web que não segue um formato específico.	
		Documentos de Texto Não Formatados: Texto em formato livre, como documentos Word não estruturados.	
		Mídias Sociais: Dados de redes sociais, incluindo postagens, comentários, vídeos e imagens.	
		E-mails: Conteúdo de e-mails, incluindo corpo do e-mail e anexos.	
		Vídeos: Dados de vídeo não estruturados, como vídeos no YouTube ou Vimeo.	
		Áudio: Gravações de áudio, como arquivos de podcast ou chamadas telefônicas.	
Imagens: Imagens não estruturadas, como fotos digitais ou imagens escaneadas.			
Caso as fontes listadas acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações.			
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.			
<p>3</p> <p>Estou interessado em entender como o ML está sendo utilizado no mercado imobiliário de valorização de Empreendimentos Residenciais de Base Imobiliária atualmente.</p> <p>Isso é algo que você/sua empresa estão utilizando? Se sim, por favor, descreva como as técnicas estão sendo aplicadas.</p>	3.1 SIM, COMO?	COLETA E FORNECIMENTO DE DADOS	
		ANÁLISE DE DADOS	
		AVALIAÇÃO DE RISCO	
		PROCESSOS DA EMPRESA	
		PROCESSAMENTO E GERAÇÃO DE "LINGUAGEM NATURAL"	
		VISÃO COMPUTACIONAL	
		PLANEJAMENTO 3D/DE ESPAÇO	
		ANÁLISE GEO-ESPACIAL	
		INTERNET DAS COISAS	
		OUTROS	
	Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.		
Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.			
3.2 NÃO			

SEÇÃO 1: FOCO EM APLICAÇÕES NA EMPRESA

SEÇÃO 1: FOCO EM APLICAÇÕES NA EMPRESA					
4	Caso a pergunta 3 tenha sido respondida com "Sim", poderia informar quais algoritmos de ML estão sendo utilizados ?		"Linear/ Logistic Regression" (LR) - Regressão Linear/ Logística		
			"Decision Trees" (DT) - Árvore de Decisão		
			"K-Nearest Neighbors" (KNN)		
			"Rule Based Learners" (RBL) - Aprendizado Baseado em Regras		
			"General Additive Models" (GMA) - Modelos Aditivos Generalizados		
			"Bayesian Models" (BM) - Modelos Bayesianos		
			Tree ensembles, random forest and multiple classifier systems" (TE, RF, MC)		
			"Support Vector Machines" (SVM) - Máquinas de Vetores de Suporte		
			"Neural Networks" (NN) - Rede Neural		
			"Multi-Layer Networks" (MLN) - Redes de Múltiplas Camadas		
			"Recurrent Neural Networks" (RNN) - Redes Recorrentes		
			"Convolutional Neural Networks" (CNN)		
			Caso as fontes listadas acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações e destaque quais técnicas acima são aplicadas dentro do campo de valoração.		
	Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.				
5	Caso a pergunta 3 tenha sido respondida com "Sim", poderia informar em quais áreas específicas tais tecnologias tem sido aplicadas na sua empresa?	5.1 SIM. QUAIS?	Análise do imóvel (Localização, Vizinhança, Infra, Especificações Técnicas, etc)		
			Legislação Urbana (Zoneamento, etc)		
			Titulação e Dispendios (Matrícula, Escritura, IPTU/ITR, Habite-se, etc)		
			Condições de Mercado (Oferta, Demanda, Preços Praticados, Tendências, etc)		
			Análise de Valor (Posicionamento do imóvel perante o mercado, Aplicações de Metodologias, Conclusão de Valor e etc.)		
			Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.		
			Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.		
		5.2 NÃO			

SEÇÃO 2: FOCO EM OPINIÃO GERAL DE MERCADO E IDENTIFICAÇÃO DE OPORTUNIDADES

6	Você está ciente de outras maneiras pelas quais o ML está sendo aplicado no mercado imobiliário de <u>avaliação de Empreendimentos Residenciais de Base Imobiliária atualmente?</u>	6.1 SIM, QUAIS?	COLETA E FORNECIMENTO DE DADOS	
			ANÁLISE DE DADOS	
			AVALIAÇÃO DE RISCO	
			PROCESSOS DA EMPRESA	
			PROCESSAMENTO E GERAÇÃO DE "LINGUAGEM NATURAL"	
			VISÃO COMPUTACIONAL	
			PLANEJAMENTO 3D/DE ESPAÇO	
			ANÁLISE GEO-ESPACIAL	
			INTERNET DAS COISAS	
			OUTROS	
			Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações. Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.	
6.2 NÃO				
7	Existem áreas específicas no mercado imobiliário de <u>avaliação de Empreendimentos Residenciais de Base Imobiliária</u> que você acredita serem boas oportunidades para aplicar técnicas de ML ou IA ?	7.1 SIM, QUAIS?	Análise do imóvel (Localização, Vizinhança, Infra, Especificações Técnicas, etc)	
			Legislação Urbana (Zoneamento, etc)	
			Titulação e Dispendios (Matrícula, Escritura, IPTU/ITR, Habite-se, etc)	
			Condições de Mercado (Oferta, Demanda, Preços Praticados, Tendências, etc)	
			Análise de Valor (Posicionamento do imóvel perante o mercado, Aplicações de Metodologias, Conclusão de Valor e etc.)	
			Caso as aplicações listadas acima não contemplem o que você gostaria de responder, utilize esse espaço para mais observações.	
			Por gentileza, utilizar esse espaço para justificar sua resposta e/ou indicar exemplos que ilustrem sua resposta.	
7.2 NÃO				
8	Que tipos de dados podem estar faltando ou ser úteis para aprimorar a análise ou aplicações de aprendizado de máquina e IA no mercado de <u>avaliação de Empreendimentos Residenciais de Base Imobiliária?</u>	Planilhas: Dados organizados em formato de planilha, como Microsoft Excel ou Google Sheets.		
		Bancos de Dados Relacionais: Fontes de dados organizadas em tabelas com esquemas predefinidos, como MySQL, PostgreSQL, SQL Server.		
		Texto da Web: Texto em páginas da web que não segue um formato específico.		
		Documentos de Texto Não Formatados: Texto em formato livre, como documentos Word não estruturados.		
		Mídias Sociais: Dados de redes sociais, incluindo postagens, comentários, vídeos e imagens.		
		E-mails: Conteúdo de e-mails, incluindo corpo do e-mail e anexos.		
		Vídeos: Dados de vídeo não estruturados, como vídeos no YouTube ou Vimeo.		
		Áudio: Gravações de áudio, como arquivos de podcast ou chamadas telefônicas.		
		Imagens: Imagens não estruturadas, como fotos digitais ou imagens escaneadas.		
Caso os tipos de dados listados acima não contemplem o que você gostaria de responder, utilize o espaço abaixo para mais observações.				