UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
FACULDADE DE ADMINISTRAÇÃO E CIÊNCIAS CONTÁBEIS
DEPARTAMENTO DE ADMINISTRAÇÃO

CAIO CESAR VIEIRA TRINTA DA SILVEIRA

REVISÃO E APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE *CHURN*

RIO DE JANEIRO
2022

CAIO CESAR VIEIRA TRINTA DA SILVEIRA

REVISÃO E APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE *CHURN*

Monografia apresentada à Faculdade de Administração e Ciências Contábeis da Universidade Federal do Rio de Janeiro (FACC/UFRJ), como requisito parcial à obtenção do grau de Bacharel em Administração.

Orientador (a): Carlos César Trucíos Maza

Rio de Janeiro

2022

CAIO CESAR VIEIRA TRINTA DA SILVEIRA

REVISÃO E APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE *CHURN*

Monografia apresentada à Faculdade de Administração e Ciências Contábeis da Universidade Federal do Rio de Janeiro (FACC/UFRJ), como requisito parcial à obtenção do grau de Bacharel em Administração.

_____

Prof. Carlos César Trucíos Maza (Orientador)

_____

Profa. Cristina Pimenta De Mello Spineti Luz (Leitora)

Rio de Janeiro, 2022

Dedico este trabalho aos meus pais Maria de Fatima Vieira Serra e Augusto Cesar Trinta da Silveira que me ensinaram cidadania e ética através do exemplo e lutaram para poder oferecer a mim a melhor educação formal possível.

AGRADECIMENTOS

Agradeço à minha família pelo apoio incondicional, suporte e ao incentivo de meus estudos ao longo da minha vida. Em especial para minha namorada Sabrina, que me deu todo o apoio emocional para a elaboração deste trabalho.

Agradeço a UFRJ por me oferecer há 6 anos uma educação pública, gratuita e de qualidade e por defenderem sempre a ciência em tempos difíceis.

Agradeço aos meus colegas e amigos de estudos, em especial para a Mariana e Mariah por compartilharem comigo as alegrias e dificuldades dos últimos quatro anos de graduação.

Agradeço aos professores que fizeram parte do meu aprendizado, em especial o professor orientador Carlos Trucíos, que ofereceu todo o suporte necessário para a realização deste trabalho.

RESUMO

Em um mundo onde os produtos, processos e relações estão gradativamente mais digitalizados, os algoritmos de aprendizado de máquina têm-se tornado ferramentas importantes em todas as áreas das organizações. Esses algoritmos ganharam relevância dentro da elaboração da estratégia de *marketing*, pois a partir de dados disponíveis eles conseguem descrever e predizer comportamentos relacionados ao *churn,* que pode ser definido pela desistência do relacionamento com a empresa por parte do cliente. Portanto, há necessidade de o profissional de administração entender essas técnicas preditivas, bem como suas vantagens e desvantagens. Este trabalho busca, através de pesquisa bibliográfica de trinta artigos científicos em inglês e português da última década: (i) identificar os principais algoritmos de aprendizado de máquina utilizados na literatura para a predição do cancelamento de serviço por parte dos clientes; (ii) apresentar as principais métricas utilizadas para avaliar o desempenho dos modelos; (iii) ilustrar os usos destes métodos através da linguagem de programação *Python*. Observou-se que, para o problema de *churn*, é preferível o uso de métodos de fácil interpretação, sendo a regressão logística e árvores de decisão são os métodos mais utilizados. Em pesquisas onde o objetivo principal era encontrar métodos com melhor desempenho, foi observado o uso de métodos mais complexos como florestas aleatórias e redes neurais, que prometem maior desempenho ao custo de menor capacidade de interpretação e geração de *"insights".* A vantagem observada em usar mais de um método para resolver problemas de classificação, é que há maneiras simples de comparar o desempenho entre modelos a partir de métricas calculadas através da matriz de confusão, como acurácia, precisão e *recall*. Por fim, este trabalho elucida que atualmente há inúmeros algoritmos desenvolvidos e pré-configurados na linguagem *Python,* logo é possível começar a aplicar com rapidez modelos complexos com conhecimento intermediário de programação. Em muitos casos, as maiores dificuldades para o pesquisador podem ser a interpretação, seleção de variáveis apropriadas para abordar o problema e escolha do melhor método preditivo.

Palavras-chave: *churn*; aprendizado de máquina; predição, classificação.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. O Problema de Pesquisa

O crescimento tecnológico exponencial das últimas décadas permitiu que capturar, armazenar e processar dados no ambiente digital se torne cada vez mais fácil e acessível. Atualmente, é muito difícil encontrar grandes empresas que não utilizem sistemas de informação para a transformação de dados em conhecimento estratégico e, por sua vez, este conhecimento, em uma ferramenta para auxiliar no processo de tomada de decisão. Entretanto, a capacidade das organizações em interpretar e gerenciar dados não acompanhou o aumento significativo do volume e variedade de dados agora sob o poder das organizações (FAYYAD *et al*., 1996).

A digitalização possibilitou o armazenamento de maior volume e variedade de dados sobre os consumidores e seus comportamentos, o que permite às organizações personalizar e customizar as ofertas e campanhas (RYAN; JONES, 2016). Nesse cenário, as empresas estão investindo cada vez mais na capacidade de transformar dados em inteligência competitiva.

Na perspectiva do marketing, um dos principais desafios enfrentados é a fidelização do consumidor. O valor resultante do consumo não é constante, mas varia de acordo com a relação desenvolvida entre cliente-empresa. É importante para as organizações a gestão, controle e desenvolvimento do ciclo de vida dos clientes, criando assim estratégias para melhorar a relação do cliente com a marca, evitando a desistência da relação. Aumentar o ciclo de vida do cliente é importante pois representa o desenvolvimento da receita líquida do cliente ao longo da relação com a empresa (STAUSS; FRIEGE, 1999).

Algumas das principais ferramentas para alcançar esse objetivo são os algoritmos que possibilitam identificar, entender e predizer o *churn*. O *churn* pode ser definido como a desistência/cancelamento de um determinado serviço por parte do cliente. A identificação, entendimento e predição desse comportamento são de vital importância nas organizações pois auxiliam na criação de políticas de marketing direcionadas para segmentos específicos de clientes. A aplicação desses algoritmos se torna importante quando há muitos fatores que podem influenciar o resultado desse evento e o esforço de analisar esses fatores manualmente é muito alto. Além disso,

em grandes organizações o volume de dados é tão grande que é humanamente impossível resolver problemas de grande escala.

O objetivo desta pesquisa é identificar e aplicar os principais algoritmos de aprendizado de máquina para a predição do comportamento de desistência/cancelamento de clientes (*churn*).

1.2. Objetivos

1.2.1 Objetivo Geral

Identificar os principais algoritmos de aprendizado de máquina utilizados na literatura para a predição do cancelamento de serviço por parte dos clientes.

1.2.2 Objetivos Específicos

1) Explicar os principais métodos, bem como suas vantagens e desvantagens.
2) Apresentar as principais métricas para avaliar a capacidade preditiva dos modelos de predição de *churn*.
3) Ilustrar os usos destes métodos através da linguagem de programação *Python*.

1.3. Justificativa

A principal motivação para a realização deste trabalho está na urgência do uso da informação como diferencial competitivo para empresas e a corrida para atingir o melhor resultado em um mercado cada vez mais tecnológico e competitivo. Por outro lado, também há urgência no mercado de trabalho para a formação de profissionais que possuam competências em análise de dados e conhecimento sobre modelos estatísticos para resolver problemas de negócio.

Para empresas de tecnologia, o principal combustível nessa corrida está na utilização de tecnologias de *big data*. É possível observar um crescimento acelerado no mercado de soluções de *big data* e análise de dados. Segundo pesquisa do *Worldwide Big Data and Analytics Spending Guide da International Data Corporation* (IDC, 2021), estimou-se que em 2021 o setor movimentaria mais de US$ 215 bilhões. De acordo com a pesquisa, em 2021 o aumento deste setor seria de aproximadamente 10,1% em relação ao ano anterior.

Segundo o relatório da consultoria *Frost & Sullivan* (2017), o Brasil já era o país líder em soluções de *big data* e análise de dados na América Latina, com 46,8% do mercado da região, gerando receita de US$ 1,16 bilhão. O mercado brasileiro de *Big Data* continua a crescer vertiginosamente, gerando competitividade entre empresas no desenvolvimento de novas soluções e na contratação de profissionais qualificados nos campos da ciência de dados.

Nesse cenário, este trabalho busca reunir, analisar e aplicar os mais relevantes algoritmos de aprendizado de máquina para a predição de *churn*, servindo como guia para novos profissionais brasileiros, dado que ainda há escassez de material relevante em português em uma área do conhecimento tão importante e atual para as organizações como a ciência de dados.

2. REFERENCIAL TEÓRICO

Esta sessão é composta pela definição e contextualização de três conceitos. Em primeiro será conceitualizado as tecnologias de Big Data, posteriormente define-se Gestão de Relacionamento com o Cliente (*CRM*) e em terceiro a métrica de análise de negócios *Churn*.

2.2. Conceitos de *Big Data*

A adoção da computação em nosso dia a dia digitalizou quase todas as informações que produzimos e consumimos como indivíduos e organizações. A digitalização dos processos, relacionamentos, conteúdo e consumo possibilitou às empresas registrarem em seus bancos de dados eventos de diversos tipos. É neste contexto que surge o termo *Big Data* para descrever grandes conjuntos de dados estruturados e não estruturados que desafiam o poder de processamento, análise e armazenamento de dados das organizações. Apesar de para Gandomi e Haider (2015) o volume de dados representar o principal fator de definição de *big data*, eles enumeram também outros fatores que caracterizam essa tecnologia: 1) a velocidade que o dado é gerado e processado dentro da organização. 2) a variedade de formatos de dados que são capturados dentro do processo de armazenamento. 3) a veracidade ou confiança em relação aos dados capturados. 4) o valor que o processo de Big Data consegue entregar para a organização, os dados armazenados precisam ser úteis

para determinado fim. Assim, embora o termo Big Data não possua uma única definição, é amplamente aceito que *Big Data* é caracterizado por um modelo 5V (volume, velocidade, variedade, veracidade e valor).

Apesar da capacidade computacional da organização em capturar, processar e armazenar grandes conjuntos de dados ser fundamental para o processo de *big data*, é apenas em sua capacidade de análise e interpretação que a tecnologia consegue entregar valor para a tomada de decisão gerencial (GANDOMI; HAIDER, 2015). A necessidade de analisar grandes volumes de dados desafia os limites de softwares empresariais tradicionais e de métodos estatísticos básicos. Para superar esses limites surgiram um conjunto de novas ferramentas baseadas em mineração de dados, inteligência artificial, processamento de linguagem natural e análises estatísticas. No mercado esse conjunto de técnicas de análise para grandes volumes de dados ganhou relevância dentro da estratégia de marketing das grandes companhias, com o objetivo de gerar conhecimento sobre sua base de clientes e melhorar a gestão de relacionamento com o cliente (*CRM*).

2.3. Gestão de Relacionamento com o Cliente (*CRM*)

Para as organizações, há um alto custo em atrair novos compradores. Logo uma ótima estratégia é investir cada vez mais em ações para reter os compradores atuais e gerar maior recorrência de compra. O abandono da relação com a empresa por parte dos consumidores afeta diretamente os resultados financeiros das organizações, além de poder afetar negativamente a percepção do público com a marca (KATELARIS; THEMISTOCLEOUS, 2017).

O uso de ferramentas e processos de *Customer Relationship Management (CRM)* são algumas das formas de aumentar a satisfação e contato do consumidor com a marca. Segundo Neslin *et al*. (2006), CRM é um processo interativo que transforma informação sobre os clientes em relacionamentos que beneficiam tanto a empresa quanto o cliente. O principal objetivo ao adotar um processo de *CRM* é aumentar a participação da empresa no hábito de consumo da sua base de clientes, em vez de maior participação de mercado. Nesse sentido, as tecnologias de *CRM* permitem capturar e armazenar em bancos de dados as informações externas e internas do consumidor que são relevantes para todas as áreas da organização. A organização, análise e distribuição desses dados permitem maior eficiência e

personalização da interação com o cliente em todos os pontos de contato com a organização (PEPPERS; ROGERS, 2001).

2.4. Churn

Segundo Katelaris e Themistocleous (2017), o *churn* ou desistência/cancelamento de clientes ocorre quando um cliente termina a relação com uma empresa para possivelmente iniciar um relacionamento com uma organização concorrente. Nas empresas, esse comportamento é medido pela taxa de *churn*.

A taxa de *churn* é um indicador muito importante para o marketing, pois evidencia a taxa de consumidores que terminaram a relação com a empresa. Esse indicador é o resultado do volume de clientes que uma empresa perdeu em um determinado período, dividido pelo total de clientes do mesmo período.

Para as organizações em que a satisfação de clientes é o centro das estratégias de marketing, a gestão da taxa de *churn* é uma constante preocupação. Essa preocupação aumenta de acordo com a competitividade do setor, bem como a conscientização de seus clientes sobre novas oportunidades de negócios.

Segundo Kotler (1999), conquistar um novo cliente custa entre 5 e 7 vezes mais do que manter um atual. Logo, investir na fidelização de clientes atuais é mais lucrativo do que investir na aquisição de novos clientes.

3. METODOLOGIA

3.1. Tipo de Pesquisa

Para a classificação da pesquisa utiliza-se a perspectiva apresentada por Vergara (1991) que observa os fins e os meios. O tipo de pesquisa pode ser determinado, quanto aos fins, como exploratório dado que ainda há necessidade de desenvolver, esclarecer e organizar o conhecimento sobre o tema na língua portuguesa. Quanto aos meios, realizou-se a pesquisa bibliográfica a partir da coleta de artigos científicos sobre modelos preditivos de *churn* disponíveis no portal de periódicos CAPES.

3.2. Coleta de Dados

Foram coletados no Portal de periódicos CAPES (Ministério da Educação) artigos científicos escritos em inglês ou português através da palavras-chave: "*Churn*" + (“*Prediction*” OU “Previsão” OU “Preditivo” OU “Predição”), todos presentes no título ou resumo do artigo. Nesta pesquisa, considerou-se todos os artigos publicados desde 01 de janeiro de 2010 até o dia 31 de dezembro de 2021 e foram encontrados 2.556 artigos na língua inglesa e 10 resultados na língua portuguesa.

A seleção dos artigos foi feita por critério de relevância da plataforma, que ordenada os artigos por quatro critérios: Quantidade de ocorrências do texto de busca em diversas partes do documento, ordem dos termos da pesquisa e sua proximidade dentro do documento, data de publicação e medidas de citação do texto por usuários em todo o mundo.

3.3. População e Amostra

Para Vergara (1998), é entendido como população o conjunto de elementos que possuem a característica de serem objetos de estudo. A população estudada na pesquisa se refere aos artigos científicos existentes na literatura que tenham como objetivo a aplicação de métodos preditivos para o fenômeno de *churn* no período de 2010 a 2021, disponíveis em inglês ou português no portal de periódicos CAPES. A amostra coletada dessa população é representada por 30 artigos científicos escolhidos por ordem de relevância do portal de periódicos CAPES e se configura como não probabilística por tipicidade, uma vez que foram selecionados unicamente por disponibilidade. O objetivo da análise dessa amostra é identificar quais foram os modelos preditivos utilizados pelos autores e quais deles foram os mais relevantes em suas pesquisas.

Abaixo, a Tabela 1 apresenta o ano, título, autor(es) e língua dos artigos da amostra utilizada.

Tabela 1 – Artigos analisados

Ano	Título	Autor	Língua
2019	A Churn Prediction Model Using Random Forest: Analysis Of Machine Learning Techniques For Churn Prediction And Factor Identification In Telecom Sector	ULLAH, I. *et al.*	Inglês
2016	Análise Preditiva De Churn Com Enfase Em Tecnicas De Machine Learning: Uma Revisao	SCHNEIDER, P.	Português
2020	Churn Modelling: Predição De Churn Para Uma Base De Dados De Instituição Financeira	GAVAZZA, I.	Português
2010	Modelagem De Probabilidade De Churn	BOTELHO, D.	Português
2011	Previsão De Churn Em Companhias De Seguros	GOMES, B.	Português
2013	Hierarchical Neural Regression Models For Customer Churn Predictio	MOHAMMADI, G. *et al.*	Inglês
2021	A Prediction Model Of Customer Churn Considering Customer Value: An Empirical Research Of Telecom Industry In China	ZHAO, M. et al.	Inglês
2021	A Novel Model Structured On Predictive Churn Methods In A Banking Organization	SILVEIRA, L.	Português
2016	Developing A Prediction Model For Customer Churn From Electronic Banking Services Using Data Mining	KERAMATI, A. *et al.*	Inglês
2017	Churn Classification Model For Local Telecommunication Company Based On Rough Set Theory	MAKHTAR, M. *et al.*	Inglês
2020	Improvised_Xgboost Machine Learning Algorithm For Customer Churn Prediction	SWETHA, P.; DAYANANDA, R.	Inglês
2011	New Insights Into Churn Prediction In The Telecommunication Sector: A Profit Driven Data Mining Approach	VERBEKE, W. *et al.*	Inglês
2012	Customer Churn Prediction In The Online Gambling Industry: The Beneficial Effect Of Ensemble Learning	COUSSEMENT, K.; BOCK, K.	Inglês
2021	Giant Fight: Customer Churn Prediction In Traditional Broadcast Industry	LI, Y.; *et al.*	Inglês
2020	Profit-Driven Churn Prediction For The Mutual Fund Industry: A Multisegment Approach	MALDONADO, S.	Inglês
2011	An Empirical Evaluation Of Rotation-Based Ensemble Classifiers For Customer Churn Prediction	BOCK, K.; POEL, K.	Inglês
2015	Including High-Cardinality Attributes In Predictive Models: A Case Study In Churn Prediction In The Energy Sector	MOEYERSOMS, J.; MARTENS, D.	Inglês
2016	Churn Prediction System For Telecom Using Filter–Wrapper And Ensemble Classification	IDRIS, A.; KHAN, A.	Inglês
2011	Customer Churn Prediction In Telecommunications	HUANG, N. *et al.*	Inglês
2021	Customer Churn Prediction For Telecommunication Industry: A Malaysian Case Study	MUSTAFA, N. *et al.*	Inglês

Ano	Título	Autor	Língua
2014	Credit Analysis Using Data Mining: Application In The Case Of A Credit Union	SOUSA, M.	Inglês
2018	Preditores De Retenção E Lealdade De Clientes Em Academias De Ginástica	ARCOVERDE, D.	Português
2015	A Comparison Of Machine Learning Techniques For Customer Churn Prediction	VAFEIADIS, T. *et al.*	Inglês
2019	Customer Churn Prediction In Telecom Using Machine Learning In Big Data Platform	AHMAD, A.; ALJOUMAA, A.	Inglês
2017	Early Churn Prediction With Personalized Targeting In Mobile Social Games	MILOSEVIC, M. *et al.*	Inglês
2018	Customer Churn Prediction In Telecommunication Industry Using Data Certainty	AMIN, A. *et al.*	Inglês
2018	Profit Driven Decision Trees For Churn Prediction	HOPPNER, S. *et al.*	Inglês
2017	Hyperparameter Optimization Of Artificial Neural Network In Customer Churn Prediction Using Genetic Algorithm	FRIDRICH, M.	Inglês
2017	Churn Prediction Of Mobile And Online Casual Games Using Play Log Data	KIM, S. *et al.*	Inglês
2021	Integrated Churn Prediction And Customer Segmentation Framework For Telco Business	WU, S. *et al.*	Inglês

3.4. Método de Análise

3.4.1 Ferramentas e Métodos Utilizados

A primeira etapa da análise foi a leitura da metodologia escolhida pelos autores com o objetivo de identificar todos os métodos quantitativos utilizados nos artigos da amostra. As conclusões também foram analisadas no caso de artigos que buscam comparar o desempenho entre os modelos, com o objetivo de identificar o melhor método avaliado pelo autor. Essas informações foram tabeladas no *software* Excel para a análise quantitativa de quais são os métodos mais utilizados na literatura e quais foram os métodos mais bem avaliados pelos autores. Depois que os principais métodos foram identificados e selecionados, realizou-se pesquisa bibliográfica para entender a definição, vantagens e limitações de cada modelo.

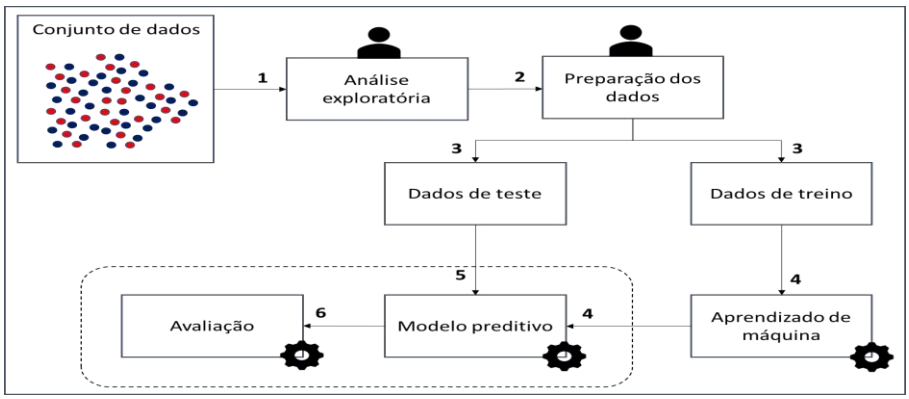
Após análise bibliográfica, a segunda etapa foi a de aplicação, descrição e avaliação dos modelos encontrados. Para a aplicação dos algoritmos de aprendizado de máquina, é recomendado que a máquina identifique e aprenda sobre as relações entre a variável de interesse e as variáveis explicativas através de dados históricos apresentados a ela. No caso da predição de *churn*, a variável de interesse (*churn*/não *churn*) já é conhecida, bem como as variáveis explicativas, como gênero do usuário, tempo de uso do produto, endereço etc. Logo, a abordagem que foi utilizada para o aprendizado da máquina foi a de aprendizagem supervisionada.

O aprendizado supervisionado é utilizado quando o algoritmo aprende a partir de valores históricos da variável de interesse para identificar as relações com as demais variáveis e aprender quais devem ser os resultados de saída. Esses valores de entrada (*churn*/não *churn*) servem como referência para o modelo ajustar suas predições com base nos erros. Para James *et al.* (2021), o objetivo dessa abordagem pode ser de responder com precisão a resposta para as observações (predição) ou a melhor compreensão da relação entre resposta e os preditores (inferência).

A figura 1 ilustra o processo de aplicação da aprendizagem supervisionada. Nesse método, é necessário que o pesquisador explore os dados e entenda as correlações e estrutura disponíveis (1), para assim preparar as variáveis de entrada no algoritmo por meio de limpeza e transformação da estrutura e tipos de dados caso necessite o algoritmo (2). Após essas duas etapas, os dados são divididos em dois conjuntos não sobrepostos (3): O conjunto de treino e o conjunto de teste. O conjunto de treino é usado para treinar o modelo e para que este aprenda as relações entre os dados (4), enquanto o conjunto de teste serve para a avaliação do modelo simulando a realidade que o modelo encontrará quando estiver em uso, usando para isso dados desconhecidos não utilizados durante o processo de aprendizagem.

Logo, o algoritmo já treinado é aplicado no conjunto de testes para avaliação dos resultados com uma menor probabilidade de viés (5). A última etapa é a de a mensuração dos resultados através de métricas de desempenho dos modelos que estão definidas na próxima seção (6).

Figura 1 – Processo de aprendizagem supervisionada



Elaborado pelo autor

As ferramentas utilizadas nesse processo foram os pacotes *pandas*, *seaborn* e *numpy*, que são ferramentas de código aberto na linguagem de programação Python (*Python Software Foundation*, 1991) para manipulação e análise de dados desenvolvidos comunitariamente. Em especial, oferecem estruturas e funções para manipular tabelas numéricas, arranjos, matrizes, séries temporais e visualização de dados.

A aplicação e avaliação dos modelos foram realizados através do pacote scikit-learn, pacote de funções desenvolvido em Python para aplicação de algoritmos de aprendizado de máquina.

3.4.2 Métricas de Avaliação dos Modelos

Como uma das etapas importantes para garantir que o modelo seja capaz de generalizar bem, o desempenho deve ser avaliado através das taxas de predição de *churn* do algoritmo (HUANG *et al*., 2011). O método de avaliação de desempenho mais presente na bibliografia analisada são as métricas que podem ser definidas através da matriz de confusão. Na avaliação de modelos de aprendizagem de máquina, a matriz de confusão nos permite mensurar e visualizar o desempenho de classificação do algoritmo utilizado através de uma tabela de contingência 2x2, também chamada de matriz de erro (Tabela 2).

Considerando uma resposta binária para os modelos (*churn* vs não *churn* ou positivo vs negativo), é possível com base na matriz de confusão definir e classificar os resultados pelos seguintes conceitos:

Tabela 2 – Matriz de confusão

Previsão	**Realidade**	
	Verdadeiro	**Falso**
Positivo	Verdadeiro Positivo	Falso Positivo
Negativo	Falso Negativo	Verdadeiro Negativo

- **Verdadeiro Positivo** – Resultado onde o modelo previu corretamente a classe positiva (*churn*).
- **Verdadeiro Negativo** – Resultado onde o modelo previu corretamente a classe negativa (Não *churn*).
- **Falso Positivo** – Resultado onde o modelo previu incorretamente a classe positiva (*churn*).
- **Falso Negativo** – Resultado onde o modelo previu incorretamente a classe negativa (Não *churn*).

A partir destas definições é possível avaliar os modelos com os seguintes indicadores de desempenho:

1) **Acurácia** – Representa a frequência relativa de valores que foram corretamente classificados no modelo. É a métrica que avalia de maneira geral a capacidade do modelo em classificar corretamente as duas opções possíveis.

$$\text{Acurácia} = \frac{\text{Verd. Pos. + Verd. Neg}}{\text{Verd. Pos. + Verd. Neg + Falso Neg. + Falso Pos.}}$$

2) ***Recall* –** Métrica que permite identificar a proporção de exemplos positivos que foram identificados corretamente. No contexto deste trabalho, essa métrica verifica quão bom o modelo é em classificar a classe positiva (*Churn*).

$$\textit{Recall} = \frac{\text{Verd. Pos.}}{\text{Verd. Pos. + Falso Neg.}}$$

3) **Precisão –** Precisão é a proporção de exemplos que foram identificados como positivos e que estavam verdadeiramente corretos. No contexto deste trabalho, a

precisão busca entender qual é a proporção de clientes classificados como *churn* que foram verificados como verdadeiros.

$$\text{Precisão} = \frac{\text{Verd. Pos}}{\text{Verd. Pos.} + \text{Falso Pos.}}$$

4) **F1 Score –** O F1 Score é necessário quando precisamos entender se há distribuição desigual entre a precisão e o *recall*, ou seja, busca medir o trade-off entre as duas taxas**.**

$$\text{F1 Score} = 2x \frac{\textit{Recall} \text{ x Precisão}}{\textit{Recall} + \text{Precisão}}$$

As métricas acima buscam mensurar o desempenho do modelo sobre os dados de teste. A acurácia compreende o desempenho geral do modelo, considerando as classes positivas e negativas. Essa métrica é usada para uma interpretação clara do modelo, mas não é confiável ao lidar com classes desbalanceadas (FRIDRICH, 2017). No contexto deste trabalho, a quantidade de clientes *churn* é bem menor do que a quantidade de clientes não *churn*, logo um modelo com alta acurácia poderia na verdade ser apenas um modelo que classifica todos os clientes como não *churn.* Portanto*,* a acurácia não é a melhor métrica de análise porque precisamos isolar o desempenho do modelo em acertar a classe positiva (Cliente é *churn*; Y = 1), para isso as métricas de precisão e *recall* são mais adequadas. Nesse sentido, os modelos precisam buscar maximizar o resultado de todos os cálculos e, ao analisá-los separadamente, é possível compreender a proporção de acertos considerando diferentes subconjuntos.

3.5. Limitações do Método

A principal limitação do método utilizado é a variedade de artigos científicos existentes que abordam o tema escolhido e não estão presentes nos buscadores utilizados, bem como o uso de outros termos sinônimos não mapeados na pesquisa e que por isso não foram coletados.

A segunda limitação é relacionada a escolha da linguagem de programação para a aplicação dos modelos. Apesar de Python (*Python Software Foundation*, 1991)

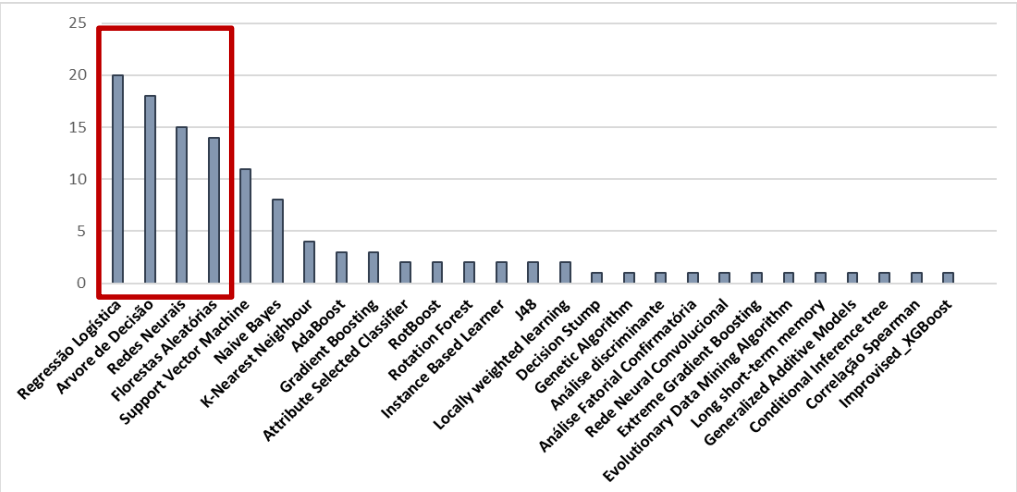
ser a linguagem mais usada atualmente na área de ciência de dados no mercado de trabalho, há outras capazes de realizar as mesmas tarefas ou mais presentes na literatura como as linguagens R (*R Core Team*, 1993) e Julia (BEZANSON *et al.*, 2012).

4. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Foram analisados 30 artigos e encontradas 120 ocorrências de uso de 27 métodos preditivos diferentes, considerando que 24 desses artigos utilizaram pelo menos 2 modelos. Foi identificado que a bibliografia sobre métodos preditivos de *churn* está altamente concentrada no setor de telecomunicações, que representa 40% das pesquisas, seguido do setor financeiro com 17%. Como dados de *CRM* são geralmente confidenciais e tratados como capital informacional para as organizações, é provável que essa concentração de pesquisas em poucos setores seja consequência de poucos conjuntos de dados disponíveis publicamente para análise. Foi possível identificar o uso da mesma base de dados de telecomunicações em quatro artigos diferentes.

No gráfico 1, foram agrupados por modelo a quantidade de usos e podemos concluir que regressão logística (20 ocorrências), árvores de decisão (18 ocorrências), redes neurais artificiais (15 ocorrências) e florestas aleatórias (14 ocorrências) são os quatro métodos mais presentes na bibliografia para a resolução de problemas que envolvem a predição e classificação de *churn*.

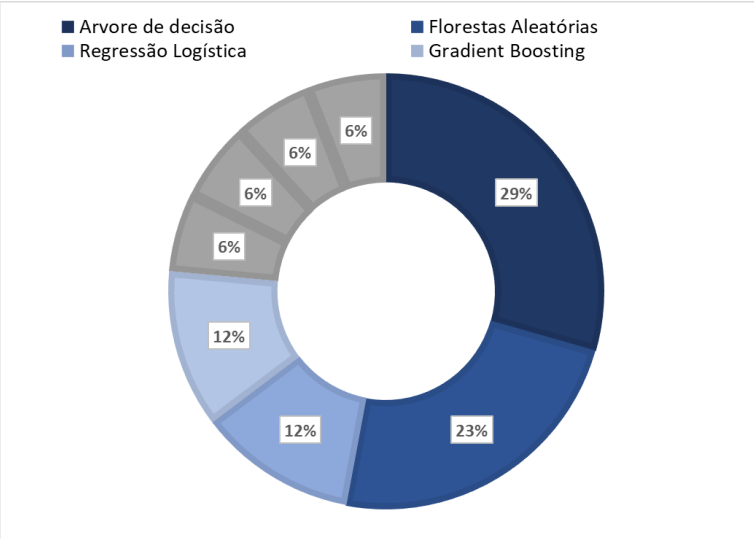
Gráfico 1 – Ocorrência por modelo preditivo nos artigos analisados



Elaborado pelo autor

Foi também possível analisar que em 17 artigos que buscaram compreender e comparar o desempenho de cada modelo para o problema, o modelo árvore de decisão foi considerado o melhor em cinco (29%), seguido por florestas aleatórias que foi avaliado como o melhor por quatro (23%). Regressão logística, apesar de ser o método mais utilizado para a resolução de problemas que envolvem *churn*, foi considerado o melhor em apenas dois artigos (12%)

Gráfico 2 – Melhores modelos de acordo com a bibliografia.



Elaborado pelo autor

Portanto, neste trabalho os modelos regressão logística, árvore de decisão, florestas aleatórias e redes neurais artificiais foram os métodos selecionados, pois são os métodos mais frequentes na literatura e tem mostrado um bom desempenho. Os três primeiros foram considerados os melhores métodos em pelo menos dois artigos de comparação e, apesar das redes neurais artificiais não terem sido consideradas as melhores em nenhum dos artigos, elas foram incluídas por serem o terceiro método mais presente na literatura.

Nas seções a seguir, serão explorados os conceitos, objetivos e vantagens de acordo com a necessidade de conhecimento para a gestão da inteligência de marketing. Para definição com maior rigor matemático dos quatro métodos, o leitor pode consultar James *et al*. (2021).

4.1 Regressão Logística

Foi através de Cox (1970) que a Regressão Logística se tornou um modelo popular no campo da estatística. Pertencente à família de modelos lineares generalizados, a regressão logística é amplamente utilizada quando o evento de interesse permite dois resultados possíveis (*churn*/não *churn*, Y=1 ou Y=0). Esse modelo permite estimar, a partir de um conjunto de observações, a probabilidade de um evento de interesse ocorrer (*churn*; Y = 1) de acordo com uma série de variáveis independentes, sejam elas qualitativas ou quantitativas, denotadas por X_1... Xn.

O modelo estima a probabilidade do evento Y = 1 dado um conjunto de variáveis independentes através de uma função de ligação de tipo logística. A probabilidade é determinada tomando como informações de entrada um conjunto de variáveis independentes e a função pode ser definida por:

$$P(Y=1) = \frac{1}{1+e^{-g(x)}}$$

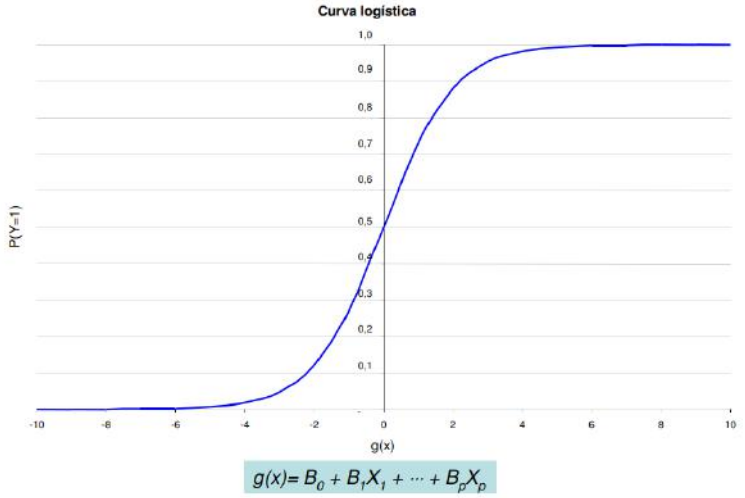
Na qual P(Y=1) é a probabilidade de um evento ocorrer e g(x) é uma combinação linear de todas as variáveis independentes. g(x) é calculada a partir dos coeficientes de regressão de todas as variáveis e pode ser definida por:

$$g(x) = \beta_o + \beta_1 X_1 + \ldots \beta n Xn$$

Onde X_1....Xn representa as variáveis independentes e β_o,β_1...βn são os parâmetros do modelo. Na prática os parâmetros nunca são conhecidos e são estimados a partir do conjunto de dados, sendo o método de máxima verossimilhança o método comumente utilizado.

É possível observar na Figura 2 que a curva logística tem um comportamento no formato da letra S (função sigmoide). Nesse sentido, para problemas de classificação, o modelo é capaz de discriminar dois grupos, onde de P(Y=1) > c, classifica-se como Y=1, caso contrário, classifica-se como Y = 0. Uma escolha comum é c= 0.5, mas outros valores podem também ser utilizados.

Figura 2 – Função Logística.



Fonte: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf

A principal vantagem da Regressão Logística está relacionada a simplicidade, facilidade de interpretação e acurácia satisfatória, o que faz deste um método bastante utilizado (SCHNEIDER, 2016). A partir do uso desse algoritmo, é possível estimar a probabilidade de um evento ocorrer para uma observação isolada contra a probabilidade de o evento não ocorrer. Por exemplo, para o problema deste trabalho o modelo é capaz de estimar se a chance de o cliente ter *churn* é maior se ele possuir mais de 50 anos e ter acesso à internet contra o fato contrário. Além disso, o modelo permite prever o efeito de uma ou mais variáveis independentes sobre a variável dependente binária. Ou seja, é possível identificar as variáveis que são significativas para explicar o *churn*/não *churn*. Por fim, o desempenho desse algoritmo pode, em alguns casos, ser tão bom quanto o da árvore de decisão (RADOSAVLJEVIK, D *et al*, 2010).

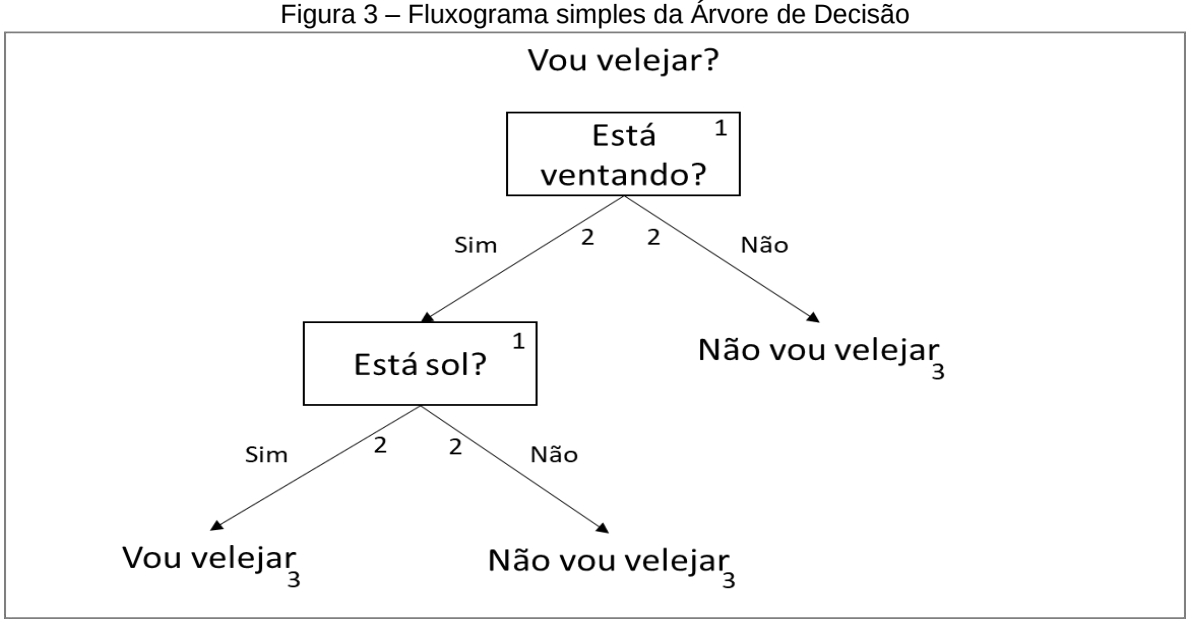
4.2 Árvore de Decisão

Em 1959, William Belson foi o primeiro pesquisador a desenvolver o método de Árvore de Decisão, que recebe esse nome porque pode ser representado através de um fluxograma que remete uma árvore. Segundo Linoff e Berry (2011) a Árvore de Decisão é “uma estrutura que pode ser usada para dividir uma grande coleção de registros em conjuntos sucessivamente menores de registros, aplicando uma sequência de regras de decisão simples”. Ao contrário do método de Regressão

Logística, esse método não é linear. Logo, ele é mais indicado para problemas que envolvam a não linearidade dos dados.

O modelo da árvore de decisão divide-se em dois, dependendo do objetivo da análise: árvores de classificação e árvores de regressão. A primeira tem como o objetivo a predição de determinada classe que a observação pertence. Por outro lado, a árvore de regressão busca entregar como resultado um número real. Como o objetivo desta pesquisa é a de resolução de um problema de classificação, a árvore de classificação será o objeto de exploração.

Figura 3 – Fluxograma simples da Árvore de Decisão



Elaborado pelo autor

A figura 3 representa um fluxograma simples de um problema com um evento de interesse “Vou velejar?” apresentado de forma dicotômica (Y = 1 se o evento ocorrer; Y = 0 caso contrário) e duas variáveis independentes “Está ventando?” (X_1) e “Está sol?” (X_2) também com possibilidade binária de resultado. A árvore é desenhada de maneira invertida, começando pela raiz e terminando em uma das opções de resultado do evento de interesse (Y) do problema (3). A partir do topo, vários pontos de decisões são feitos a partir das variáveis independentes e representados por nós (1), que cada um deles dá origem aos possíveis caminhos/resultados de cada variável, que são representados pelos “ramos” (2).

Para cada decisão a ser feita, serão definidas duas opções de resposta: “não” ou “sim”. A resposta levará a outros nós e ramos subsequentes, construindo assim vários possíveis caminhos que levam a um dos dois resultados.

A escolha da estrutura da árvore é uma tarefa muito importante para o aumento do desempenho e precisão desse método. O conjunto de decisões/variáveis do problema pode ser ordenado de diversas formas. Os algoritmos de aprendizagem de máquina irão buscar quais atributos possuem maior relação com o evento de interesse (Y=1), colocando-as no início de sua cascata de nós.

Os cálculos mais utilizados por esses algoritmos são a Entropia e o Índice Gini. O cálculo de entropia verifica a probabilidade de ocorrer o evento de interesse a partir de seleção aleatória de uma variável. O resultado verifica o nível de desordem dos dados ao calcular a variação do evento de interesse (Y=1) em cada variável independente. Quanto maior a desordem, menor o ganho de informação que a variável independente analisada oferece ao algoritmo.

Em outras palavras, a entropia revela o nível de relação que cada variável possui com o evento de interesse do problema. Variáveis com maior ganho de informação e ordem são variáveis que possuem forte relação com o resultado final, logo são ordenadas pelos algoritmos como as primeiras decisões a serem tomadas, deixando as variáveis com maior desordem e menor ganho de informação ao fim.

O Índice Gini também evidencia a distribuição dos dados do evento de interesse em relação à cada variável independente, mas com um cálculo diferente. Nesse método, quanto menor o valor do índice Gini, maior será a prioridade da variável ser escolhida para os primeiros nós da árvore.

O uso da árvore de decisão se tornou popular porque não requer grande configuração de variáveis e análise técnica em sua aplicação e, em geral, o algoritmo apresenta boa precisão em problemas de classificação (HAN; KAMBER, 2006). Além disso, comparados a outras técnicas como redes neurais, é possível observar e compreender as operações lógicas até o resultado. O processo de decisão do algoritmo se assemelha mais à tomada de decisão humana do que outros métodos de classificação (JAMES, G. *et al*, 2021).

As desvantagens do uso do método são o menor desempenho na interpretação de relacionamentos não lineares entre atributos e a característica de que se há muitas variáveis preditivas com baixa qualidade no conjunto de dados, o algoritmo tende a criar árvores demasiadamente adaptadas, gerando alto custo de processamento e

padrões altamente irregulares, levando, muitas vezes, a problema de sobreajuste (*overfitting)*, que é quando o modelo tem um bom desempenho nos dados de treino, mas baixo desempenho nos dados de teste.

4.3 Florestas Aleatórias

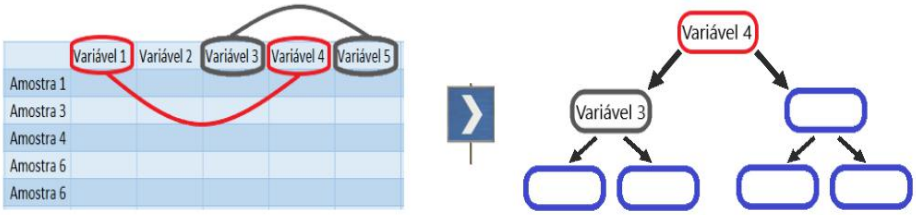
Desenvolvido por Leo Breiman em 2001, Florestas Aleatórias ou, como conhecido em inglês, *Random Forest*, consiste em um método de aprendizagem de máquina combinado, ou seja, é um modelo formado por um conjunto de modelos mais simples (MISHRA; REDDY, 2017). Comumente utilizados em problemas de classificação, os algoritmos de métodos combinados criam muitos modelos não relacionados e o resultado é a combinação de todos os resultados dos modelos gerados. Esta metodologia torna esses algoritmos mais complexos, podendo levar a melhores resultados, acompanhados de um maior custo computacional.

No algoritmo de florestas aleatórias, são criadas múltiplas árvores de decisão nos dados de treinamento e retorna como resultado a classe que tem maior frequência dentre todas as árvores geradas.

Como etapa inicial da criação da primeira árvore, o algoritmo seleciona aleatoriamente algumas amostras do conjunto de dados de treino. Para isso é utilizado um método de amostragem com reposição, ou seja, onde as amostras selecionadas podem ser repetidas na seleção seguinte.

A segunda etapa é a de seleção e ordenação das variáveis para cada nó. Ao contrário do método de apenas uma árvore de decisão onde o cálculo de entropia é feito comparando todas as variáveis do conjunto, as florestas aleatórias realizam esse cálculo com base em cada amostra, selecionando aleatoriamente apenas um subconjunto de variáveis e definindo entre essas qual será utilizada no primeiro nó. Essa seleção é repetida para os nós subsequentes, excluindo as variáveis que já foram selecionadas anteriormente.

Figura 4 – Seleção e ordenação de variáveis, Florestas Aleatórias.



Fonte: https://didatica.tech/wp-content/uploads/2019/11/Selecao_de_variaveis.png

Após a criação da primeira árvore, as duas etapas anteriores são feitas novamente, criando árvores que provavelmente serão diferentes das anteriores, uma vez que as amostras e variáveis são selecionadas de forma aleatória. Quanto maior o número de árvores criadas, melhor serão os resultados do algoritmo. Entretanto, a partir de um determinado ponto, a adição de nova árvore trará uma melhoria marginal ao modelo e pode não compensar o aumento do custo computacional.

O principal objetivo das florestas aleatórias é minimizar o problema de sobreajuste da árvore de decisão. Quanto maior a quantidade de variáveis no problema, maior será a profundidade da árvore, o que tende a construir um modelo que responde muito bem a dados conhecidos, mas perde poder de generalização e passa a responder mal a dados desconhecidos (o já mencionado problema de *overfitting*). As Florestas Aleatórias minimizam esse problema porque trabalham com subconjuntos aleatórios de variáveis, constroem árvores menores a partir desses subconjuntos e só após o treinamento essas árvores são combinadas. Nesse sentido, como uma única árvore de decisão pode ser sensível ao ruído do conjunto de dados, o resultado médio de muitas árvores criadas a partir de amostras aleatórias e não correlacionadas dilui esse problema (SCHNEIDER, 2016).

É interessante utilizar esse método quando o conjunto de dados possui grande quantidade de variáveis, mas o seu uso pode gerar um maior custo computacional e tempo de processamento, portanto pode não ser recomendado em problemas que requerem a execução e resultado da predição em tempo real. Além disso, o modelo torna as operações lógicas complexas demais para interpretação clara do resultado, ao contrário da utilização da árvore de decisão simples.

4.4 Redes Neurais Artificiais

O primeiro modelo de Redes Neurais Artificiais (RNA) foi desenvolvido em 1958 pelo psicólogo Frank Rosenblatt. Chamado de *Perceptron*, o modelo buscava simular como o cérebro humano processa os dados visuais e reconhece objetos. A ideia era de criar uma unidade lógica de processamento que recebesse diferentes dados de entrada e na saída informasse o sinal de 1 ou 0. No caso de resultado igual a 1, o neurônio seria ativado, enviando sinal para outros neurônios. Caso contrário, esse neurônio não enviaria informação para a rede.

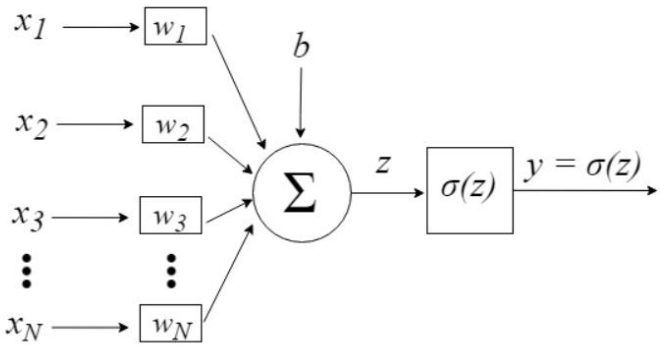
Apesar de revolucionário para a época, foi apenas na década de 1980, com a evolução da capacidade de processamento computacional, que os métodos de aprendizagem de máquina envolvendo RNA começaram a ser aplicados em uma variedade de propósitos.

Redes Neurais Artificiais é uma família de métodos e são utilizados para estimar ou aproximar funções que dependem de muitos parâmetros de entrada. Esse processo é realizado por uma estrutura em camadas constituída por “neurônios”, que são elementos computacionais ligados em rede de tal forma que a primeira camada de neurônios recebe os dados de treino, executa cálculos e envia esses resultados para o neurônio da camada seguinte, que por sua vez, buscando ajustar os erros, repete o processo até a última camada que resulta na decisão de classificação do algoritmo. Nesta seção, foram explorados o conceito de funcionamento do neurônio artificial e como os algoritmos operam a estrutura de combinação de neurônios em camadas.

Estrutura de um neurônio

A figura 5 representa o funcionamento da unidade computacional das Redes Neurais. Na área da inteligência artificial, esse modelo computacional foi construído em referência aos neurônios biológicos:

Figura 5 – Modelo de um neurônio



Fonte: https://miro.medium.com/max/705/1*KDiqpWOgtCnO8x3wZJHmDA.png

O vetor X = [X1, X2, ... , Xn] representa os dados de entrada no neurônio, como por exemplo variáveis de um conjunto de dados sobre *churn* ou pixels de uma imagem. Uma vez inseridos ao sistema, esses dados são multiplicados pelos respectivos pesos, chamados de pesos sinápticos, representados pelo vetor W = [W1, W2, ... , Wn] que precisam ser estimados, gerando o potencial de ativação do neurônio, representado pelo valor z, que pode ser expresso pela expressão:

$$z = \sum_{i=1}^{N} x_i w_i + b$$

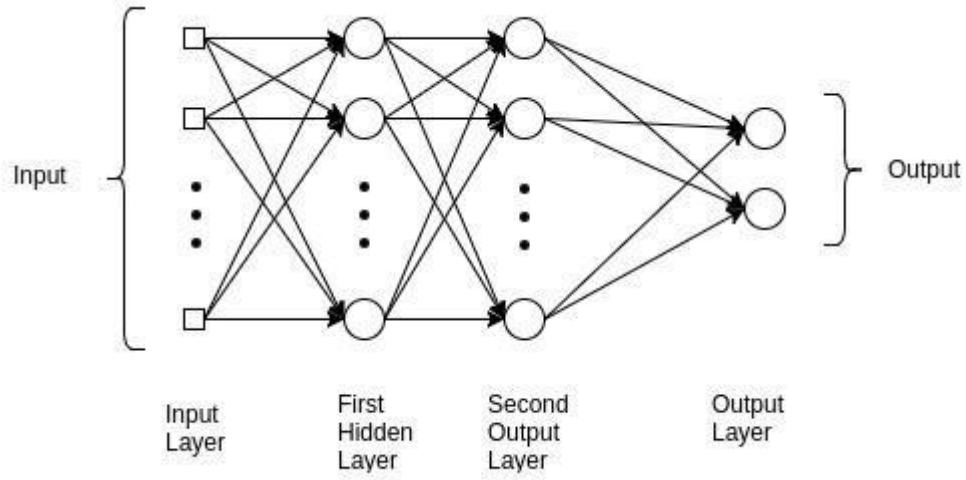
O parâmetro b é uma constante chamada de viés que não depende da entrada nesse cálculo. Após calculado, é aplicado ao valor z, a função responsável por limitar o valor encontrado a um certo intervalo, que geralmente é a função sigmoide, responsável por transformar o z em um valor entre 0 e 1 – assim como no modelo de Regressão Logística (JAMES et al., 2021). Por fim, o resultado dessa função é definido como parâmetro de saída (Y) do neurônio e, ao contrário do primeiro modelo de *Perceptron* criado, a abordagem atual possibilita como saída não apenas resultado binário, mas um valor entre 0 e 1 que, quanto mais próximo do 1, maior é a contribuição e potencial de ativação do neurônio para a rede.

Estrutura de neurônios em rede

Apesar de haver muitas abordagens na literatura para a construção de algoritmos envolvendo o conceito de neurônios artificiais, o mais presente na bibliografia de classificação de *churn* é o modelo *Multilayer Perceptron*, que consiste

em uma rede neuronal totalmente conectada, ou seja, cada neurônio recebe como dados de entrada os dados de saída de todos os neurônios que compõe a camada anterior. Os cálculos realizados para determinar o potencial de ativação são os mesmos, o que difere é que agora o valor de entrada pode não ser o valor da variável inicial, mas sim os valores de saída dos neurônios da camada anterior. Na figura 6, é possível visualizar um modelo de rede com quatro camadas. O vetor x representa os dados que entram no modelo pela camada inicial, que por sua vez gera valores de saída z que servirão como entradas da camada seguinte. Esse processo segue até a última camada, que fornece os resultados de saída para o problema.

Figura 6 – Multicamadas de neurônios



Fonte: https://miro.medium.com/max/724/1*eloYEyFrblGHVZhU345PJw.jpeg

No exemplo, as duas camadas do meio são denominadas como “camadas ocultas”, pois o valor de cada neurônio dessas camadas é geralmente desconhecido, não sendo possível sua interpretação humana. Quando as redes neurais possuem muitas camadas, trata-se da área de aprendizado de máquina conhecida como aprendizado profundo. Nesse caso, a quantidade de cálculos e ajustes de pesos e vieses se torna exponencialmente maior, ganhando a possibilidade de encontrar resultados mais precisos, em contrapartida o custo computacional é muito maior e o conhecimento do processo se torna desconhecido, tendo apenas como evidências do modelo os valores de entrada e saída.

A etapa de treinamento é essencial para a construção de uma rede neural de boa precisão. Assim como o processo cognitivo humano, as redes neurais aprendem

a classificar a partir da identificação de padrões de um conjunto de dados inseridos no sistema. Para isso, é necessário que seja inserido um grande volume de dados com rótulos informando a qual classe aquele registro pertence.

Uma das formas de aprendizado da rede neural é a de retroalimentação. Nessa estratégia, o algoritmo calcula a função erro na camada de saída da rede neural (resultado da predição *vs* rótulo esperado), e com isso recalcula o valor dos pesos do vetor W no sentido da última camada para a primeira camada com o objetivo de diminuir o valor do erro. De forma simplificada, o processo lógico de aprendizado por retroalimentação consiste em:

1) Como primeiro passo, o modelo atribui valores aleatórios para todos os parâmetros da rede.
2) O pesquisador fornece ao modelo os resultados esperados para cada observação (conjunto de dados de treino). Com esses resultados é possível calcular o valor da função de erro total obtida no modelo.
3) Para diminuir o valor de erro total, o algoritmo utiliza os valores dos gradientes para cada parâmetro da rede. O valor de gradiente fornece ao modelo a direção de maior crescimento de uma função, logo é utilizado o sentido contrário ao do gradiente para encontrar o caminho de redução da função erro.
4) Por fim, os pesos de cada neurônios são atualizados de modo iterativo, recalculando os gradientes em cada iteração até diminuir o erro de forma ótima ou acabar a quantidade de iterações possíveis da estrutura estabelecida.

Atualmente as Redes Neurais podem ser utilizadas em problemas de classificação assim como o de *churn*. Para o problema desta pesquisa, a unidade de saída da última camada do modelo pode representar as duas classes possíveis (Y = 1;Y = 0). A utilização deste método para a predição de *churn* possui algumas desvantagens em relação a outros métodos. O algoritmo de RNA chega a conclusões de forma empírica, ou seja, através de ajustes dos parâmetros de forma cíclica até encontrar o resultado ótimo. Isso gera a uma enorme complexidade de operações dentro da estrutura, necessitando grande capacidade de processamento computacional e grande volume de dados de treino. Além disso, a principal desvantagem para a predição de *churn* é que este método é a dificuldade de interpretação das operações lógicas até o resultado, gerando poucos insights e

conhecimento de negócio sobre o perfil do consumidor em *churn*. Apesar dessas desvantagens, os modelos de Redes Neurais Artificiais são capazes de entregar boa acurácia mesmo quando as relações entre as variáveis são desconhecidas ou não lineares, conseguindo assim processar grandes bases de dados como geralmente são os casos que envolvem *churn*.

4.5 Exemplo de aplicação

Esta seção busca exemplificar ao leitor a resolução de um problema de predição de *churn* utilizando os algoritmos de aprendizado de máquina descritos neste trabalho. O objetivo não é gerar novo conhecimento sobre a população estudada, mas servir como Ilustração de como os métodos podem ser aplicados utilizando a linguagem de programação Python, servindo como guia prático no uso da linguagem *Python* para a aplicação dos algoritmos.

4.5.1 Análise Exploratória

A base de dados utilizada contém informações sobre 7.043 clientes de uma empresa de telecomunicações disponibilizada publicamente pela IBM (https://www.kaggle.com/blastchar/telco-customer-churn). A Tabela 3 informa as 21 variáveis de informações dos clientes presentes no conjunto de dados, bem como a descrição e o tipo de variável:

Tabela 3 – Variáveis do conjunto de dados de telecomunicações

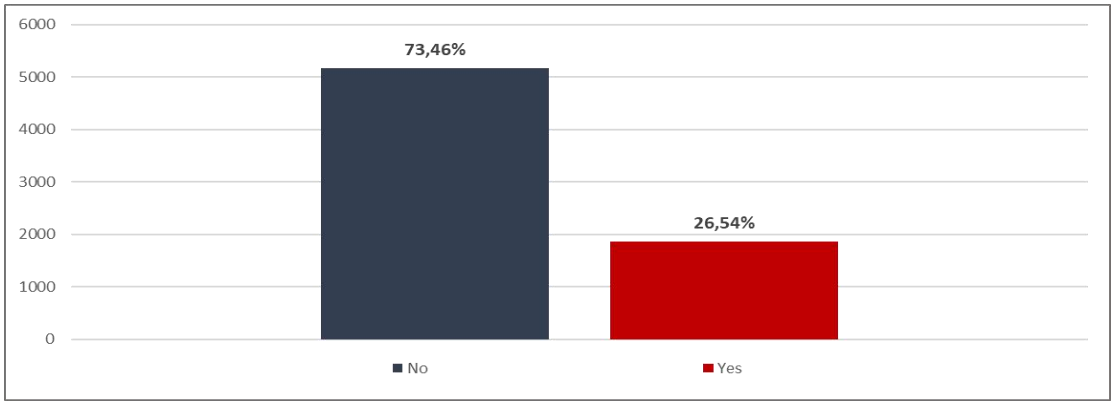
Campo	Descrição	Tipo de variável
ID	Identificação do Cliente	Numérica
gênero	Se o cliente é homem ou mulher	Categórica
Senioridade	Se o cliente é idoso ou não (1, 0)	Categórica
Parceiro	Se o cliente tem um parceiro ou não (Sim, Não)	Categórica
Dependentes	Se o cliente tem dependentes ou não (Sim, Não)	Categórica
Tempo de casa	Número de meses que o cliente permaneceu na empresa	Numérica
Telefonia	Se o cliente tem serviço telefônico ou não (Sim, Não)	Categórica
Múltiplas linhas	Se o cliente tem várias linhas ou não (Sim, Não)	Categórica

Campo	Descrição	Tipo de variável
Internet	Provedor de serviços de Internet do cliente (DSL, Fibra ótica, Não)	Categórica
Segurança Online	Se o cliente tem segurança online ou não (Sim, Não)	Categórica
Backup	Se o cliente tem backup online ou não (Sim, Não)	Categórica
Proteção dispositivo	Se o cliente tem proteção do dispositivo ou não (Sim, Não)	Categórica
Support Tec.	Se o cliente tem suporte técnico ou não (Sim, Não)	Categórica
TV Streaming	Se o cliente tem streaming de TV ou não (Sim, Não)	Categórica
Filme Streaming	Se o cliente tem streaming de filmes ou não (Sim, Não, Sem serviço de internet)	Categórica
Prazo Contrato	O prazo do contrato do cliente (mês a mês, um ano, dois anos)	Numérica
Fatura física	Se o cliente tem faturamento sem papel ou não (Sim, Não)	Categórica
Método de Pagamento	O método de pagamento do cliente (Débito automático, Cheque enviado, Transferência bancária (automática), Cartão de crédito (automático))	Numérica
Cobrança mensal	O valor cobrado do cliente mensalmente	Numérica
Cobrança Total	O valor total cobrado do cliente	Numérica
Churn	Se o cliente é *churn* ou não (Sim ou Não)	Categórica

Realizou-se análise exploratória do conjunto dos dados sobre os clientes do setor de telecomunicações com o objetivo de entender a distribuição das categorias em relação à variável *churn*, identificar necessidades de tratamento dos dados para a aplicação dos algoritmos de aprendizado de máquina e, além disso, é importante mencionar que a análise exploratória de dados é um processo importante antes de qualquer processo de modelagem de dados, pois nos permite conhecer melhor as variáveis a serem analisadas, estabelecer algumas hipóteses e/ou encontrar possíveis observações atípicas

A distribuição de clientes do rótulo *Churn* é desbalanceada, há aproximadamente 27% (1869 usuários) de clientes em *churn* e 73% (5.174 usuários) em não *churn*. Essa distribuição desbalanceada pode ser um problema para obter o melhor resultado de um modelo que possui como principal objetivo identificar os clientes que não estão mais ativos.

Gráfico 3 – Distribuição do rótulo *Churn*



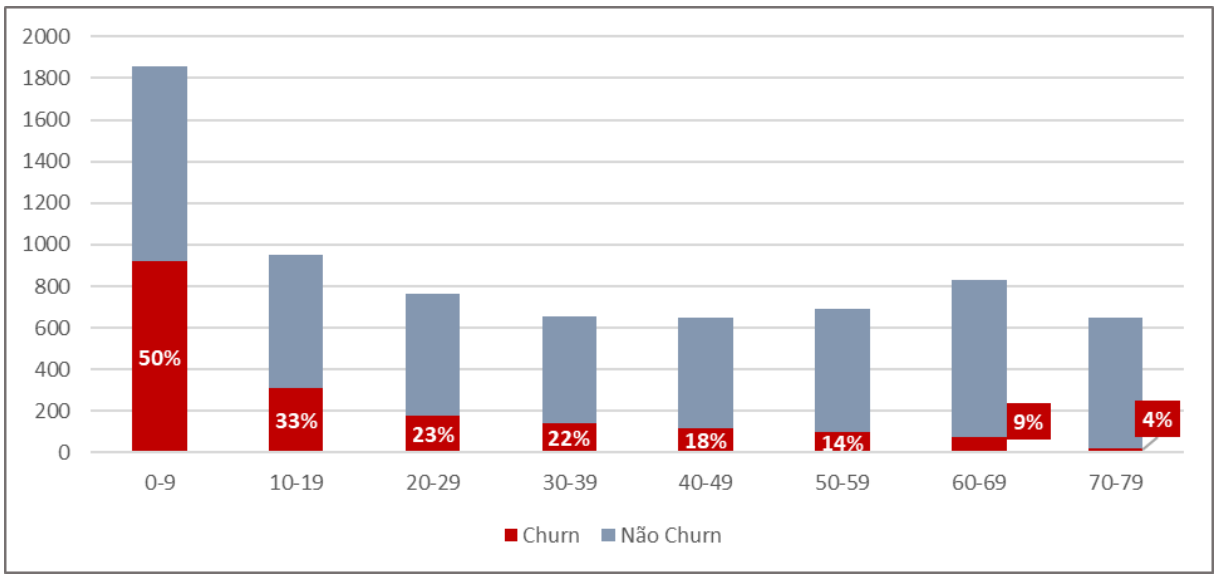
Elaborado pelo autor

Foram identificadas variáveis que não possuem forte correlação linear com a variável de interesse *churn*. A quantidade de clientes mulheres e homens é quase a mesma e a proporção de *churn* entre eles é quase idêntica. O fato de o cliente possuir serviço de telefonia ou múltiplas linhas telefônicas também não possui forte correlação com a probabilidade de *churn*, dado a distribuição dos dados. Caso o algoritmo precisasse ser otimizado e uma das estratégias fosse a de redução de variáveis, esses campos poderiam ter sido desconsiderados após novos testes de associação entre as variáveis. Como não é o objetivo desta aplicação, foram considerados nos modelos todas as variáveis explicativas.

Na análise exploratória também foram analisadas as variáveis com maior correlação com a variável dependente. É possível identificar que a taxa de *churn* é maior no caso de clientes que possuem método de débito automático em comparação aos demais métodos, visto que há maior facilidade no cancelamento. O pagamento por débito automático possui 45,29% em *churn*, enquanto por transferência, cartão de crédito e cheque enviado possuem 17%, 15% e 19% respectivamente.

A variável de tipo de serviço de internet também possui forte correlação com o *churn*. Clientes que possuem serviço de fibra óptica são o grupo com maior proporção de *churn* com 42%, enquanto o serviço por linha telefônica (*DSL*) possui apenas 19% de *churn*. Além disso, a taxa de *churn* é maior em usuários que são clientes a pouco tempo.

Gráfico 4 – *Churn* por tempo de contrato em meses.



Elaborado pelo autor

4.5.2 Preparação dos Dados

Após a exploração dos dados, realizou-se a preparação dos dados de entrada dos algoritmos. Alguns algoritmos aceitam apenas valores numéricos como dados de entrada, logo adotou-se a estratégia de transformar todos as variáveis em variáveis numéricas. Para isso, foi utilizada a função "*LabelEncoder*()" da biblioteca sklearn para enumerar as variáveis que possuem uma ordenação natural entre as possibilidades. Para as variáveis categóricas que não possuem uma ordenação natural, foi utilizada a função *"get_dummies()"* da biblioteca pandas, que transforma todos os possíveis valores de um campo em uma nova coluna com resultado binário (ver Figura 7).

Figura 7 – Transformação de variáveis em numéricas.

```
dummies = ['Contract', 'PaymentMethod', 'InternetService']
df2 = pd.get_dummies(df[dummies])

categorias = ['gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'OnlineSecurity',
              'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',
              'PaperlessBilling', 'tenure', 'TotalCharges', 'MonthlyCharges']

for cat in categorias:
    num_label = LabelEncoder().fit_transform(df[cat])
    df2[cat + '_cat'] = num_label
```

Elaborado pelo autor

Ao final da transformação dos dados, foi utilizada a função *"train_test_split()"* (ver Figura 8) que pertence a biblioteca *sklearn* para a divisão aleatória do conjunto

de dados em dois: 80% dois dados foram determinados como o conjunto de treino (*x_train* e *y_train*) e 20% como o conjunto de teste (*x_valid*, *y_valid*).

Figura 8 – Função divisão do conjunto de dados

```
x_train, x_valid, y_train, y_valid = train_test_split(data, target, test_size=0.2)
```

Elaborado pelo autor

4.5.3 Aplicação e Avaliação dos Modelos

Após preparação dos dados, realizou-se o treino (método "*.fit*()") e predição de novos clientes "*churn*"/"não *churn*" (método "*.predict*()") dos algoritmos escolhidos através das funções "*RandomForestClassifier*().", "*DecisionTreeClassifier*()", "*MLPClassifier*()" e "*LogisticRegression*(*penalty='none'*)", que são algoritmos pertencentes à biblioteca sklearn e representam os modelos estatísticos Florestas Aleatórias, Árvore de Decisão, Multi-layer Perceptron (Redes Neurais Artificiais) e Regressão Logística, respectivamente. Também foi utilizada a função "*DummyClassifier(strategy="most_frequent")",* que gera um modelo com classificação do valor mais frequente para os dados de saída, ou seja, o modelo classifica todos os resultados como "não *churn*", que representa a maior parte do conjunto analisado. O objetivo é conseguir comparar a performance dos demais modelos com um modelo simples e pouco eficaz (*baseline model*), ver Figura 9.

Figura 9 – Treinamento e aplicação dos modelos

```
# Treinando os modelos #
LR_method = LogisticRegression(penalty='none')
LR_method.fit(x_train, y_train)

AD_method = DecisionTreeClassifier()
AD_method.fit(x_train, y_train)

RF_method = RandomForestClassifier()
RF_method.fit(x_train, y_train)

RNA_method = MLPClassifier()
RNA_method.fit(x_train, y_train)

BM_method = DummyClassifier(strategy="most_frequent")
BM_method.fit(x_train, y_train)

# Aplicando os dados de teste #
LR_previsao = LR_method.predict(x_valid)
AD_previsao = AD_method.predict(x_valid)
RF_previsao = RF_method.predict(x_valid)
RNA_previsao = RNA_method.predict(x_valid)
BM_method = BM_method.predict(x_valid)
```

Elaborado pelo autor

Por último, foram utilizadas as funções “accuracy_score()” e “classification_report()” da biblioteca sklearn para a validação dos resultados com o rótulo classificador de *churn* do conjunto de teste e cálculo das métricas de desempenho: Acurácia, precisão, *recall* e F-1 *Score*. Abaixo (Tabela 4), os resultados obtidos por modelo:

Tabela 4 – Métricas obtidas dos modelos.

Modelo	Acurácia	Precisão	Recall	F-1 Score
Regressão Logística	80,2%	60%	66%	63%
Arvore de Decisão	72,2%	48%	47%	47%
Florestas Aleatórias	78,4%	50%	60%	55%
Redes Neurais	79,3%	53%	62%	57%
Baseline Model	74,3%	-	-	-

Elaborado pelo autor

Ao analisar os resultados das métricas, é possível concluir que a Regressão Logística foi o modelo com o melhor desempenho para esse conjunto de dados. Ao comparar com o *baseline model*, todos os modelos apresentaram acurácia um pouco superior, que avalia de maneira geral a capacidade do modelo em classificar corretamente as duas opções possíveis.

Entretanto, a escolha da melhor métrica depende do problema de negócio e da distribuição dos dados. No caso deste problema, a acurácia não é a melhor métrica para avaliar a qualidade dos modelos, já que a distribuição de clientes entre os rótulos “*churn*” e “não *churn*” é muito desbalanceada. Portanto, como nosso objetivo principal é a classificação correta do rótulo positivo “*churn*”, as métricas “precisão” e “*recall*” são as que melhor avaliam a qualidade do modelo em prever especificamente esse rótulo. No caso do *baseline model*, as métricas precisão e recall não se aplicam, dado que o modelo classifica todos os rótulos como negativos.

A precisão do melhor modelo aplicado – Regressão Logística - é de média 60%, ou seja, o modelo acertou mais da metade dos clientes que foram identificados como “*churn*”.

O resultado de 66% em *recall* nos dá a informação que a Regressão Logística foi capaz de classificar corretamente dois terços dos clientes que são de fato “*churn*”, logo um terço desses clientes foram classificados como “não *churn*” e representam a classe de falsos negativos.

5 CONSIDERAÇÕES FINAIS

Em um mundo onde os produtos, processos e relações estão cada vez mais digitalizados, problemas de *big data* e soluções que envolvem aprendizado de máquina estão cada vez mais presentes em todas as áreas das organizações. Portanto, há necessidade de o profissional de administração entender as técnicas preditivas que dão suporte a tomada decisão gerencial na análise dos mais variados problemas, dentre eles a retenção de clientes.

Foi possível observar a partir de análise bibliográfica que há grande concentração de uso em poucos métodos preditivos quando o objetivo é resolver problemas de classificação de *churn*. Para esse problema, os cientistas de dados optam primeiramente por métodos de fácil interpretação das operações do algoritmo e que consigam descrever o perfil do cliente, bem como as variáveis independentes que possuem maior relação com o *churn*. Por isso, a Regressão Logística e a Árvore de Decisão são os métodos mais utilizados.

Entretanto, em pesquisas onde o objetivo principal era encontrar os métodos com melhor desempenho, foi observado o uso de métodos mais complexos como Florestas Aleatórias e Redes Neurais, que prometem maior desempenho e menor capacidade de interpretação e geração de *insights.*

A vantagem em testar e usar mais de um método para resolver problemas de classificação é que há maneiras simples de comparar o desempenho entre modelos a partir de métricas calculadas através da matriz de confusão, como acurácia, precisão e *recall*. É importante entender a distribuição de dados do conjunto e a necessidade do problema de negócio, pois para cada situação uma das métricas pode ser a mais relevante na escolha do melhor modelo.

Hoje há inúmeros algoritmos desenvolvidos e pré-configurados pela comunidade de ciência de dados. Através da linguagem *Python* e suas bibliotecas de aprendizado de máquina*,* é possível começar a aplicar com rapidez modelos complexos como o de Redes Neurais e Florestas Aleatórias com conhecimento intermediário de programação. Em muitos casos, a maior dificuldade para o pesquisador/analista/administrador, pode ser a interpretação, seleção de variáveis apropriadas para abordar o problema e escolha do melhor método preditivo. No caso da aplicação e utilizando os dados de teste deste trabalho, o melhor modelo foi a

regressão logística, mas isso não significa que esse método sempre dê os melhores resultados. Na prática, para cada caso será necessário realizar a análise exploratória de dados para melhor utilizar os modelos disponíveis. Nesse sentido, este trabalho busca ajudar a organizar o conhecimento e elucidar as vantagens e desvantagens dos métodos mais populares na literatura para resolver problemas de fidelização de clientes.

Por fim, para próximos passos, é possível expandir a pesquisa através de novos termos de busca, tais como "*machine-learning"*, *"classification"*, *"*modelagem*"*, que também podem fazer referência ao problema desta pesquisa. Há também outros métodos de avaliação de desempenho dos modelos que podem ser abordados, como a métrica de curva ROC e validação cruzada, porém não é o foco deste trabalho. Além disso, há a oportunidade de exemplificar as aplicações através de outras linguagens também populares, como R e Julia.

REFERÊNCIAS

1. GANDOMI, T; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics**. Science Direct**, Toronto, 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401214001066. Acesso em: 10 de outubro de 2021.

2. FAYYAD, U *et. al*. Advances in knowledge discovery and data mining: Towards a Unifying Framework. **Association for the Advancement of Artificial Intelligence**, Redmond, 1996. Disponível em: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf. Acesso em: 10 de outubro de 2021.

3. IDC Corporate**. Worldwide Big Data and Analytics Spending Guide**. IDC Corporate, 2021. Disponível em: https://www.idc.com/getdoc.jsp?containerId=prUS48165721. Acesso em: 10 de outubro de 2021.

4. NESLIN, S. *et al*. Defection detection: measuring and understanding the predictive accuracy of customer *churn* models. **Journal of Marketing Research**, Nova Iorque, 2006. v. 43, n. 2, p. 204-211.

5. PEPPERS, D; ROGERS, M. **Marketing 1 to 1: um guia executivo para entender e implantar estratégias de Customer Relationship Management (CRM Series). 2. ed**. São Paulo: Makron Books, 2001.

6. VERGARA, Sylvia Constant. **Sugestão para estruturação de um projeto de pesquisa. Caderna de Pesquisa**. Rio de Janeiro: EBAP/FGV n°2, 1991.

7. VERGARA, Sylvia Constant. **Projetos e Relatórios de Pesquisa em Administração. São Paulo**: ATLAS S.A., 1998. 87 p.

8. KOTLER, P; ARMSTRONG, G. **Princípios de Marketing**. São Paulo: Atlas, 1999.

9. JAMES, G. *et al*. **An Introduction to Statistical Learning with Applications in R. 2 ed.** Palo Alto: Springer, 2021. Disponível em: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf Acesso em: 15 de janeiro de 2022

10. KATELARIS, L; THEMISTOCLEOUS, M. **Predicting Customer *Churn*: Customer Behavior Forecasting for Subscription-Based Organizations**. Atenas: Springer, 2017. p. 128-135.

11. RADOSAVLJEVIK, D; PUTTEN, P; LARSEN, K. **The impact of experimental setup in prepaid churn prediction for mobile telecommunications: what to predict, for whom and does the customer experience matter?**. Bonn, 2010. Disponível em: https://www.researchgate.net/publication/220593770_The_Impact_of_Experimental_Setup_in_Prepaid_Churn_Prediction_for_Mobile_Telecommunications_What_to_Predict_for_Whom_and_Does_the_Customer_Experience_MatterTrans MLDM Acesso em: 10 de janeiro de 2022

12. LINOFF, G; BERRY, M. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management**. Indianapolis: John Wiley & Sons, 2011.

13. R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing**. Vienna, 1993. Disponível em: https://www.R-project.org. Acesso em: 05 de janeiro de 2022

14. PYTHON SOFTWARE FOUNDATION. **Python Language**. Amsterdã, 1991. Disponível em: https://www.python.org/doc/. Acesso em: 06 de dezembro de 2021.

15. BEZANSON, J. *et al***. Julia Language**. Massachusetts: Massachusetts Institute of Technology, 2012. Disponível em: https://julialang.org/. Acesso em: 05 de fevereiro de 2022.

16. RYAN, D; JONES, C. **Understanding digital marketing: marketing strategies for engaging thedigital generation.** Londres e Filadélfia: Kogan Page Publishers, 2016. Disponível em: https://link.springer.com/content/pdf/10.1057/dddmp.2009.7.pdf Acesso em: 15 de dezembro de 2021.

17. STAUSS, B; FRIEGE, C. **Regaining Service Customers: Costs and Benefits of Regain Management**. Eichstätt: Journal of Service Research, 1999. p. 347-361. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/109467059914006. Acesso em: 15 de dezembro de 2021.

18. FROST & SULLIVAN. **O Mercado de Big Data na América Latina irá triplicar até 2022, impulsionado por soluções de análise em tempo real**. Frost & Sullivan, 2017. Disponível em: https://www.frost.com/news/press-releases/o-mercado-de-big-data-na-america-latina-ira-triplicar-ate-2022-impulsionado-por-solucoes-de-analise-em-tempo-real/. Acesso em: 05 de dezembro de 2021.

19. COX, D. **The Analysis of Binary Data.** Londres: Chapman & Hall, 1970.

20. FRIDRICH, M. **Hyperparameter Optimization of Artificial Neural Network in Customer Churn Prediction using Genetic Algorithm.** Brno: Trends Economics and Management, 2017. Disponível em: https://trends.fbm.vutbr.cz/index.php/trends/article/view/385 Acesso em: 05 de dezembro de 2021.

21. HUANG, N. *et al.* **Customer churn prediction in telecommunications.** Dublin: Elsevier, 2011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417411011353. Acesso em: 05 de dezembro de 2021.

22. SCHNEIDER, P**. Análise preditiva de Churn com ênfase em técnicas de Machine Learning: Uma Revisão**. Rio de Janeiro: FGV, 2016

23. HAN, J.; KAMBER, M**. Data Mining:Concepts and Techniques.** Morgan Kaufmann, Elsevier, 2006.

24. MISHRA, A.; REDDY, U. **A comparative study of customer churn prediction in telecom industry using ensemble based classifiers.** Coimbatore: ICICI, 2017.